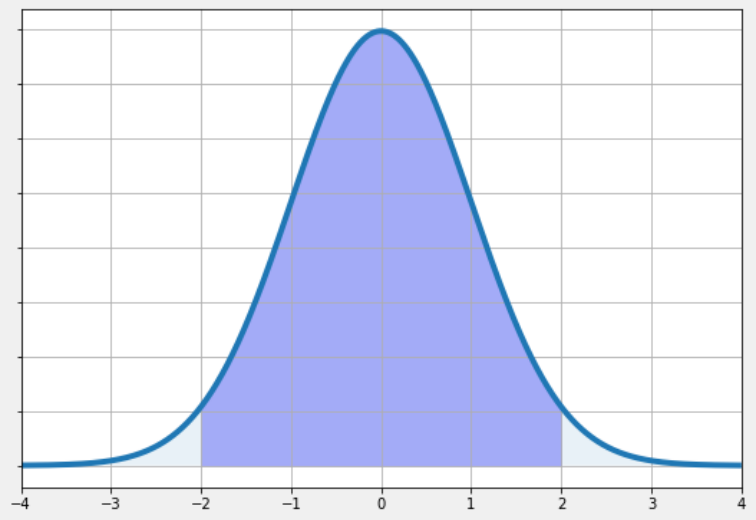
Notas de Probabilidad y Estadística

Por Pablo De Nápoli



;

Versión 2023.5.2

Departamento de Matemática - FCEyN

Universidad de Buenos Aires

Nota: No es una publicación oficial del Departamento de Matemática.

Prólogo a la versión 2022:

Las presentes *Notas de Probabilidad y Estadística* (para matemáticos) las he ido escribiendo a lo largo de las diferentes cursadas en las que estuve como profesor en esta materia (2006, 2010, 2016 y 2021).

Su objetivo es ser una ayuda para facilitar el seguimiento y la comprensión de las clases teóricas, y también que los estudiantes interesados puedan profundizar en algunos temas (que en muchos casos están desarrollados con más extensión que en las clases teóricas).

El contendio del curso ha ido evolucionando a lo largo del tiempo, entre otras cosas para adaptarse a los cambios en las materias anteriores. Es probable por ello que en la cursada 2022 no siga exactamente el mismo enfoque en algunas partes, aunque este material va a seguir sirviendo como referencia.

Consideraciones sobre el enfoque elegido

Una seria dificultad que se presenta en esta asignatura es que el desarrollo riguroso de la teoría de probabilidades está indisolublemente ligado a la *teoría de la medida* (Integral de Lebesgue). Este teoría se desarrolla en la asignatura *Análisis Real / Medida y probabilidad*. Pero el Departamento de Matemática ha decidido (ya hace años) que la presente asignatura esté antes en el régimen de correlatividades. Por lo que en el momento de cursarla, los estudiantes no conocen esta herramienta.

Teniendo en cuenta esta dificultad, en el curso he seguido el siguiente enfoque: en el capítulo 3 se presentan primero los conceptos fundamentales de la materia en el contexto de variables aleatorias discretas. En dicho contexto, se opera con sumas finitas o infinitas (series), por lo que las demostraciones no presentan dificultades técnicas.

Posteriormente, en el capítulo , se generalizan estos conceptos al caso de variables aleatorias continuas, pero por lo general se omiten las demostraciones (que muchas veces presentan dificultades técnicas fáciles de resolver si uno conoce la teoría de la integral de Lebesgue). Si bien haremos énfasis en el caso más importante en la práctica de las variables absolutamente continuas (que poseen una densidad de probabilidad), el desarrollo de la

teoría general requiere una definición general de la esperanza matemática que se aplique en todos los casos. Esto puede hacerse utilizando la integral de Riemman-Stieltjes (como haremos en estas notas, siguiendo a V. Yohai) o la integral de Lebesgue.

Se aclara que el material de los apéndices no forma parte del contenido del curso, y no se toma en los exámenes finales. En particular, en el apéndice D se presenta un resumen de los resultados esenciales de la teoría de la integración de Lebesgue, algunos de los cuáles se utilizarán (sin demostrarlos) durante el curso, y se explica porqué la definición de la esperanza con la integral de Lebesgue es equivalente a la que utiliza la integral de Riemman-Stieltjes.

Agradecimientos

Aún a riesgo de olvidarme de alguien, no quiero dejar de agradecer a todos los que de alguna manera me ayudaron a dar la materia y a redactar este apunte.

- A N. Fava y V. Yohai (con quienes en su momento cursé esta materia, dado que mis cursos estuvieron inspirados en gran parte en lo que aprendí de ellos)
- A G. Boente Boente (quien generosamente me prestó el material de sus clases, a M. A. García Álvarez (por regalarme su excelente libro), y a todos los colegas que me hicieron comentarios críticos sobre estas notas.
- A todos los que fueron mis ayudantes en las diferentes cursadas:
  - En 2006: Marcela Svarc, Julieta Molina, Analía Ferrari.
  - En 2010: Daniela Rodriguez, Julián Martinez, Alejandro Lugea.
  - En 2016: Analía Ferrari, Julieta Molina, Florencia Statti.
  - En 2021: Agustín Damonte, Octavio Duarte.
  - En 2022: Mariano Merzbacher, Nicolas Igolnikov, Eugenia Belén.
- A todos mis estudiantes, quienes en muchas veces han aportado correcciones u observaciones que han contribuido a mejorar este apunte.

Pablo L. De Nápoli

Índice general

Capítulo 1

El Espacio Muestral

1.1. Experimentos Aleatorios

La teoría de probabilidades trata con experimentos aleatorios, es decir con experimentos cuyo resultado no resulta posible prever de antemano. Denominamos **espacio muestral** al conjunto de los posibles resultados de un experimento aleatorio, y lo simbolizamos con la letra Ω.

Históricamente, la teoría de probabilidades se desarrolló para estudiar los juegos de azar, pero posteriormente encontró otras innumerables aplicaciones. En estos casos el espacio muestral es usualmente finito:

Ejemplos de experimentos aleatorios:

- Se arroja una moneda. Hay dos resultados posibles:

$$\Omega = \{cara, ceca\}$$

- Se arroja un dado. Hay seis resultados posibles:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Sin embargo, en otras aplicaciones del cálculo de probabilidades, aparecen espacios muestrales de naturaleza más compleja. Veamos algunos ejemplos:

- Se elije un individuo al azar de una población humana y se mide su altura. El resultado es un número real positivo (dentro de un cierto rango). Podemos por lo tanto pensar que el espacio muestral Ω es un intervalo de la recta real.

- Se observa la trayectoria de una partícula que se mueve sobre la superficie de un líquido siguiendo una trayectoria de apariencia caótica durante un cierto intervalo

de tiempo $[0, T]$ (movimiento Browniano). En este caso, cada posible resultado del experimento es una curva continua. Por ello el espacio muestral podría tomarse como el espacio de funciones continuas $C([0, T], \mathbb{R}^2)$.

Un **evento** o **suceso** es algo que puede ocurrir o no ocurrir en cada realización del experimento aleatorio. Los eventos corresponden a subconjuntos del espacio muestral. Por ejemplo: si el experimento consiste en arrojar un dado, el evento "sale un número par" está representado por el subconjunto $A = \{2, 4, 6\}$ del espacio muestral.

Las operaciones booleanas con los conjuntos tienen una interpretación natural en este contexto. Recordamos cuáles son estas operaciones y su significado:

- La intersección $A \cap B$ representa el evento que ocurre si ocurre A y también B.
- La unión $A \cup B$ representa el evento que ocurre si ocurre A o ocurre B (pero pueden ocurrir ambos simultáneamente, es un "o" no exclusivo).
- El complemento A^c de un evento, representa el evento que ocurre si no ocurre A.
- La diferencia de conjuntos $A - B = A \cap B^c$ representa el evento que ocurre si ocurre A pero no ocurre B.
- La diferencia simétrica $A \Delta B$ representa el evento que ocurre si ocurre A o ocurre B pero no ocurren ambos simultáneamente (Es un "o" exclusivo). Recordamos que:

$$A \Delta B = (A \cup B) - (A \cap B) = (A - B) \cup (B - A)$$

También notamos que la condición de que dos eventos A y B sean disjuntos ($A \cap B = \emptyset$) significa que ambos eventos no pueden ocurrir simultáneamente.

1.2. La definición clásica de Laplace

La idea básica del cálculo de probabilidades será asignar a cada evento $A \subset \Omega$, un número real entre 0 y 1 que llamaremos su probabilidad y simbolizaremos por $P(A)$. Este número medirá qué tan probable es que ocurra el evento A.

El matemático francés Pierre-Simon Laplace (1749–1827) propuso la siguiente definición del concepto de probabilidad: consideremos un experimento aleatorio que tiene un número finito de resultados posibles

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

y supongamos que dichos resultados son *equiprobables* (es decir que consideramos que cada uno de ellos tiene las mismas chances de ocurrir o no que los demás), entonces la probabilidad de un evento $A \subset \Omega$ se define por

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

Por ejemplo, supongamos que nos preguntamos ¿cuál es la probabilidad de obtener un número par al arrojar un dado?. En este caso hay 6 casos posibles, que corresponden a los elementos del espacio muestral

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

y 3 casos favorable, que corresponden a los elementos del evento

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

Si suponemos que el dado no está cargado (de modo que asumimos que los seis resultados posibles del experimento son equiprobables), entonces

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

¿Cuál es el significado intuitivo de esta probabilidad?. Intuitivamente, esperamos que si repetimos el experimento muchas veces, observemos que aproximadamente la mitad de las veces sale un número par (y la otra mitad de las veces sale un número impar).

Notemos algunas propiedades de la noción de probabilidad, introducida por la definición de Laplace:

1. La probabilidad de un evento es un número real entre 0 y 1.

   $$0 \leq P(A) \leq 1$$

2. La probabilidad de un evento imposible es 0:

   $$P(\emptyset) = 0$$

   mientras que la probabilidad de un evento que ocurre siempre es 1:

   $$P(\Omega) = 1$$

   Por ejemplo; al tirar un dado, la probabilidad de sacar un 7 es cero mientras que la probabilidad de sacar un número menor que 10 es uno (Los eventos imposibles corresponden como conjuntos al conjunto vacío, y los que ocurren siempre corresponden a todo el espacio muestral Ω).

   Notemos que para el concepto de probabilidad introducido por la definición clásica de Laplace, es cierta la recíproca de esta afirmación: si $P(A) = 0$, el suceso A es

imposible, mientras que si $P(A) = 1$ el suceso ocurre siempre. Sin embargo, esto no será cierto para otras extensiones del concepto de probabilidad que introduciremos más adelante.

3. Si A y B son dos eventos que no pueden ocurrir simultáneamente, entonces la probabilidad de que ocurra A u ocurra B (lo que corresponde como conjunto a $A \cup B$), es cero

$$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$$

Observación 1.2.1

Notemos que el aparente azar en este ejemplo del dado, se debe en realidad a nuestra ignorancia. Porque la mecánica clásica (Newtoniana) nos dice que el movimiento del dado es en realidad un proceso completamente determinístico (no aleatorio). Y si conociéramos la posición y velocidad iniciales y las fuerzas que actúan, podríamos calcular (en principio) en forma exacta, cómo se va a mover el dado.

Un tipo diferente de azar, mucho más fundamental, aparece en la mecánica cuántica, una de las teorías fundamentales de la física moderna. Esta teoría postula que el azar es un componente esencial e irreductible de la naturaleza a nivel microscópico. Así por ejemplo, no podemos predecir con exactitud donde vamos a encontrar un electrón, sino solamente calcular la probabilidad de que el electrón esté en una cierta región del espacio.

El uso de la definición de Laplace en ejemplos concretos requiere frecuentemente contar los elementos de un conjunto. Para ello son fundamentales las nociones de *combinatoria* que se ven en la materia Álgebra I. Pueden encontrar un resumen en el apéndice A. También pueden consultar [Wil04].

1.3. La interpretación frecuencial de la probabilidad

Supongamos que tenemos un evento A en un espacio muestral Ω y que tiene una cierta probabilidad $p = P(A)$.

Repetimos nuestro experimento aleatorio muchas veces, y designamos por f_n a la frecuencia relativa de éxitos en las primeras n realizaciones de nuestro experimento. Es decir:

$$f_n = \frac{\text{número de éxitos en las primeras n repeticiones}}{n}$$

Intuitivamente, esperamos que f_n aproxime a la probabilidad $P(A)$ cuando n es grande. Matemáticamente nos gustaría poder decir que

$$f_n \to p \text{ cuando } n \to \infty$$

en algún sentido.

Este enunciado se conoce como **ley de los grandes números** y toda la teoría del cálculo de probabilidades surgió del intento de formalizarlo como un teorema matemático (como veremos más adelante).

Observación 1.3.1 *(experimentos compuestos) Supongamos que tenemos dos experimentos aleatorios, a los que corresponden los espacios muestrales Ω_1 y Ω_2. ¿Qué espacio muestral corresponderá al experimento compuesto donde realizamos primero un experimento y después el otro?. Será el* producto cartesiano

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

donde ω_1 corresponderá al resultado del primer experimento y ω_2 al del segundo.

Veamos un ejemplo: Supongamos que tenemos una moneda equilibrada y la arrojamos n veces en forma sucesiva. Podemos asociarle a este experimento el espacio muestral

$$\Omega_n = \{(\omega_1, \omega_2, \ldots, \omega_n)\}$$

donde

$$\omega_i = \begin{cases} 0 & \text{si en la } i\text{-ésima tirada de la modena sale ceca} \\ 1 & \text{si en la } i\text{-ésima tirada de la modena sale cara} \end{cases}$$

Como Ω tiene 2^n elementos que consideramos equiprobables, de acuerdo con la defición de Laplace

$$P(\{\omega\}) = \frac{1}{2^n}$$

Sea S_n el número de veces en que sale cara en n tiradas de la moneda. S_n es nuestro primer ejemplo de una **variable aleatoria**, esto es: un número que depende del resultado de un experimento aleatorio. Matemáticamente, S_n es una función $S_n : \Omega \to \mathbb{N}_0$, dada por

$$S_n(\omega) = \omega_1 + \omega_2 + \ldots + \omega_n$$

Nos podemos preguntar ¿cuál es la probabilidad de que S_n tome un determinado valor k? Notamos que que para que $S_n = k$ debemos elegir k lugares ω_k entre los n donde pondremos un 1, y en los restantes $n - k$ lugares pondremos un cero. Luego, de acuerdo a la definición de Laplace:

$$P\{S_n = k\} = \frac{\binom{n}{k}}{2^n}$$

Este es un ejemplo de cómo especificar la **distribución de probabilidades** de la variable S_n. Volveremos sobre estos conceptos más adelante, en el siguiente capítulo.

Sea $f_n = \frac{S_n}{n}$ la frecuencia relativa de caras. Entonces, esperamos que en algún sentido

$$\frac{S_n}{n} \to \frac{1}{2}$$

1.4. Definición axiomática de la probabilidad (provisional)

La definición clásica de Laplace, aunque tiene un claro significado intuitivo presenta algunas limitaciones. En primer lugar, su aplicación está limitada a problemas donde el espacio muestral es finito. Sin embargo como hemos mencionado al comienzo, en muchas aplicaciones importantes del cálculo de probabilidades, nos encontramos con espacios muestrales que no lo son.

Por otra parte, la definición clásica de Laplace hace la suposición de que los posibles resultados del experimento aleatorio (los puntos del espacio muestral) son equiprobables, pero es fácil imaginar experimentos en los que esta suposición no se verifica, por ejemplo si arrojamos un dado que no está equilibrado ("está cargado").

Por los motivos expresados, será conveniente generalizar la noción de probabilidad. Por ello, introduciremos la siguiente definición axiomática (provisional).

Definición 1.4.1 *Sea Ω un espacio muestral, por una probabilidad definida en Ω entenderemos una función P que a cada parte de Ω (evento) le asigna un número real de modo que se cumplen las propiedades enunciadas en la sección anterior:*

1. *La probabilidad de un evento A es un número real entre 0 y 1:*

$$0 \leq P(A) \leq 1$$

2. *La probabilidad del evento imposible es 0:*

$$P(\emptyset) = 0$$

mientras que la probabilidad de un evento que ocurre siempre es 1:

$$P(\Omega) = 1$$

3. *La probabilidad es finitamente aditiva:*

$$A \cap B = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$$

Más adelante, nos veremos obligados a modificar esta definición, ya que en muchos ejemplos no es posible asignar probabilidades a todas las posibles partes de Ω (por lo que deberemos restringir la noción de evento).

Veamos algunos ejemplos:

Supongamos que tenemos un espacio muestral finito

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

pero que no queremos asumir que los posibles resultados de nuestro experimento aleatorio son equiprobables. Entonces supondremos que cada uno de ellos tiene una probabilidad $p_i \in [0,1]$:

$$P(\{\omega_i\}) = p_i$$

Entonces dado un evento $A \subset \Omega$, le asignamos la probabilidad

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

Si suponemos que

$$\sum_{i=1}^{n} p_i = 1$$

entonces la probabilidad así definida, verifica los axiomas de nuestra definición axiomática de probabilidad.

Notemos que en particular, si los resultados r_i $(1 \leq i \leq n)$ son equiprobables:

$$p_1 = p_2 = \ldots = p_n$$

entonces $p_i = \frac{1}{n}$ para todo i, y recuperamos la definición clásica de Laplace:

$$P(A) = \frac{\#(A)}{n}$$

El ejemplo anterior, fácilmente puede generalizarse al caso de un espacio muestral numerable

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$$

Nuevamente supongamos que a cada resultado ω_i (con $i \in \mathbb{N}$) le hemos asignado una probabilidad $p_i \in [0,1]$, de modo que

$$\sum_{i=1}^{\infty} p_i = 1$$

entonces si definimos

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

obtenemos una probabilidad definida en Ω.

Es importante notar, que para esta nueva noción de probabilidad que hemos definido ya no se verifica en general que $P(A) = 0$ implique que A sea un evento imposible, o que si $P(A) = 1$ entonces A es un evento que ocurre siempre (porque algunos p_i podrían ser cero).

Veamos algunas consecuencias de estas definiciones:

Proposición 1.4.2 *Si A y B son dos eventos y $A \subset B$ entonces*

$$P(B - A) = P(B) - P(A).$$

En particular, la probabilidad es creciente:

$$A \subset B \Rightarrow P(A) \leq P(B)$$

Prueba: Como $A \subset B$,

$$B = A \cup (B - A) \text{ unión disjunta}$$

luego

$$P(B) = P(A) + P(B - A)$$

de donde, despejando $P(B - A)$ obtenemos el resultado. □

En particular, elijiendo $B = \Omega$ obtenemos

Corolario 1.4.3 *Si A es un evento y $A^c = \Omega - A$ su complemento, entonces*

$$P(A^c) = 1 - P(A)$$

Proposición 1.4.4 *Si A y B son dos eventos, entonces*

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

En particular, la probabilidad es subaditiva:

$$P(A \cup B) \leq P(A) + P(B)$$

Prueba:

$$A \cup B = (A - A \cap B) \cup (A \cap B) \cup (B - A \cap B) \text{ (unión disjunta)}$$

luego

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

□

Ejemplo 1.4.5 *Supongamos que arrojamos dos dados y queremos calcular la probabilidad de que nos salga al menos un 6. Consideramos $A =$"sale un 6 al arrojar el primer dado´´ y $B =$"sale un 6 al arrojar el segundo dado´´. Entonces la probabilidad buscada es*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36} = 0{,}30555\ldots$$

Este resultado se puede generalizar para una unión de n eventos:

Proposición 1.4.6 (Fórmula de inclusiones y exclusiones) *Sean $A_1, A_2, \dots A_n$ eventos. Entonces*

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \left\{ \sum P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) \right\}$$

donde para cada k, la segunda suma recorre las $\binom{n}{k}$ formas de elegir k conjuntos entre los (A_i).

Observación: Hay entonces

$$\sum_{k=1}^{n} \binom{n}{k} = 2^n - 1$$

términos en total en la suma del segundo miembro.

La demostración se hace por inducción en n. Del mismo modo tenemos:

Proposición 1.4.7 (Subaditividad Finita) *Sean $A_1, A_2, \dots A_n$ eventos, entonces:*

$$P\left(\bigcup_{k=1}^{n} A_k \right) \leq \sum_{k=1}^{n} P(A_k)$$

1.5. El marco de Kolmogorov

Como hemos dicho, en muchas situaciones importantes, no es posible asignar probabilidades a todos los subconjuntos del espacio muestral.

El ejemplo más sencillo de esta situación es el siguiente: supongamos que realizamos el experimento de elegir un número real del intervalo $[0, 1]$ con "distribución uniforme". Con esto queremos decir que si $I \subset [0, 1]$ es un intervalo, queremos que:

$$P(I) = |I| \tag{1.1}$$

donde I designa la longitud del intervalo I.

Un experimento equivalente es el siguiente (ruleta continua): imaginemos que tenemos una rueda y la hacemos girar. Nos interesa medir cual es la posición de la rueda. Dado que esta está determinada por un ángulo $\theta \in [0, 2\pi)$ respecto de la posición inicial, podemos pensar este experimento como elegir un número al azar en el intervalo $[0, 2\pi)$. La distribución uniforme, corresponde a postular que todas las posiciones finales de la rueda son igualmente probables.

Se demuestra en análisis real que no es posible definir una medida (probabilidad) σ-aditiva, que esté definida para todos los posibles subconjuntos del intervalo $[0,1]$ de modo que se verifique la relación (1.1) para cada subintervalo $I \subset [0,1]$.

Lebesgue propuso la siguiente solución a este problema: restringir la clase de los conjuntos a los que asignaremos medida (probabilidad) a lo que se llama una σ-álgebra.

Definición 1.5.1 *Sea Ω un conjunto (espacio muestral). Una σ-álgebra $\mathcal{E}$ de partes de Ω, es una colección de partes de Ω con las siguientes propiedades:*

1. *$\emptyset \in \mathcal{E}$.*
2. *Si A está en $\mathcal{E}$, entonces su complemento $A^c = \Omega - A \in \mathcal{E}$.*
3. *Si $(A_n)_{n\in\mathbb{N}}$ es una familia numerable de conjuntos de $\mathcal{E}$ entonces $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{E}$.*

Obviamente, el conjunto de todas las partes de Ω, $\mathcal{P}(\Omega)$ es una σ-álgebra, pero existen σ-álgebras más pequeñas.

Ejemplo: Si Ω es un conjunto no numerable, por ejemplo $\Omega = \mathbb{R}$ entonces

$$\mathcal{E} = \{A \subset \Omega : A \text{ es numerable o } A^c \text{ es numerable}\}$$

es una σ-álgebra más pequeña que $\mathcal{P}(\Omega)$.

Algunas observaciones importantes:
Si $\mathcal{E}$ es una σ-álgebra de partes de Ω, entonces

1. $\Omega \in \mathcal{E}$.
2. Si $(A_n)_{n\in\mathbb{N}}$ es una familia numerable de subconjuntos de Ω entonces $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{E}$.
   Prueba: por la fórmula de De Morgan
   $$\bigcap_{n\in\mathbb{N}} A_n = \left(\bigcup_{n\in\mathbb{N}} A_n^c\right)^c$$
3. Si $A, B \in \mathcal{E}$ entonces $A - B \in \mathcal{E}$.

Definición 1.5.2 *Observemos que la intersección de una familia cualquiera de σ-álgebras de partes de Ω, también es una σ-álgebra. Deducimos que para cualquier $\mathcal{C} \subset \mathcal{P}(\Omega)$, existe una menor σ-álgebra $\sigma(\mathcal{C})$ que la contiene. Dicha σ-álgebra se denomina la σ-álgebra generada por $\mathcal{A}$.*

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{A}} \{\mathcal{C} \subset \mathcal{A} \subset \mathcal{P}(\Omega) : \mathcal{A} \text{ es una sigma álgebra }\}$$

Definimos la σ-álgebra de Borel de $\mathbb{R}$, como la σ-álgebra generada por los intervalos abiertos de $\mathbb{R}$. **Notación:** $\mathcal{B}(\mathbb{R})$

Definición 1.5.3 *Sean Ω un conjunto y $\mathcal{E} \subset \mathcal{P}(\Omega)$. Una medida sobre $\mathcal{E}$ es una función $\mu : \mathcal{E} \to [0, +\infty]$. con las siguientes propiedades:*

1. $$\mu(\emptyset) = 0$$

2. *Si $(A_n)_{n\in\mathbb{N}}$ es una familia disjunta numerable de conjuntos de $\mathcal{E}$, entonces:*

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

Si además se verifica que $\mu(\Omega) = 1$, μ se denomina una medida de probabilidad sobre Ω.

El matemático ruso Andréi Kolmogórov propuso en 1931 el siguiente marco para la teoría moderna de probabilidades, en el que vamos a trabajar:

Definición 1.5.4 *Un espacio de probabilidad es una terna $(\Omega, \mathcal{E}, P)$ donde Ω es un conjunto (espacio muestral), $\mathcal{E}$ es una σ-álgebra de partes de Ω (la σ-álgebra de los eventos) y P es una medida de probabilidad sobre $\mathcal{E}$.*

El siguiente es un resultado fundamental de análisis real:

Teorema 1.5.5 *(Existencia de la medida de Lebesgue) Existen una única σ-álgebra $\mathcal{M}$ de partes de $\mathbb{R}$ y una única medida $m : \mathcal{M} \to [0, +\infty)$ con las siguientes propiedades:*

1. *$\mathcal{M}$ contiene a los intervalos abiertos (por lo tanto $\mathcal{M}$ contiene a la σ-álgebra de Borel).*

2. *$m(I) = |I|$ para cualquier intervalo de la recta.*

3. *Para cualquier conjunto $A \in \mathcal{M}$, la medida de A es el supremo de las medidas de los compactos contenidos en A:*

$$m(A) = \sup\{m(K) : K\, compacto, K \subset A\}$$

y es el ínfimo de las medidas de los abiertos que contienen a A:

$$m(A) = \text{ínf}\{m(U) : U\, abierto, U \supset A\}$$

(Se dice que la medida m es regular).

4. *La medida m es invariante por traslaciones:*

$$m(A + x) = m(A)\ \forall A \in \mathcal{M}$$

5. *Si $A \in \mathcal{M}$, $m(A) = 0$ y $B \subset A$; entonces $B \in \mathcal{M}$ y $m(B) = 0$. (Se dice que la σ-álgebra de Lebesgue es completa).*

$\mathcal{M}$ se denomina la σ-álgebra de Lebesgue y m se denomina la medida de Lebesgue. Los conjuntos de la σ-álgebra $\mathcal{M}$ se denominan conjuntos medibles Lebesgue.

Corolario 1.5.6 *Si consideramos la restricción de la medida de Lebesgue y de la σ-álgebra de Lebesgue al intervalo $[0, 1]$, entonces obtenemos un espacio de probabilidad.*

1.5.1. Consecuencias de la σ-aditividad

En lo sucesivo, trabajaremos en un espacio de probabilidad $(\Omega, \mathcal{E}, P)$

Proposición 1.5.7 (uniones crecientes) *Si tenemos una sucesión infinita creciente de eventos*

$$A_1 \subset A_2 \subset \ldots \subset A_k \subset A_{k+1} \subset \ldots$$

entonces

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{k \to +\infty} P(A_k)$$

Prueba: Utilizamos el truco de disjuntar los eventos y notamos que como son crecientes:

$$C_k = A_k - \bigcup_{j=1}^{k-1} A_j = A_k - A_{k-1} \quad \text{poniendo } A_0 = \emptyset$$

Ahora los C_k son disjuntos. Entonces por la σ-aditividad

$$\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} C_k \Rightarrow P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(C_k) = \sum_{k=1}^{\infty} [P(A_k) - P(A_{k-1})]$$

Miremos una suma parcial: ¡es una serie telescópica!

$$\sum_{k=1}^{n} [P(A_k) - P(A_{k-1})] = P(A_n) - P(A_0) = P(A_n)$$

deducimos que

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{n \to \infty} P(A_n)$$

□

Proposición 1.5.8 (intersecciones decrecientes) *Si tenemos una sucesión infinita decreciente de eventos*

$$B_1 \supset B_2 \supset \dots B_k \supset B_{k+1} \supset \dots$$

entonces

$$P\left(\bigcap_{k=1}^{\infty} B_k\right) = \lim_{k\to+\infty} P(B_k)$$

Prueba: Tomamos complemento $A_k = B_k^c$. Entonces si los B_k eran decrecientes, los (A_k) serán crecientes. Y observamos que por las leyes de De Morgan

$$\bigcap_{k\in\mathbb{N}} B_k = \left(\bigcup_{k\in\mathbb{N}} A_k\right)^c$$

Luego:

$$\begin{aligned} P\left(\bigcap_{k=1}^{\infty} B_k\right) &= 1 - P\left(\bigcup_{k=1}^{\infty} A_k\right) \\ &= 1 - \lim_{k\to+\infty} P(A_k) \\ &= \lim_{k\to+\infty} [1 - P(A_k)] \\ &= \lim_{k\to+\infty} P(B_k) \end{aligned}$$

□

Proposición 1.5.9 (σ-subaditividad) *Si $(A_k)_{n\in\mathbb{N}}$ es una familia numerable de eventos*

$$P\left(\bigcup_{k\in\mathbb{N}} A_k\right) \leq \sum_{k=1}^{\infty} P(A_k).$$

En particular,

$$si\ P(A_k) = 0\ para\ todo\ k \Rightarrow P\left(\bigcup_{k\in\mathbb{N}} A_k\right) = 0.$$

Tomando complemento,

$$si\ P(A_k) = 1\ para\ todo\ k \Rightarrow P\left(\bigcap_{k\in\mathbb{N}} A_k\right) = 1.$$

Prueba: Pongamos

$$D_n = \bigcup_{k=1}^{n} A_k$$

Por la subaditividad finita

$$P(D_n) \leq \sum_{k=1}^{n} P(A_k) \leq \sum_{k=1}^{\infty} P(A_k)$$

Pero los D_n son crecientes y

$$\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} D_k.$$

Luego al tomar límite cuando $n \to +\infty$ deducimos que

$$P\left(\bigcup_{k\in\mathbb{N}} A_k\right) = \lim_{k\to\infty} P(D_k) \leq \sum_{k=1}^{\infty} P(A_k).$$

□

Nota: Volveremos más adelante (en el lema de Borel-Cantelli 8.4.1) a analizar con mayor profundidad las consecuencias probabilísticas de la hipótesis de σ-aditividad.

Capítulo 2

Probabilidad Condicional e Independencia

2.1. Probabilidad Condicional

En muchas situaciones tendremos que estimar la probabilidad de un evento pero disponemos de alguna información adicional sobre su resultado.

Por ejemplo supongamos que arrojamos un dado (equilibrado) y nos preguntamos ¿Qué probabilidad le asignaríamos a sacar un dos, si supiéramos de antemano que el resultado será un número par?. Para formalizar esta pregunta consideramos en el espacio muestral

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

los eventos

$$A = \text{sale un } 2 \ = \{2\}$$

$$B = \text{sale un número par } = \{2, 4, 6\}$$

Entonces vamos a definir la probabilidad condicional de que ocurra el evento A sabiendo que ocurre el evento B que notaremos $P(A/B)$.

Si estamos en una situación como la anterior donde la definición clásica de Laplace se aplica podemos pensarlo del siguiente modo: los resultados posibles de nuestro experimento son ahora sólo los elementos de B (es decir: hemos restringido nuestro espacio muestral a B), mientras que los casos favorables son ahora los elementos de $A \cap B$ luego

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

Si dividimos numerador y denominador por $\#(\Omega)$, tenemos:

$$P(A/B) = \frac{\frac{\#(A\cap B)}{\#(\Omega)}}{\frac{\#(B)}{\#(\Omega)}} = \frac{P(A\cap B)}{P(B)}$$

Aunque hemos deducido esta fórmula de la definición clásica de Laplace, la misma tiene sentido en general siempre que $P(B) > 0$. Adoptamos pues la siguiente definición:

Definición 2.1.1 *La probabilidad condicional $P(A/B)$ de un evento A suponiendo que ocurre el evento B se define por:*

$$P(A/B) = \frac{P(A\cap B)}{P(B)} \tag{2.1}$$

siempre que $P(B) > 0$.

Otra manera de comprender esta definición es la siguiente: para definir la probabilidad condicional $P(A/B)$ queremos reasignar probabilidades a los eventos $A \subset \Omega$ de modo que se cumplan tres condiciones:

1. La función $A \mapsto P(A/B)$ debe ser una probabilidad (o sea satisfacer los requisitos de nuestra definición axiomática).
2. $P(A\cap B/B) = P(A/B)$ (Esta fórmula dice que la probabilidad condicional de que ocurran los eventos A y B simultaneamente sabiendo que ocurre B debe ser igual a la probabilidad condicional de A sabiendo que ocurre B).
3. Si $A \subset B$ la probabilidad condicional $P(A/B)$ debe ser proporcional a la probabilidad de A de modo que
$$P(A/B) = kP(A) \text{ si } A \subset B$$
siendo k una constante de proporcionalidad fija.

Entonces a partir de estas dos condiciones tenemos:

$$P(A/B) = P(A\cap B/B) = kP(A\cap B)$$

y como queremos que $P(A/B)$ sea una probabilidad debe ser $P(\Omega/A) = 1$, luego

$$1 = kP(\Omega\cap B) = kP(B)$$

con lo que:

$$k = \frac{1}{P(B)}$$

y vemos que la definición (2.1) es la única que satisface estas condiciones.

2.1.1. Fórmula de la probabilidad total

Si ahora consideramos una partición del espacio muestral Ω en eventos disjuntos $B_1, B_2, \dots B_k, ...$ en una cantidad finita o infinita numerable de eventos,

$$\Omega = \bigcup_{k \in I} B_k$$

con $P(B_k) > 0$ para todo k, tenemos que:

$$P(A) = \sum_{k \in I} P(A \cap B_k)$$

por la σ-aditividad de la probabilidad, y como

$$P(A \cap B_k) = P(B_k)P(A/B_k)$$

en virtud de la definición de probabilidad condicional, deducimos la siguiente fórmula:

$$P(A) = \sum_{k \in I} P(B_k)P(A/B_k) \tag{2.2}$$

Ejemplo: Supongamos que realizamos el siguiente experimento compuesto. Primero arrojamos una moneda equilibrada. Luego, si sale cara arrojamos un dado, pero si sale seca arrojamos dos dados. ¿cuál es la probabilidad de obtener al menos un 6 al arrojar los dados?

Llamamos B_1= "sale cara´´, y $B_2 = B_1^c$ = "sale ceca´´. A = "sale al menos un 6´´. Entonces, según la fórmula de la probabilidad total, y teniendo en cuenta el resultado del ejemplo 1.4.5, tenemos que

$$P(A) = P(A/B_1) \cdot P(B_1) + P(A/B_2) \cdot P(B_2) = \frac{1}{6} \cdot \frac{1}{2} + \frac{11}{36} \cdot \frac{1}{2} = \frac{17}{72} = 0{,}2361111\dots$$

2.2. Independencia

Definición 2.2.1 *Decimos que el evento A es independiente del evento B con $P(B) > 0$ si*

$$P(A/B) = P(A)$$

Intuitivamente este concepto significa que; saber si el evento B ocurre o no, no nos dará una mejor estimación de la probabilidad de que ocurra el evento A.

Teniendo en cuenta la definición de la probabilidad condicional, vemos que la condición para que el evento A sea independiente de B es que:

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

Esta manera de escribir la definición tiene dos ventajas: se ve que tiene sentido aún si $P(B) = 0$, y muestra que los roles de los eventos A y B son simétricos. Reescribimos pues la definición en la siguiente forma:

Definición 2.2.2 *Decimos que los eventos A y B son (estocásticamente) independientes si*

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

Ejemplo 2.2.3 *Consideramos el experimento de extraer una carta de un mazo de 48 cartas españolas, y consideramos los eventos:*

- *$A =$ "sale un 1".*
- *$B =$ 'sale una carta de espadas"*

$$P(A) = \frac{4}{48} = \frac{1}{12}, \quad P(B) = \frac{12}{48} = \frac{1}{4}$$

Entonces $A \cap B$ es "sale el uno de espadas" y

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

Luego A y B son independientes.

2.2.1. Una aplicación a la ecología

Supongamos que tenemos una población de animales en un territorio y queremos **estimar** cuántos animales hay. Un método posible es el de **captura / recaptura**. Se utiliza mucho para poblaciones de micro mamíferos y reptiles. Mediante trampas se capturan individuos que son marcados y devueltos a su ambiente. Después de un cierto período de tiempo, suficiente para que los marcados se mezclen con el resto de la población, se realiza una nueva captura

Nuestro espacio muestral Ω serán los individuos de la población. Consideramos, para un individuo elegido al azar, los eventos:
$A =$ "el animal es capturado en la primera captura (y marcado)."
$B =$ "el animal es capturado en la segunda captura."
$C = A \cap B =$ "el animal es capturado en la segunda captura y estaba marcado."

Llamemos n_A al número de inviduos capturados en la primera captura, n_B a los capturados en la segunda captura, n_C a los capturados en la segunda captura y n_Ω a los capturados en ambas. n_A, n_B y n_C son conocidos. Queremos determinar n_Ω.

Si la población es grande, la probabilidad de un individuo de ser capturado será parecida a la frecuencia observada. Entonces podemos **estimar** las probabilidades

$$P(A) \approx \frac{n_A}{n_\Omega}, P(B) \approx \frac{n_B}{n_\Omega}, P(C) \approx \frac{n_C}{n_\Omega}$$

Ahora como las capturas son **independientes** esperamos que

$$P(C) = P(A) \cdot P(B)$$

por lo que tenemos la igualdad aproximada:

$$\frac{n_C}{n_\Omega} \approx \frac{n_A}{n_\Omega} \cdot \frac{n_B}{n_\Omega}$$

de donde podemos estimar el tamaño de la población como

$$n_\Omega \approx \frac{n_A \cdot n_B}{n_C}$$

2.2.2. Propiedades de la independencia de eventos

Proposición 2.2.4 *Si A y B son eventos independientes, A y B^c también lo son.*

Prueba: Notamos que

$$A \cap B^c = A - B = A - A \cap B$$

Luego como $A \cap B \subset A$, y la hipótesis

$$P(A \cap B^c) = P(A - A \cap B) = P(A) - P(A \cap B) = P(A) - P(A)P(B)$$

Entonces

$$P(A \cap B^c) = P(A) \cdot [1 - P(B)] = P(A) \cdot P(B^c)$$

□

2.2.3. Independencia con tres eventos

Si tenemos tres eventos A,B y C ¿cuándo diremos que son independientes?. No sólo vamos a querer que tengamos independencia de a pares:

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

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

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

Si no que también queremos por ejemplo que

$$P(A/B \cap C) = P(A)$$

Esto significa que queremos que A sea independiente de $B \cap C$ por lo que vamos a pedir que

$$P(A \cap (B \cap C)) = P(A) \cdot P(B \cap C)$$

o sea

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

2.2.4. Generalización a familias arbitrarias de eventos

En general, la definición de independencia es la siguiente:

Definición 2.2.5 *Decimos que una familia cualquiera de eventos $\mathcal{A} \subset \mathcal{E}$ es independiente si*

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_n}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_n})$$

para cualquier elección de una cantidad finita A_{i_1}, ..., A_{i_n} de eventos distintos de la familia $\mathcal{A}$.

Ejercicio 2.2.6 (Ejercicio 9 de la práctica 2) *Si $A_1, \ldots, A_n$ son eventos independientes y $B_1, \ldots, B_n$ son tales que para cada $i = 1, \ldots, n$ se tiene $B_i = A_i$ o $B_i = A_i^c$ entonces los eventos $B_1, \ldots, B_n$ también resultan independientes.*

En este ejercicio, $\mathcal{A} = \{A_1, \ldots, A_n\}$.

2.3. Cadenas de Markov

Consideramos un sistema que puede tener una cantidad finita de estados $\Omega = \{E_1, E_2, \ldots, E_n\}$ y que evoluciona con tiempo discreto $t \in \mathbb{N}_0$. Llamemos X_t al estado del sistema en el tiempo t.

La evolución del sistema estaría descripta por el **espacio muestral**

$$\Omega^\infty = \{X = (X_0, X_1, X_2, \ldots, X_t, \ldots) : X_t \in \Omega \text{ para todo } t \in \mathbb{N}_0\}$$

Suponemos que tenemos una cierta probabilidad de pasar del estado E_i al E_j

$$p_{ij} = P\{X_{t+1} = E_i / X_t = E_j\}$$

y que esta probabilidad es independiente de t (no varía en el tiempo)- Los números $P = (p_{ij})$ forman una matriz $den \times n$ que se denomina **matriz de transición**. Notamos que

$0 \leq p_{ij} \leq 1$ y

$$\sum_{i=1}^{n} p_{ij} = 1$$

(Las columnas de la matriz son vectores de probabilidad). P se dice una **matriz estocástica**.

2.3.1. Un ejemplo de una cadena de Markov

Consideramos un modelo simple de 3 estados para un mercado financiero con 3 estados:

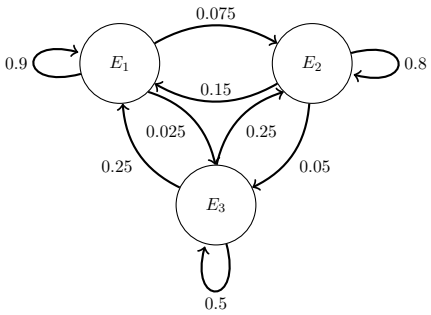
- E_1: Mercado en crecimiento.
- E_2: Mercado en decrecimiento.
- E_3: Mercado estancado.

Supongamos que tenemos la siguiente matriz de transición:

$$P = \begin{bmatrix} 0{,}9 & 0{,}15 & 0{,}25 \\ 0{,}075 & 0{,}8 & 0{,}25 \\ 0{,}025 & 0{,}05 & 0{,}5 \end{bmatrix}$$

(Esto podría aplicarse a todo un mercado o a un activo particular negociado en ese mercado)

Podemos representar esta situación por medio del siguiente grafo:



Pueden encontrar más información sobre este tipo de modelos para mercados financieros en [LJ20].

2.3.2. Propiedades de la matriz de transición

Notamos que por la **fórmula de probabilidad total**:

$$P(X_{t+1} = E_i) = \sum_{j=1}^{n} P(X_{t+1} = E_i / X_t = E_j) \cdot P(X_t = E_j)$$

Esto quiere decir que si consideramos el **vector de probabilidades**

$$U_t = \begin{bmatrix} P(X_t = E_1) \\ P(X_t = E_2) \\ \cdots \\ P(X_t = E_n) \end{bmatrix}$$

tendremos que

$$U_{t+1} = P \cdot U_t \quad \text{para todo } t \in \mathbb{N}$$

Entonces por inducción

$$U_t = P^t U_0$$

2.3.3. Comportamiento a largo plazo

Volvamos a nuestro ejemplo. Diagonalizemos la matriz M. Sus autovalores son:

$$\lambda_1 = 1, \lambda_2 = 0{,}741421356237310, \lambda_3 = 0{,}458578643762690$$

Y podemos encontrar una matriz de cambio de base C tal que

$$C^{-1}PC = D = \begin{pmatrix} 1 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}7414213562373091 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}4585786437626905 \end{pmatrix}$$

$$C = \begin{pmatrix} 0{,}5773502691896265 & 0{,}44371856511363317 & -0{,}03400257431430442 \\ 0{,}5773502691896251 & -0{,}8113070602072252 & -0{,}13017637781608127 \\ 0{,}5773502691896256 & -0{,}3806503501002046 & 0{,}9909076322234505 \end{pmatrix}$$

Luego

$$\lim_{t \to +\infty} P^t = \begin{pmatrix} 0{,}625 & 0{,}625 & 0{,}625 \\ 0{,}3125 & 0{,}3125 & 0{,}3125 \\ 0{,}0625 & 0{,}0625 & 0{,}0625 \end{pmatrix}$$

Entonces la ecuación

$$U_t = P^t U_0$$

nos muestra que no importa cuál sea el estado inicial U_0, el sistema converge al estado estacionario:

$$U_\infty = \begin{pmatrix} 0{,}625 \\ 0{,}3125 \\ 0{,}0625 \end{pmatrix}$$

Esto significaría que el mercado va a estar un $62,5\,\%$ del tiempo en estado alcista, un $32,15\,\%$ en estado bajista y un $6,25\,\%$ estancado. Este vector es un autovector de autovalor 1 de P es decir:

$$PU_\infty = U_\infty = 1 \cdot U_\infty$$

2.3.4. Otros ejemplos de cadenas de Markov

- Juegos de tablero con dados, como el **Monopoly**.
- El experimento de tirar infinitas veces la moneda es una cadena de Markov con dos estados $\{\text{cara}, \text{ceca}\}$. En este caso la matriz de transición es:

$$P = \begin{bmatrix} 0{,}5 & 0{,}5 \\ 0{,}5 & 0{,}5 \end{bmatrix}$$

  Esta matriz verifica $P^2 = P$. Luego $P^t = P$ para todo t, y tenemos que el estado estacionario es

$$U_\infty = \begin{bmatrix} 0{,}5 \\ 0{,}5 \end{bmatrix}$$

- El algoritmo *page rank* usado por Google para asignar a cada página un ranking a cada página web. Este simula una un navegante que va visitando aleatoriamente las páginas, y puede pensarse como una cadena de Markov (ver [KGS13]). La distribución estacionaria de esta cadena de Markov determina el ranking que será asignado a cada página.
- También se utiliza con frecuencia modelos basados en cadenas de Markov para estudiar la propagación de epidemias. Ver por ejemplo el capítulo 2 de [Ige20].

Pueden encontrar más información sobre las cadenas de Markov en el apéndice B.

Capítulo 3

Variables Aleatorias Discretas

3.1. Variables aleatorias discretas

En muchas situaciones, nos interesa un número asociado al resultado de un experimento aleatorio: por ejemplo, el resultado de una medición.

Para evitar por el momento, algunas dificultades técnicas, comenzaremos con el caso de variables aleatorias discretas, que resulta más sencillo de entender.

Definición 3.1.1 *Sea $(\Omega, \mathcal{E}, P)$ un espacio de probabilidad. Una variable aleatoria discreta es una función $X : \Omega \to \mathbb{R}$ tal que la imagen de X es un conjunto finito o numerable de $\mathbb{R}$:*

$$Im(X) = \{x_1, x_2, \ldots, x_i, \ldots\}$$

(donde la sucesión (x_i) puede ser finita o infinita), y tal que $X^{-1}(\{x_i\}) \in \mathcal{E}$ sea un evento para cada $x_i \in Im(X)$.

Como $X^{-1}(\{x_i\}) = \{\omega \in \Omega : X(\omega) = x_i\}$ es un evento para cada i, esto significa que están definidas las probabilidades:

$$p_i = P(\{X = x_i\})$$

Dichas probabilidades se conocen como la *distribución de probabilidades* de la variable X.

Ejemplo 3.1.2 *Tiramos dos dados. Nuestro espacio muestral es:*

$$\Omega = \{\omega = (\omega_1, \omega_2) : \omega_i \in D\}$$

donde $D = \{1, 2, 3, 4, 5, 6\}$. Consideramos la suma S de los puntos obbtenidos

$$S : \Omega \to \mathbb{N}, \quad S(\omega) = \omega_1 + \omega_2$$

*Es una **variable aleatoria discreta**. ¿Cuál es la distribución de probabilidades de S?*

x_i	p_i	
2	$1/36$	$= 0{,}028$
3	$2/36$	$= 0{,}056$
4	$3/36$	$= 0{,}083$
5	$4/36$	$= 0{,}111$
6	$5/36$	$= 0{,}139$
7	$6/36$	$= 0{,}167$
8	$5/36$	$= 0{,}139$
9	$4/36$	$= 0{,}111$
10	$3/36$	$= 0{,}083$
11	$2/36$	$= 0{,}056$
12	$1/36$	$= 0{,}028$

3.2. La Esperanza

Un concepto de fundamental importancia asociado a las variables aleatorias, es el de esperanza (o valor esperado). Para variables aleatorias discretas, este concepto resulta sencillo de definir:

Definición 3.2.1 *Sea $X : \Omega \to \mathbb{R}$ una variable aleatoria discreta. Diremos que X es integrable (o que tiene esperanza finita) si la serie*

$$\sum_i p_i x_i$$

es absolutamente convergente, es decir si:

$$\sum_i p_i |x_i| < +\infty$$

En este caso definimos, la esperanza de X como el valor de dicha suma.

$$E[X] = \sum_i p_i x_i \tag{3.1}$$

Notemos que una variable aleatoria discreta con imagen finita (o sea que tome sólo un número finito de valores) siempre es integrable ya que la suma (3.1) es finita en este caso.

Ejemplo: Supongamos que arrojamos un dado ¿cuál es la esperanza del valor obtenido X ?

$$E[X] = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3,5$$

Ejemplo: Supongamos que jugamos un peso a la ruleta y apostamos a un color (por ej. negro). Sea X nuestra ganancia (o pérdida) ¿cuánto debemos esperar ganar (o perder) ?

Aquí

$$X = \begin{cases} 1 & \text{si sale negro} \quad (\text{con probabilidad } \frac{18}{37}) \\ -1 & \text{si sale rojo o cero} \quad (\text{con probabilidad } \frac{19}{37}) \end{cases}$$

En consecuencia:

$$E[X] = \frac{18}{37} - \frac{19}{37} = \frac{-1}{37} = -0,027\ldots$$

Así pues, al jugar a la ruleta, debemos esperar perder un 27 por mil.

Ejemplo: Sea A un evento, consideramos la función $I_A : \Omega \to \mathbb{R}$ definida por

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

Intuitivamente I_A vale 1 cuando el evento A ocurre, y 0 sino. Se denomina el indicador del evento A. (En la teoría de la medida, esta función se llama la función característica del conjunto A y se suele denotar por χ_A, pero en la teoría de probabilidades la expresión "función característica" tiene un significado diferente).

I_A es una variable aleatoria discreta pues su imagen consta de dos valores (0 y 1) y sus pre-imágenes son $I_A^{-1}(0) = \Omega - A$ y $I_A^{-1}(1) = A$, que son eventos.

La esperanza de I_A es:

$$E[I_A] = 0 \cdot P(\Omega - A) + 1 \cdot P(A) = P(A)$$

Es decir, la esperanza del indicador de un evento, coincide con su probabilidad.

Ejemplo:(un ejemplo de una variable aleatoria que toma infinitos valores). Consideremos el experimento consistente en arrojar infinitas veces una moneda (en forma independiente).

Como vimos anteriormente, podemos modelizar este experimento utilizando el espacio muestral $\Omega = \{0,1\}^{\mathbb{N}}$ de las sucesiones de ceros y unos, y representando cada realización del experimento por la sucesión $\omega = (X_i)_{i\in\mathbb{N}}$ donde

$$X_i = \begin{cases} 1 & \text{si en la i-ésima realización del experimento sale cara} \\ 0 & \text{si en la i-ésima realización del experimento sale ceca} \end{cases}$$

Notemos que las X_i son variables aleatorias. Estamos interesados ahora en la siguiente variable aleatoria, T = cuántas tiradas tengo que esperar hasta que salga una cara por primera vez. Formalmente

$$T(\omega) = \min_{x_i=1} i$$

Hay un caso especial, que es cuando siempre sale ceca, esto es: ¿qué valor de T le asignaremos a la sucesión $\omega = (0, 0, 0, \ldots, 0, \ldots)$? Lo razonable es poner:

$$T((0, 0, 0, \ldots, 0, \ldots)) = +\infty$$

Esto muestra que a veces resulta conveniente admitir variables aleatorias que pueden tomar el valor $+\infty$ (o también $-\infty$).

Ahora debemos calcular cuál es la distribución de probabilidades de T, es decir cuál es la probabilidad de que T tome cada valor.

$$P\{T = k\} = P\{X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1\}$$

y dado que los ensayos son independientes a este evento le asignamos la probabilidad dada por el producto de las probabilidades:

$$P\{T = k\} = P\{X_1 = 0\} \cdot P\{X_2 = 0\} \cdot \ldots \cdot P\{X_{k-1} = 0\} \cdot P\{X_k = 1\} = \frac{1}{2^k}$$

Mientras que al evento "siempre sale ceca" le asignamos probabilidad 0,

$$P\{T = +\infty\} = P\{T((0, 0, 0, \ldots, 0, \ldots)\} = 0$$

Entonces la esperanza de T se calcularía por:

$$E[T] = \sum_{k=1}^{\infty} kP\{T = k\} + (+\infty) \cdot P\{T = +\infty\} = \sum_{k=1}^{\infty} \frac{k}{2^k} + (+\infty) \cdot 0$$

Hacemos la convención de que:

$$0 \cdot (+\infty) = 0$$

Entonces la esperanza de T es:

$$E[T] = \sum_{k=1}^{\infty} \frac{k}{2^k}$$

Utilizando la fórmula,

$$\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2} \text{ si } |x| < 1$$

que se deduce de derivar la serie geométrica, con $x = \frac{1}{2}$, deducimos que $E[T] = 2$.

Así pues, en promedio, habrá que esperar dos tiradas, para que salga cara.

3.2.1. Esperanzas en la computadora

Consideremos una variable aleatoria discreta con 3 valores $1, 2, 3$ tal que

$$P\{X = 1\} = \frac{1}{2}, P\{X = 2\} = P\{X = 3\} = \frac{1}{4}$$

$$E[X] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} = \frac{7}{4} = 1{,}75$$

Veamos cómo se calcularía esto en la computadora usando **Python 3** y el paquete **SciPy**.

```
import scipy.stats
xk = (1, 2, 3)
pk = (0.5, 0.25, 0.25)
distribucion = scipy.stats.rv_discrete(values=(xk, pk))
print(distribucion.mean())
```

3.2.2. Esperanzas infinitas

A veces resulta conveniente admitir esperanzas infinitas. Si $X \geq 0$ diremos que $E[X] = +\infty$ si

$$\sum_i x_i P\{X = x_i\}$$

diverge.

Si X es una variable aleatoria discreta cualquiera, escribimos

$$X = X^+ - X^-$$

donde

$$X^+ = \begin{cases} X & \text{si} \quad X \geq 0 \\ 0 & \text{si} \quad X < 0 \end{cases}$$

y

$$X^- = \begin{cases} -X & \text{si} \quad X < 0 \\ 0 & \text{si} \quad X \geq 0 \end{cases}$$

Notamos que X^+ y X^- son variables aleatorias no negativas.

Decimos que $E[X] = +\infty$ si $E[X^+] = +\infty$ y $E[X^-] < \infty$. Similarmente diremos que $E[X] = -\infty$ si $E[X^-] = +\infty$ y $E[X^+] < \infty$. Si $E[X^+]$ y $E[X^-]$ son ambas infinitas, $E[X]$ no está definida.

3.2.3. Propiedades de la Esperanza

Proposición 3.2.2 (linealidad de la esperanza) *1. Si $X, Y : \Omega \to \mathbb{R}$ son variables aleatorias discretas con esperanza finita, entonces*

$$E[X+Y] = E[X] + E[Y]$$

2. Si $X : \Omega \to \mathbb{R}$ es una variable aleatoria discreta con esperanza finita, entonces:

$$E[\lambda X] = \lambda E[X]$$

Prueba: Sean (x_i) los valores que toma X, e (y_j) los valores que toma Y: entonces

$$E[X] = \sum_i x_i P\{X = x_i\} = \sum_{i,j} x_i P\{X = x_i, Y = y_j\}$$

ya que

$$\{X = x_i\} = \bigcup_j \{X = x_i, Y = y_j\} \text{ (unión disjunta)}$$

y el reordenamiento de la serie está justificado por la convergencia absoluta, de la serie:

$$\sum_{i,j} x_i P\{X = x_i, Y = y_j\}$$

Similarmente,

$$E[Y] = \sum_j y_j P\{X = x_i\} = \sum_{i,j} y_j P\{X = x_i, Y = y_j\}$$

En consecuencia,

$$E[X] + E[Y] = \sum_{i,j} (x_i + y_j) P\{X = x_i, Y = y_j\}$$

Sea $Z = X+Y$ y sean $z_1, z_2, \ldots, z_k, \ldots$ los valores de Z. Entonces los z_k son exactamente los valores $x_i + y_j$ (pero estos últimos pueden repetirse). Entonces,

$$E[Z] = \sum_k z_k P\{Z = z_k\} = \sum_k \sum_{i,j: x_i + y_j = z_k} z_k P\{X = x_i, Y = y_j\}$$

pues

$$\{Z = z_k\} = \bigcup_{i,j: x_i + y_j = z_k} \{X = x_i, Y = y_j\} \text{ (unión disjunta)}$$

Deducimos que

$$E[Z] = \sum_k (x_i + y_j) P\{X = x_i, Y = y_j\} = E[X] + E[Y]$$

Esto completa la prueba de la primera afirmación. En cuanto a la segunda afirmación, λX es una variable aleatoria discreta que toma los valores λx_i, por lo tanto:

$$E[\lambda X] = \sum_i \lambda x_i P\{\lambda X = \lambda x_i\} = \lambda \sum_i x_i P\{X = x_i\} = \lambda E[X]$$

□

Proposición 3.2.3 (Monotonía de la esperanza) *1. Si X es una variable aleatoria con esperanza finita y $X \geq 0$ con probabilidad 1, entonces $E[X] \geq 0$.*

2. *Sean X e Y variables aleatorias con esperanza finita. Entonces, si $X \leq Y$ con probabilidad 1, tenemos que $E[X] \leq E[Y]$*

3. *Si X es una variable aleatoria acotada, entonces:*

$$\inf_\Omega X \leq E[X] \leq \sup_\Omega X$$

.

4. *Si X es una variable aleatoria discreta con esperanza finita, entonces:*

$$|E[X]| \leq E[|X|]$$

Proposición 3.2.4 *Sean X una variable aleatoria discreta y $\varphi : \mathbb{R} \to \mathbb{R}$. Entonces*

$$E[\varphi(X)] = \sum_i \varphi(x_i) P\{X = x_i\}$$

siempre que esta serie sea absolutamente convergente.

Prueba: Sea $Y = \varphi(X)$, y sean (y_j) los valores de Y, entonces:

$$E[Y] = \sum_j y_j P\{Y = y_j\} = \sum_j y_j \sum_{i:\varphi(x_i)=y_j} P\{X = x_i\} = \sum_i \varphi(x_i) P\{X = x_i\}$$

(El reordenamiento se justifica usando la convergencia absoluta de la serie.) □

Esta propiedad se puede generalizar a funciones de **vectores aleatorios**. Este concepto es una generalización natural del de variable aleatoria discreta:

Definición 3.2.5 *Un vector aleatorio discreto n-dimensional es una función $X : \Omega \to \mathbb{R}^n$ tal que $Im(X)$ sea finita o infinita numerable, y $P\{X = x\}$ sea un evento para todo $x \in \mathbb{R}^n$. Dar un vector aleatorio discreto $X = (X_1, X_2, \ldots, X_n)$ es equivalente a dar n variables aleatorias discretas $x_1, x_2, \ldots, x_n$*

Con esta terminología tenemos [con la misma demostración de antes]:

Proposición 3.2.6 *Sean X un vector aleatorio n-dimensional y $\varphi : \mathbb{R}^n \to \mathbb{R}$, entonces*

$$E[\varphi(X)] = \sum_i \varphi(x_i) P\{X = x_i\}$$

donde x_i recorre la imagen de X, siempre que esta serie sea absolutamente convergente.

3.2.4. Independencia

Definición 3.2.7 *Sean X e Y dos variables aleatorias discretas definidas en un mismo espacio muestral. Diremos que son* **independientes**, *si para cada x_i, y_j los eventos $\{X = x_i\}$ e $\{Y = y_j\}$ son independientes, es decir de acuerdo a la definición de eventos independientes si,*

$$P\{X = x_i, Y = y_j\} = P\{X = x_i\} \cdot \{Y = y_j\}$$

Observación: Remarcamos que esta definición solamente se aplica a variables discretas, cuando generalicemos esta noción a variables aleatorias no discretas, nos veremos en la necesidad de adoptar una definición diferente.

Proposición 3.2.8 *Si X e Y son variables aleatorias discretas independientes, y $f, g : \mathbb{R} \to \mathbb{R}$ son funciones, entonces $Z = f(X)$ y $W = g(Y)$ también son variables aleatorias discretas independientes.*

Prueba: Calculemos la distribución conjunta de Z y W:

$$\begin{aligned} P\{Z = z, W = w\} &= \sum_{x,y: f(x)=z, g(y)=w} P\{X = x, Y = y\} \\ &= \sum_{x,y: f(x)=z, g(y)=w} P\{X = x\} P\{Y = y\} \\ &= \left(\sum_{x: f(x)=z} P\{X = x\} \right) \left(\sum_{y: g(y)=w} P\{Y = y\} \right) = P\{Z = z\} P\{W = w\} \end{aligned}$$

□

Proposición 3.2.9 *Si X e Y son variables aleatorias discretas independientes con esperanza finita, entonces:*

$$E(XY) = E(X)E(Y)$$

Prueba:

$$E[XY] = \sum_{i,j} x_i y_j P\{X = x_i, Y = y_j\} = \sum_{i,j} x_i y_j P\{X = x_i\} P\{Y = y_j\}$$

$$\left(\sum_i x_i P\{X = x_i\}\right)\left(\sum_j y_j P\{Y = y_j\}\right) = E[X]E[Y]$$

Observación: En el caso en que X e Y toman infinitos valores, la aplicación de la propiedad distributiva, está justificada por el hecho de que las series que intervienen son absolutamente convergentes, por hipótesis. □

3.2.5. Desigualdad de Jensen

Definición 3.2.10 *Sea $f : \mathbb{R} \to \mathbb{R}$ una función. Diremos que f es convexa, si dados $x, y \in \mathbb{R}$ y $\alpha \in [0, 1]$, se verifica que:*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y)$$

Observación: Si f es de clase C^2, entonces f es convexa, si y sólo si $f''(x) \geq 0$.

Observación: Una función convexa en $\mathbb{R}$ es necesariamente continua. Además es posible probar que su derivada $f'(x)$ existe salvo quizás para un conjunto a lo sumo numerable de valores de x, y que f' es creciente (ver [WZ77], teorema 7.40).

Ejercicio: Una **combinación convexa** de los x_i es una combinación lineal

$$\sum_{i=1}^{n} \alpha_i x_i$$

en la que $0 \leq \alpha_i$ y $\sum_{i=1}^{n} \alpha_i = 1$. Probar que si $f : \mathbb{R} \to \mathbb{R}$ es una función convexa y $\sum_{i=1}^{n} \alpha_i x_i$ es una combinación convexa, entonces:

$$f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \leq \sum_{i=1}^{n} \alpha_i f(x_i)$$

Proposición 3.2.11 (Desigualdad de Jensen) *Si $g : \mathbb{R} \to \mathbb{R}$ es una función convexa, entonces:*

$$g(E[X]) \leq E[g(X)])$$

en los siguientes casos: si X es no negativa y $g(x) \geq 0$ para $x \geq 0$, o si X y g son arbitrarias y $E(|g(X)|) < \infty$.

Prueba: Hagamos la demostración primero, en el caso que X toma sólo finitos valores. Sea $p_i = P\{X = x_i\}$. Entonces

$$E[X] = \sum_{i=1}^{n} p_i x_i$$

es una combinación convexa de los valores de X. Como X es una función convexa,

$$g(E[X]) = g\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i g(x_i) = E[g(X)]$$

Si X toma un número numerable de valores, x_i con probabilidades p_i, entonces hacemos lo siguiente: para cada $n \in \mathbb{N}$ definamos,

$$s_n = \sum_{i=1}^{n} p_i$$

y notamos que

$$\sum_{i=1}^{n} \frac{p_i}{s_n} x_i$$

es una combinación convexa. Entonces, como g es convexa:

$$g\left(\sum_{i=1}^{n} \frac{p_i}{s_n} x_i\right) \leq \sum_{i=1}^{n} \frac{p_i}{s_n} g(x_i)$$

Cuando $n \to +\infty$, tenemos que $s_n \to 1$. Entonces, utilizando la continuidad de g, obtenemos que:

$$g(E[X]) = g\left(\sum_{i=1}^{\infty} p_i x_i\right) \leq \sum_{i=1}^{\infty} p_i g(x_i) = E[g(X)]$$

□

Ejemplo: $f(x) = |x|^p$ es una función convexa si $p \geq 1$. En consecuencia, en este caso:

$$|E[X]|^p \leq E[|X|^p]$$

3.3. Momentos - Varianza

Definición 3.3.1 *Sea X una variable aleatoria (discreta). Definimos el k-ésimo momento de X entorno de b como $E[(X-b)^k]$. El k-ésimo momento absoluto entorno de b se define como $E[|X-b|^k]$.*

Algunas observaciones:

1. Si $E[|X|^t] < \infty$ y $0 \leq s \leq t$, entonces $E[|X|^s] < +\infty$. En efecto según la desigualdad de Jensen,
$$(E[|X|^s])^p \leq E[|X|^t]$$
donde $p = \frac{t}{s} \geq 1$. Es más, vemos que:

2. $E[|X|^p]^{1/p}$ es una función creciente de p.

3. Si $E[|X|^p] < +\infty$ y $E[|Y|^p] < +\infty$ entonces $E[|X+Y|^p]^{1/p} < +\infty$ **Prueba:**
$$|X+Y|^p \leq (|X| + |Y|)^p = (2 \operatorname{máx} |X|, |Y|)^p$$
$$\leq 2^p \operatorname{máx}(|X|^p, |Y|^p) \leq 2^p(|X|^p + |Y|^p)$$
Por lo tanto,
$$E[|X+Y|^p] \leq 2^p(E[|X|^p] + E[|Y|^p]) < +\infty$$
□

4. En consecuencia, el conjunto
$$L_d^p(\Omega, \mathcal{E}, P) = \{X : \Omega \to \overline{\mathbb{R}} \text{ variable aleatoria discreta} : E[|X|^p] < +\infty\}$$
(siendo $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$) es un espacio vectorial.

5. Si $p \geq 1$, es posible probar que
$$\|X\|_p = E[|X|^p]^{\frac{1}{p}}$$
es una norma en dicho espacio.

En lo sucesivo, nos van a interesar especialmente dos clases L^p:

$$L_d^1(\Omega) = \{X : \Omega \to \overline{\mathbb{R}} : \text{ variable aleatoria (discreta) con esperanza finita}\}$$

$$L_d^2(\Omega) = \{X : \Omega \to \overline{\mathbb{R}} : \text{ variable aleatoria (discreta) con segundo momento finito}\}$$

Ejemplo: Notemos que $L_d^2 \subset L_d^1$ por lo anterior. Veamos un ejemplo de una variable aleatoria que está en L_d^1 pero no en L_d^2: Consideramos un espacio muestral numerable
$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\}$$
en el que
$$P\{\omega_n\} = \frac{1}{n(n+1)}$$

Verifiquemos que esta asignación efectivamente define una distribución de probabilidades en Ω:

$$\sum_{n=1}^{\infty} P\{\omega_n\} = \sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \left[\frac{1}{n} - \frac{1}{n+1}\right] = 1$$

(serie telescópica). Definamos la variable aleatoria $X : \Omega \to \mathbb{R}$, dada por $X(\omega_n) = \sqrt{n}$. Entonces,

$$E(X) = \sum_{n=1} X(\omega_n) P\{\omega_n\} = \sum_{n=1}^{\infty} \frac{\sqrt{n}}{n(n+1)} \leq \sum_{n=1}^{\infty} \frac{1}{n^{3/2}} < +\infty$$

pero

$$E(X^2) = \sum_{n=1} X(\omega_n)^2 P\{\omega_n\} = \sum_{n=1}^{\infty} \frac{n}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{n+1} = +\infty$$

Definición 3.3.2 *El segundo momento de X entorno de su media se llama la* **varianza** *(o* **variancia**[1]*) de X, es decir:*

$$Var(X) = E[(X - E(X))^2]$$

Por lo anterior $Var(X) < +\infty$ si y sólo si el segundo momento de X es finito, es decir si $X \in L^2_d$.

Notamos que si X es una variable aleatoria discreta, podemos calcular su varianza usando la fórmula

$$\begin{aligned} Var(X) &= E[\varphi(X)] \\ &= \sum_i (x_i - \mu_X)^2 \cdot p_i \end{aligned}$$

donde $\mu_X = E[X]$ y $\varphi(x) = (x - \mu_X)^2$.

$$\sigma_X = \sqrt{Var(X)}$$

se demomina ***desviación estándar*** *o* ***desviación típica****.*

Ejemplo: Sea A un evento con probabilidad p, e I_A su indicador. Calculemos su varianza. Ya vimos que:

$$E[I_A] = P(A) = p$$

En consecuencia:

$$\mathrm{Var}(I_A) = E[(I_A - p)^2]$$

[1]Según el diccionario de la RAE, ambas grafías son aceptables.

La distribución de probabilidades de $(I_A - p)^2$ es:

$$(I_A - p)^2 = \begin{cases} (1-p)^2 & \text{si ocurre } A \quad \text{(con probabilidad } p\text{)} \\ p^2 & \text{si no ocurre } A \quad \text{(con probabilidad } q = 1-p\text{)} \end{cases}$$

En consecuencia,

$$\text{Var}(I_A) = (1-p)^2 p + p^2(1-p) = p - p^2 = pq$$

Proposición 3.3.3 *1. Si $X = c$ es constante, entonces $Var(X) = 0$.*

2. $Var(aX + b) = a^2\, Var(X)$.

Prueba: Si $X = c$ es constante, $E[X] = c$ luego $Var(X) = E[0] = 0$.

$$E[aX + b] = aE[X] + E[b] = a \cdot E[X] + b$$

$$\begin{aligned} \text{Var}(aX + b) &= E[[(aX + b) - (a \cdot E[X] + b)]^2] \\ &= E[(aX - aE(X))^2] = E[a^2(X - E(X))^2] = a^2 E[(X - E(X))^2] \\ &= a^2 \text{Var}(X) \end{aligned}$$

□

Proposición 3.3.4 (Otra manera de escribir la varianza)

$$Var(X) = E[X^2] - E[X]^2$$

Prueba: Sea $\mu_X = E[X]$

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu_X)^2] \\ &= E[X^2 - 2\mu_X X + \mu^2] \\ &= E[X^2] - 2\mu_X E[X] + E[\mu_X^2] \\ &= E[X^2] - 2\mu_X^2 + \mu_X^2 \\ &= E[X^2] - \mu_X^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

□

3.3.1. Desigualdades de Chebyshev y de Markov

Proposición 3.3.5 (Desigualdad básica) *Sea X una variable aleatoria no negativa, entonces*

$$P(X \geq \lambda) \leq \frac{1}{\lambda} E(X) \tag{3.2}$$

Prueba: Sea $A = \{\omega \in \Omega : X(\omega) \geq \lambda\}$. Entonces $X \geq \lambda I_A$, en consecuencia: $E[X] \geq \lambda E[I_A] = \lambda P(A)$ □

Proposición 3.3.6 (Desigualdad de Markov) *Si X es una variable aleatoria (discreta) entonces*

$$P\{|X| \geq \lambda\} \leq \frac{1}{\lambda^p} E(|X|^p)$$

Prueba: Si cambiamos X por $|X|^p$ en la desigualdad anterior tenemos que:

$$P\{|X| \geq \lambda\} = P\{|X|^p > \lambda^p\} \leq \frac{1}{\lambda^p} E(|X|^p)$$

□

Proposición 3.3.7 (desigualdad de Chebyshevclásica) *Sea X una variable (discreta) entonces*

$$P\{|X - E(X)| > \lambda\} \leq \frac{Var(X)}{\lambda^2}$$

Prueba: Usamos la desigualdad anterior con $p = 2$ y cambiamos X por $X - E(X)$. □

Intuitivamente, la desigualdad de Chebyshevdice que la varianza de la variable X nos da una estimación de la probabilidad de que X tome valores alejados de su esperanza. Si $\text{Var}(X)$ es pequeña, entonces es poco probable que X tome un valor alejado de $E(X)$.

3.3.2. Covarianza

Definición 3.3.8 *Sean X e Y dos variables aleatorias. Definimos la* **convarianza** *o* **covariancia** *de X e Y por*

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y)]$$

Digamos que $Im(X) = \{x_1, x_2, \ldots, x_i, \ldots\}$ e $Im(Y) = \{y_1, y_2, \ldots, y_j, \ldots\}$. y

$$p_{i,j} = P\{X = x_i, Y = y_j\}$$

denota la ***distribución conjunta*** *de X e Y entonces:*

$$Cov(X, Y) = E[\varphi(X, Y)] = \sum_{i,j} p_{i,j}(x_i - \mu_X) \cdot (y_j - \mu_Y)$$

donde

$$\varphi(x, y) = (x - \mu_X) \cdot (y - \mu_Y)$$

.

Observación: Si X e Y son variables aleatorias independientes entonces Cov(X,Y) = 0. La recíproca no es cierta, como muestra el siguiente ejemplo:

Ejemplo (Barry James, pag. 130) Sean X e Y dos variables aleatorias con valores -1, 0, 1 con la siguiente función de probabilidad conjunta:

	-1	0	1
-1	$\frac{1}{5}$	0	$\frac{1}{5}$
0	0	$\frac{1}{5}$	0
1	$\frac{1}{5}$	0	$\frac{1}{5}$

entonces $E[XY] = E[X] = E[Y] = 0$, pero X e Y no son independientes pues

$$P\{X = 0, Y = 0\} = \frac{1}{5} \neq \frac{1}{25} = \frac{1}{5} \cdot \frac{1}{5} = P\{X = 0\}P\{Y = 0\}$$

Definición 3.3.9 *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias discretas. Diremos que no están correlacionadas si $Cov(X_i, X_j) = 0$ para $i \neq j$.*

Observación 3.3.10 *Recordamos que introdujimos el espacio vectorial*

$$L^2_d = \{X : \Omega \to \mathbb{R} \text{ variable aleatoria discreta con } E(|X|^2) < +\infty\}$$

*Sus elementos se llaman **variables aleatorias con segundo momento finito**. Notemos que:*

$$X \in L^2_d \Rightarrow E(X) \text{ y } Var(X) \text{ son finitas.}$$

$$\|X\|_2 = E(|X|^2)^{1/2}$$

*es una norma en este espacio, que proviene del **producto interno***

$$\langle X, Y \rangle = E(X \cdot Y)$$

Entonces la desigualdad de Cauchy-Schwarz (aplicada a $X - \mu_X$ y $Y - \mu_Y$) nos da que

$$|Cov(X, Y)| \leq \sqrt{Var(X)} \cdot \sqrt{Var(Y)} \tag{3.3}$$

*Notemos que decir que X e Y no están correlacionadas equivale a decir que $X - \mu_X$ e $Y - \mu_Y$ son **ortogonales** en L^2_d. Esto nos sugiere considerar el número*

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}$$

*que sería geométricamente el coseno del ángulo entre $X - \mu_X$ y $Y - \mu_Y$, para medir que tan lejos están las variables X e Y de estar correlacionadas. Se denomina **coeficiente de correlación entre X e Y**. La desigualdad (3.3) nos dice que $0 \leq |\rho| \leq 1$, o sea que $-1 \leq \rho \leq 1$.*

Proposición 3.3.11 *Si X e Y son variables aleatorias (discretas) con segundo momento finito:*

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Prueba:

$$Var(X+Y) = E[(X+Y-E[X]-E[Y])^2] = E[((X-E(X)+(Y-E(Y))^2] =$$

$$= E[(X-E(X))^2] + E[(Y-E(Y))^2] + 2E[(X-E(X))(Y-E(Y))] =$$

$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

□

Corolario 3.3.12 *Si $X_1, X_2, \ldots, X_n$ son variables aleatorias (discretas) con segundo momento finito, que no están correlacionadas, entonces*

$$Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} Var(X_i)$$

Dem: Sale de la fórmula anterior por inducción.

3.4. Ensayos de Bernoulli - La Distribución Binomial

En esta sección presentaremos un esquema conceptual, que fue introducido por Bernoulli, y que es útil para modelizar muchas situaciones.

El esquema de ensayos de Bernoulli consiste en lo siguiente: Consideramos un experimento aleatorio con dos resultados, que convencionalmente llamamos "éxito" y "fracaso". Supongamos que la probabilidad de obtener un éxito en una realización del experimento es $p \in [0,1]$, y naturalmente la de obtener un fracaso será $q = 1-p$

Imaginemos que repetimos el experimento una cantidad n de veces, de manera independiente. Para modelizar este experimento consideramos el espacio muestral $\Omega = \{0,1\}^n$ compuesto por las n-uplas de números 0 y 1 con la siguiente interpretación: codificaremos una realización del experimento por una n-upla $\omega = (x_1, x_2, \ldots, x_n) \in \Omega$ de modo que:

$$x_i = \begin{cases} 1 & \text{si la } i\text{-ésima realización del experimento fue un "éxito"} \\ 0 & \text{si la } i\text{-ésima realización del experimento fue un "fracaso"} \end{cases}$$

Es un espacio muestral finito, con cardinal 2^n. Notemos que las funciones $X_i : \Omega \to \mathbb{R}$ (proyecciones) dadas por $X_i(\omega) = x_i$ son variables aleatorias.

¿De qué modo asignaremos las probabilidades en este espacio?. Puesto que consideramos que los ensayos son independientes, a una determinada n-upla $\omega = (x_1, x_2, \ldots, x_n)$ le asignamos la probabilidad

$$P\{\omega\} = P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \prod_{i=1}^{n} P\{X_i = x_i\}$$

Ahora la probabilidad de que $X_i = x_i$ es p si $x_i = 1$ (es un éxito) y q si $x_i = 0$ (es un fracaso). De modo que

$$P\{\omega\} = p^k q^{n-k}$$

donde $k = \sum_{i=1}^{n} x_i$ es el número de éxitos que ocurren en esa realización del experimento. Notemos que esta forma de asignar las probabilidades dice precisamente que las X_i son variables aleatorias independientes.

Por otra parte, notemos que si definimos $S_n : \Omega \to \mathbb{R}$ como el número de éxitos en los n ensayos de Bernoulli, es una variable aleatoria (en la notación anterior $S_n(\omega) = k$). Tenemos que:

$$S_n = X_1 + X_2 + \ldots + X_n \tag{3.4}$$

Nos interesa cuál es la distribución de probabilidades de S_n, es decir queremos determinar para cada k (con $0 \leq k \leq n$) cuál es la probabilidad de que S_n tome el valor k.

Observamos que el evento $\{S_n = k\} = \{\omega \in \Omega : S_n(\omega) = k\}$ se compone de las n-uplas que tienen exactamente k éxitos y $n - k$ fracasos, y que hay exactamente

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

de tales n-uplas, y cada una de ellas tiene probabilidad $p^k q^{n-k}$. En consecuencia la probabilidad del evento $S_n = k$ será

$$P\{S_n = k\} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}$$

Esta distribución de probabilidades se conoce como la **distribución binomial**, dado que viene dada por los términos del desarrollo del binomio de Newton:

$$(p+q)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}$$

Definición 3.4.1 *Sea $X : \Omega \to \mathbb{N}_0$ una variable aleatoria con valores enteros. Diremos que X tiene* **distribución binomial** *si:*

$$P\{X = k\} = b(k, n, p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \tag{3.5}$$

y $P\{X = k\} = 0$ si $k \notin \{0, 1, \ldots, n\}$. **Notación:** $X \sim Bi(n, p)$

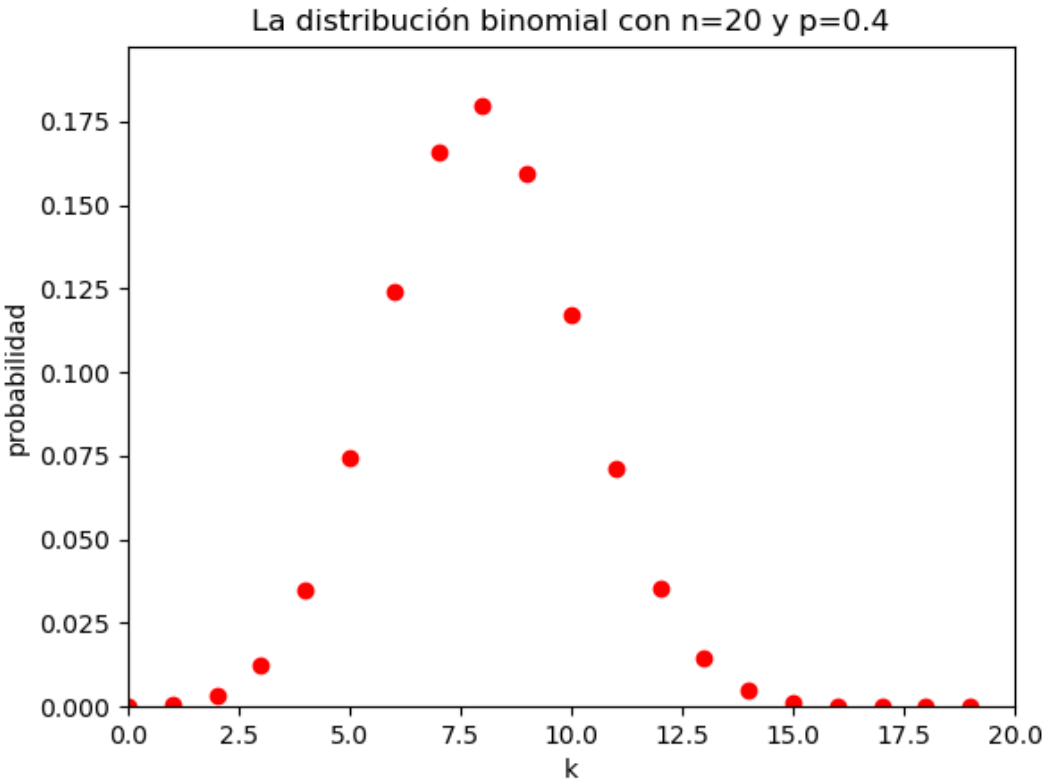


Figura 3.1: La distribución binomial con $n = 20$ y $p = 0{,}4$

Observación 3.4.2 *Si $p = 0$, tenemos que*

$$b(k,n,p) = \begin{cases} 1 & si \quad k = 0 \\ 0 & si \quad k = 1, 2, \ldots n \end{cases},$$

mientras que si $p = 1$ tenemos que

$$b(k,n,p) = \begin{cases} 0 & si \quad k = 0, 1, 2, \ldots, n-1 \\ 1 & si \quad k = n \end{cases}$$

Esto está de acuerdo con la fórmula (3.5), definiendo $0^0 = 1$.

Necesitamos calcular la esperanza y la varianza de S_n. Para ello utilizamos la representación (3.4) de S_n como suma de las variables X_i. Notamos que cada X_i es de hecho

el indicador del evento "ocurre un éxito en la i-ésima realización del experimento". En consecuencia:

$$E[X_i] = p, \quad Var(X_i) = pq$$

Por la linealidad de la esperanza,

$$E[S_n] = np$$

y por otro lado, como las X_i son variables aleatorias independientes, también se verifica que

$$\text{Var}(S_n) = npq$$

3.5. Convoluciones discretas

Consideramos dos variables aletorias discretas X e Y **independientes** con valores enteros. La distribuciones puntuales vienen dadas por las sucesiones

$$p_k = p(k) = P\{X = k\}, q_k = q(k) = P\{Y = k\} \quad k \in \mathbb{Z}$$

que podemos pensar como funciones $p, q : \mathbb{Z} \to \mathbb{R}$. ¿Cuál es la distribución de $X + Y$?

Definimos la **convolución discreta** de p y q por

$$(p * q)(k) = \sum_{m,n \in Z : m+n=k} p(m) \cdot q(n) = \sum_{m \in \mathbb{Z}} p(m) \cdot q(k-m)$$

Si X e Y toman valores naturales con probabilidad 1 (en $\mathbb{N}_0$ la fórmula se simplifica

$$(p * q)(k) = \sum_{m=0}^{k} p(m) \cdot q(k-m)$$

Proposición 3.5.1 *Si X e Y **independientes** son variables aletorias discretascon valores enteros, entonces la distribución puntual de probabilidades de $X + Y$ viene dada por $p * q$ es decir*

$$P\{X + Y = k\} = (p * q)(k)$$

Prueba:

$$\begin{aligned} P\{X + Y = k\} &= \sum_{m,n:m+n=k} P\{X = m, Y = n\} \\ &= \sum_{m,n:m+n=k} P\{X = m\} \cdot P\{Y = n\} \textbf{ por independencia} \\ &= \sum_{m,n:m+n=k} p(m) \cdot q(n) = (p * q)(k) \end{aligned}$$

□

3.6. La aproximación de Poisson a la distribución binomial

La aproximación de Poisson es una aproximación de la distribución binomial para el caso en que k es pequeño comparado con n y p es también pequeño pero $\lambda = np$ es moderado.

Empecemos desarrollando el combinatorio que aparece en la distribución binomial:

$$b(k,n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} p^k (1-p)^{n-k}$$

Notamos que en el desarrollo del combinatorio, hay k factores en el numerador. Multiplicando y dividiendo por n^k queda:

$$b(k,n,p) = \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \cdot \frac{(np)^k}{k!} (1-p)^{n-k}$$

Pongamos $\lambda = np$, entonces

$$b(k,n,p) = \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \cdot \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Como

$$\lim_{n \to +\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

deducimos que si k es pequeño en comparación con n, entonces

$$b(k,n,p) \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

Como formalización de esta idea, obtenemos el siguiente teorema:

Teorema 3.6.1 (Teorema de Poisson) *Si k está fijo, y $n \to +\infty$ de modo que $\lambda = np$ permanece fijo, entonces:*

$$lim_{n \to +\infty} b(k,n,p) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Lo que obtuvimos en el límite, es otra distribución de probabilidades que se utiliza con frecuencia y se conoce como **distribución de Poisson**:

Definición 3.6.2 *Sea $X : \Omega \to \mathbb{N}_0$ una variable aleatoria entera. Diremos que X tiene distribución de Poisson de parámetro $\lambda > 0$, si*

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

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

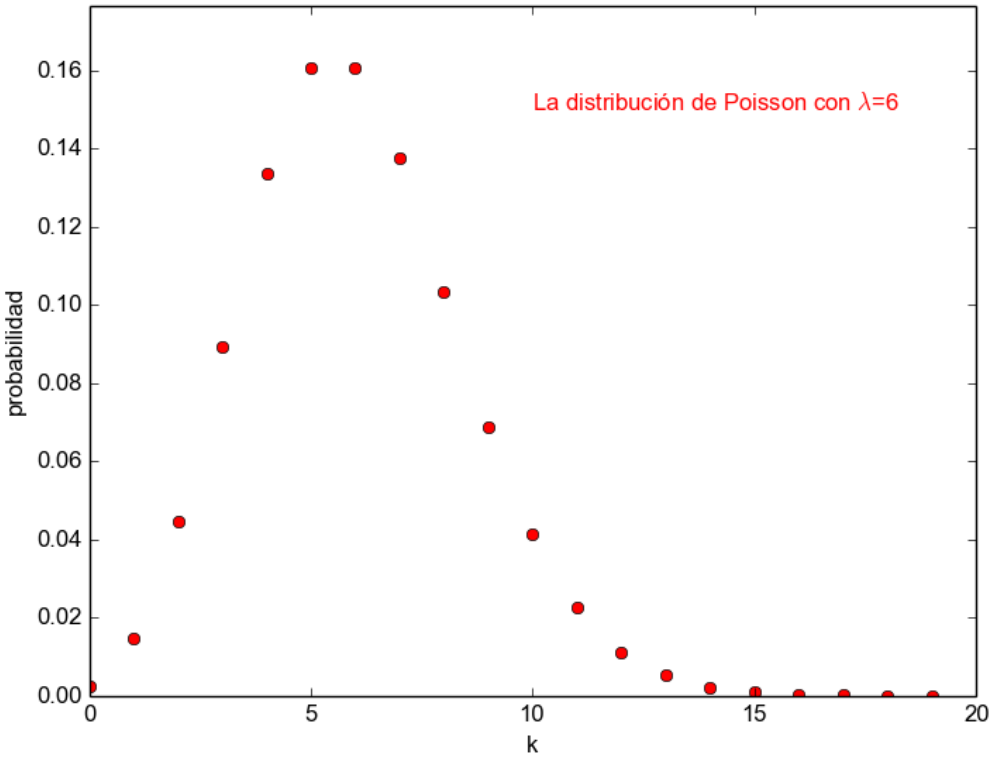


Figura 3.2: La distribución de Poisson con $\lambda = 6$.

Hay que verificar que efectivamente tenemos una distribución de probabilidades, es decir que:

$$\sum_{k=0}^{\infty} P\{X = k\} = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = 1$$

pero esto es inmediato, considerando el desarrollo en serie de e^{λ}.

Vamos a calcular ahora la esperanza y la varianza de la distribución de Poisson:

$$E[X] = \sum_{k=0}^{\infty} k \cdot P\{X = k\} = \sum_{k=1}^{\infty} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda$$

Generalizando este truco

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\cdot P\{X=k\} = \sum_{k=2}^{\infty} k(k-1)\cdot e^{-\lambda}\frac{\lambda^k}{k!}$$
$$= e^{-\lambda}\lambda^2 \sum_{k=1}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda}\cdot\lambda^2\cdot e^{\lambda} = \lambda^2$$

y por lo tanto:

$$E[X^2] = E[X(X-1)] + E[X] = \lambda^2 + \lambda$$

y

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Proposición 3.6.3 *Si $X \sim \mathcal{P}(\lambda_1)$, $Y \sim \mathcal{P}(\lambda_2)$ y son independientes, entonces $X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$.*

Prueba:

$$(p * q)(k) = \sum_{m=0}^{k} p(m)\cdot q(k-m) = \sum_{m=0}^{k} \frac{\lambda_1^m}{m!}\cdot e^{-\lambda_1}\cdot\frac{\lambda_2^{k-m}}{(k-m)!}\cdot e^{-\lambda_2}$$
$$= e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}\sum_{m=0}^{k}\frac{k!}{m!\cdot(k-m)!}\cdot\lambda_1^m\cdot\lambda_2^{k-m}$$
$$= e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}\sum_{m=0}^{k}\binom{k}{m}\cdot\lambda_1^m\cdot\lambda_2^{k-m} = e^{-(\lambda_1+\lambda_2)}\cdot\frac{(\lambda_1+\lambda_2)^k}{k!}$$

□

3.7. El método de las funciones generatrices

En algunas situaciones, el método que expondremos a continuación resulta de utilidad para operar con distribuciones de probabilidad discretas. Lo usaremos para obtener de otro modo la distribución binomial, y calcular su esperanza y su varianza.

Definición 3.7.1 *Sea $X : \Omega \to \mathbb{N}_0$ una variable aleatoria que toma valores enteros. Llamamos función generatriz de la distribución de probabilidades de X a*

$$g_X(z) = \sum_{k=0}^{\infty} P\{X=k\}z^k \ (z \in \mathbb{C})$$

suponiendo que esta serie tenga un radio de convergencia $r_X > 0$ (entonces convergerá absolutamente en $|z| < r_X$). **Observación:** *La notación g_X que usaremos en estas notas, no es una notación estándar.* [2]

Notemos que si $0 < |z| < r_X$,

$$g_X(z) = E[z^X]$$

Cuando $z = 0$ esta fórmula es problemática si X toma el valor 0. Si usamos la definición $0^0 = 1$, tiene sentido pues $g_X(0) = P\{X = 0\}$.

Observación: En virtud de la unicidad del desarrollo en serie de potencias, la distribución de probabilidades de una variable aleatoria entera está unívocamente determinada por su función generatriz.

Proposición 3.7.2 *Si X e Y son variables aleatorias independientes, entonces:*

$$g_{X+Y}(z) = g_X(z) \cdot g_Y(z)$$

para $|z| < \min(r_X, r_Y)$.

Prueba: Como X e Y son independientes, z^X y z^Y son independientes. En consecuencia, si $0 < |z| < r_X$:

$$g_{X+Y}(z) = E[z^{X+Y}] = E[z^X \cdot z^Y] = E[z^X] \cdot E[z^Y] = g_X(z) \cdot g_Y(z)$$

Cuando $z = 0$,

$$\begin{aligned} g_{X+Y}(0) &= P\{X + Y = 0\} = P\{X = 0, Y = 0\} \\ &= P\{X = 0\} \cdot P\{Y = 0\} = g_X(0) \cdot g_Y(0) \end{aligned}$$

□

Esta proposición puede generalizarse sin dificultad a varias variables independientes: si $X_1, X_2, \ldots, X_n$ son independientes, entonces

$$g_{X_1+X_2+\ldots+X_n}(z) = g_{X_1}(z) \cdot g_{X_2}(z) \cdots g_{X_n}(z)$$

Aplicación: Otra prueba de que el número de éxitos S_n en n ensayos de Bernoulli tiene distribución binomial.

Utilicemos la representación (3.4) de S_n como suma de n variables independientes que valen 1 con probabilidad p y 0 con probabilidad $q = 1 - p$. La función generatriz de cada X_i es:

$$g_{X_i}(z) = pz + q$$

[2] En clase y en versiones anteriores de estas notas utilicé la notación f_X, pero decidí cambiarla por g_X, ya que en la teoría de probabilidades la notación f_X suele utilizarse para la densidad de probabilidad para variables aleatorias absolutamente continuas.

y como S_n es la suma de las X_i y son independientes:

$$g_{S_n}(z) = (pz+q)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k z^k q^{n-k} \tag{3.6}$$

Notemos que la probabilidad de que S_n tome el valor k viene dado por el coeficiente de z^k en g_{S_n}. En consecuencia:

$$P\{S_n = k\} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \ (0 \leq k \leq n)$$

Las funciones generatrices pueden usarse para calcular esperanzas y varianzas (y más generalmente momentos) de variables aleatorias enteras:

Proposición 3.7.3 *Si la serie que define la función generatriz g_X tiene radio de convergencia $r_X > 1$, entonces*

$$E(X) = g'_X(1)$$
$$Var(X) = g''_X(1) + g'_X(1) - g'_X(1)^2$$

Prueba: Como las series de potencia pueden derivarse término a término en el interior de su disco de convergencia, tenemos que:

$$g'_X(z) = \sum_{k=1}^{\infty} kP\{X=k\}z^{k-1}$$

con convergencia absoluta si $|z| < r_X$. En particular si $z = 1$,

$$g'_X(1) = \sum_{k=1}^{\infty} kP\{X=k\} = E[X]$$

Volviendo a derivar tenemos que

$$g''_X(z) = \sum_{k=2}^{\infty} k(k-1)P\{X=k\}z^{k-2}$$

con convergencia absoluta si $|z| < r_X$, y haciendo $z = 1$,

$$g''_X(1) = \sum_{k=2}^{\infty} k(k-1)P\{X=k\} = E[X(X-1)] = E[X^2] - E[X]$$

Luego

$$Var(X) = E[X^2] - E[X]^2 = g''_X(1) + g'_X(1) - g'_X(1)^2$$

□

3.7.1. Cálculo de la esperanza y la varianza de la distribución binomial (de otra manera)

Sea como antes S_n el número de éxitos en n ensayos de Bernoulli. Como vimos antes $g_{S_n}(z) = (pz+q)^n$. En consecuencia, como

$$g'_{S_n}(z) = n(pz+q)^{n-1}p$$
$$g''_{S_n}(z) = n(n-1)(pz+q)^{n-2}p^2$$

deducimos que

$$E[S_n] = np$$

y que:

$$\mathrm{Var}(S_n) = n(n-1)p^2 + np - n^2p^2 = -np^2 + np = np(1-p) = npq$$

Ejercicio: Si $X \sim \mathrm{Bi}(n,p)$ e $Y \sim \mathrm{Bi}(m,p)$ y son independientes, entonces $X+Y \sim \mathrm{Bi}(n+m,p)$.

3.7.2. Otra aplicación: otra forma de deducir las propiedades de la distribución de Poisson

Si X tiene distribución de Poisson de parámetro λ, la función generatriz de su distribución de probabilidades es:

$$g_X(z) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k z^k}{k!} = e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)} \tag{3.7}$$

Tenemos que

$$g'_X(z) = \lambda e^{\lambda(z-1)}$$
$$g''_X(z) = \lambda^2 e^{\lambda(z-1)}$$

En consecuencia por la proposición 3.7.3, deducimos que:

$$E(X) = g'_X(1) = \lambda$$
$$\mathrm{Var}(X) = g''(1) + g'(1) - g'(1)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

También podemos dar otra prueba de la proposición 3.6.3, cuyo enunciado recordamos:

Proposición 3.7.4 *Si $X \sim \mathcal{P}(\lambda_1)$, $Y \sim \mathcal{P}(\lambda_2)$ y son independientes, entonces $X+Y \sim \mathcal{P}(\lambda_1+\lambda_2)$.*

Prueba: Por la proposición 3.7.2,

$$g_{X+Y}(z) = g_X(z)\cdot g_Y(z) = e^{\lambda_1(z-1)}e^{\lambda_2(z-1)} = e^{(\lambda_1+\lambda_2)(z-1)}$$

En consecuencia, $X+Y \sim \mathcal{P}(\lambda_1+\lambda_2)$, ya que la distribución de probabilidades de $X+Y$ está determinada por su función generatriz. □

3.7.3. El teorema de Bernoulli

Imaginemos que realizamos una sucesión ilimitada de ensayos de Bernoulli. Sea $f_n = \frac{S_n}{n}$ la frecuencia de éxitos que obtenemos en los n primeros ensayos. Es intuitivamente razonable que conforme $n \to +\infty$, f_n tienda a la probabilidad p de obtener un éxito.

Nos gustaría transformar esta idea intuitiva en un teorema matemático. El siguiente teorema debido a Jacques Bernoulli, y publicado en 1713 en su libro *Ars Conjectandi*, constituye una formalización de esta idea:

Teorema 3.7.5 (Teorema de J. Bernoulli) *Sea f_n la frecuencia de éxitos en los n primeros ensayos de una sucesión ilimitada de ensayos de Bernoulli. Entonces dado cualquier $\delta > 0$,*

$$P\{|f_n - p| > \delta\} \to 0 \text{ conforme } n \to \infty$$

Prueba: Notemos que $E[f_n] = p$. Luego, por la desigualdad de Chebyshev,

$$P\{|f_n - p| > \delta\} \leq \frac{\text{Var}(f_n)}{\delta^2}$$

pero

$$\text{Var}(f_n) = \text{Var}\left(\frac{S_n}{n}\right) = \frac{pq}{n}$$

En consecuencia:

$$P\{|f_n - p| > \delta\} \leq \frac{pq}{n\delta^2} \to 0 \text{ cuando } n \to +\infty \tag{3.8}$$

□

Una generalización del teorema de Bernoulli (que se prueba con el mismo argumento) es la siguiente, conocida (al igual que a veces el teorema de Bernoulli) como la ley débil de los grandes números:

Teorema 3.7.6 (Ley débil de los grandes números - caso de variancia finita) *Sean $X_1, X_2, \ldots, X_n, \ldots$ una secuencia infinita de variables aleatorias independientes e idénticamente distribuidas, con*

$$E[X_i] = \mu$$

$$Var(X_i) = \sigma^2 < +\infty$$

Entonces si llamamos

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

y tomamos cualquier $\delta > 0$, tenemos que

$$P\{|\overline{X}_n - \mu| > \delta\} \to 0 \text{ cuando } n \to +\infty$$

Prueba: Por linealidad de la esperanza, $E[\overline{X}_n] = \mu$, y por otro lado

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

ya que las X_i son independientes. La desigualdad de Chebyshev, dice entonces que:

$$P\{|\overline{X}_n - \mu| > \delta\} \leq \frac{\sigma^2}{n\delta^2} \to 0 \text{ cuando } n \to +\infty$$

□

Algunas observaciones sobre el teorema de Bernoulli:

- Si bien la prueba del teorema de Bernoulli, resulta muy sencilla hoy en día, J. Bernoulli dice en su libro que estuvo pensando en este teorema durante más de 20 años, lo cuál muestra que el resultado no es para nada trivial.
- Como todo teorema matemático, el teorema de Bernoulli no afirma nada sobre la realidad, es solamente una afirmación sobre el modelo matemático

  (La cuestión de la validez práctica de un modelo matemático sólo se puede decidir sobre bases empíricas, es decir contrastándolo con la experiencia). Sin embargo, podemos interpretarlo como una muestra de la consistencia interna de nuestro modelo matemático.
- La ley débil de los grandes números recibe este nombre, porque, como veremos más adelante, existe otro teorema conocido como la ley fuerte de los grandes números, que afirma que en realidad $S_n \to p$ (o $\overline{X}_n \to \mu$) con probabilidad 1.

  (Pero notemos que para darle sentido a la afirmación de que $S_n \to p$ con probabilidad 1, debemos asignar probabilidades a secuencias de infinitos ensayos de Bernoulli, como en el experimento que consideramos anteriormente de arrojar infinitas veces una moneda. Esto introduce ciertas dificultades relacionadas con la teoría de la medida, como por ejemplo que ya no podremos asignarle probabilidad a cualquier parte del espacio muestral Ω, y que por lo tanto debemos restringir el dominio de la función probabilidad a una σ-álgebra de eventos.)

3.8. Ley débil de los grandes números: caso general

La hipótesis de que las variables aleatorias X_i tengan varianza finita no es realmente necesaria para la validez de la ley débil de los grandes números, pudiéndose probar para variables que tengan solamente esperanza finita, por medio de un método de truncamiento. Sin embargo, para fijar ideas, hemos optado por enunciarla y demostrarla primero en este caso en el que la demostración resulta más sencilla. Veamos ahora el caso general:

Teorema 3.8.1 (Ley débil de los grandes números - caso general) *Sean $X_1, X_2, \ldots, X_n, \ldots$ una secuencia infinita de variables aleatorias independientes e idénticamente distribuidas, con*

$$E[X_i] = \mu < +\infty$$

Entonces si llamamos

$$S_n = X_1 + X_2 + \ldots + X_n$$

y tomamos cualquier $\delta > 0$, tenemos que

$$P\left\{\left|\frac{S_n}{n} - \mu\right| > \delta\right\} \to 0 \text{ cuando } n \to +\infty$$

Prueba: Para simplificar la notación, notemos que podemos asumir sin pérdida de generalidad, que

$$E(X_i) = 0 \; \forall \; i$$

(cambiando si no X_i por $X_i - \mu$).

La demostración en el caso de variancia infinita, se basa en el **método de truncamiento**, que consiste en descomponer X_i como suma de dos variables aleatorias. Para cada $k = 1, 2, \ldots, n$, escribimos:

$$X_k = U_{n,k} + V_{n,k} \; (k = 1, 2, \ldots, n) \tag{3.9}$$

donde

$$U_{n,k} = \begin{cases} X_k & \text{si} \quad |X_k| \leq \lambda n \\ 0 & \text{si} \quad |X_k| > \lambda n \end{cases}$$

y

$$V_{n,k} = \begin{cases} 0 & \text{si} \quad |X_k| \leq \lambda n \\ X_k & \text{si} \quad |X_k| > \lambda n \end{cases}$$

donde $\delta > 0$ es una constante que especificaremos después. Y pongamos:

$$U_n = U_{n,1} + U_{n,2} + \ldots + U_{n,n}$$
$$V_n = V_{n,1} + V_{n,2} + \ldots + V_{n,n}$$

De la desigualdad triangular $|S_n| \leq |U_n| + |V_n|$, y de la subaditividad de la probabilidad, deducimos que:

$$\begin{gathered} P\{|S_n| > \delta n\} \\ \leq P\{|U_n| > \delta n/2\} + P\{|V_n| > \delta n/2\} \end{gathered} \tag{3.10}$$

Entonces hemos de probar que cada una de las probabilidades del segundo miembro tiende a cero cuando $n \to +\infty$.

Comencemos acotando:

$$P\{|U_n| > \delta n/2\}$$

Observemos que las variables $U_{n,k}$ están acotadas ($|U_{n,k}| \leq \lambda n$) y en consecuencia tienen segundo momento finito. Más explícitamente, si llamemos $a = E(|X_i|)$, tenemos que

$$E(U_{n,k}^2) \leq n\lambda a$$

En consecuencia las $U_{k,n}$ tienen variancia finita:

$$\mathrm{Var}(U_{n,k}) \leq E(U_{n.k}^2) \leq n\lambda a$$

Por otra parte las $U_{n,k}$ son variables independientes e idénticamente distribuidas (pues $U_{n,k}$ es función de X_k, y las X_k eran independientes e idénticamente distribuidas). En consecuencia:

$$\mathrm{Var}(U_n) = \mathrm{Var}(U_{n,1} + U_{n,2} + \ldots + U_{n,n}) = \sum_{k=1}^{n} \mathrm{Var}(U_{n,k}) \leq n^2\lambda a$$

Además de la definición de las $U_{n,k}$ deducimos que

$$E(U_{n,k}) = E(U_{n,1}) = \sum_{i:|x_i|>\lambda n} x_i P\{X_1 = x_i\} \to E(X_1) = 0$$

conforme $n \to +\infty$. En consecuencia para $n \geq n_0(\varepsilon)$ será:

$$E(U_n^2) = \mathrm{Var}(U_n) + E(U_n)^2 < 2\lambda n^2 a$$

y entonces por la desigualdad de Chebyshev, tenemos que:

$$P\{|U_n| > \delta n/2\} < \frac{8a\lambda}{\delta^2} < \frac{\varepsilon}{2}$$

si elegimos λ suficientemente pequeño.

En cuanto al segundo término: obviamente

$$P\{|V_n| > \delta n/2\} \leq P\{V_{n,1} + V_{n,2} + \ldots + V_{n,n} \neq 0\}$$

y como

$$\{V_{n,1} + V_{n,2} + \ldots + V_{n,n} \neq 0\} \subset \bigcup_{k=1}^{n} \{V_{n,k} \neq 0\}$$

tenemos que:

$$P\{|V_n| > \delta n/2\} \leq \sum_{k=1}^{n} P\{V_{n,k} \neq 0\} = nP\{V_1 \neq 0\}$$

ya que las V_k tienen todas la misma distribución de probabilidades. Pero por definición de V_1, esto dice que

$$P\{|V_n| > \delta n/2\} \leq nP\{|X_1| > \lambda n\} = n \sum_{i:|x_i|>\lambda n} P\{X_1 = x_i\}$$

donde $\mathrm{Im}(X_1) = \{x_1, x_2, \ldots, x_n \ldots\}$. Deducimos que:

$$P\{|V_n| > \delta n/2\} \leq \frac{1}{\lambda} \sum_{|x_i|>\lambda n} |x_i| P\{X_1 = x_i\}$$

Dado entonces cualquier $\varepsilon > 0$, como la esperanza de X_1 es finita por hipótesis, deducimos que si elegimos n suficientemente grande, digamos si $n \geq n_0(\varepsilon)$, tendremos que:

$$P\{|V_n| > \delta n/2\} < \frac{\varepsilon}{2}$$

(ya que las colas de una serie convergente tienden a cero).

Por (3.10), deducimos que:

$$P\{|S_n| > \delta n\} \leq \varepsilon$$

si $n \geq n_0(\varepsilon)$. □

3.9. Polinomios de Bernstein: Una prueba del teorema de Weierstrass

En esta sección expondremos una prueba del teorema de Weierstrass sobre aproximación a funciones continuas por polinomios, debida a S.N. Bernstein:

Teorema 3.9.1 (Weierstrass) *Sea $f \in C[0,1]$ una función continua $f : [0,1] \to \mathbb{R}$, entonces existe una sucesión de polinomios $P_n(t)$ tal que $P_n(t) \to f(t)$ uniformemente para $t \in [0,1]$.*

En un lenguaje más moderno, el teorema de Weierstrass dice que los polinomios son densos en el espacio $C[0,1]$ de las funciones continuas (con la norma del supremo).

La prueba de S.N. Berstein (1912) [Ber12] de este teorema, consiste en utilizar la distribución binomial, para construir explícitamente una sucesión de polinomios que converge uniformemente a f.

Veamos primero la idea intuitiva de la demostración: sea $p \in [0,1]$ y sea como antes S_n el número de éxitos en n ensayos de Bernoulli con probabilidad p. La ley de los grandes números afirma que:

$$\frac{S_n}{n} \to p \text{ (en probabilidad)}$$

y como f es continua es razonable esperar que:

$$f\left(\frac{S_n}{n}\right) \to f(p)$$

(De vuelta, esto no es estrictamente cierto para toda sucesión de ensayos de Bernoulli, pero sí vale en probabilidad.) Por lo que esperamos que:

$$E\left[f\left(\frac{S_n}{n}\right)\right] \to E[f(p)] = f(p)$$

Notemos que:

$$B_n(p) = E\left[f\left(\frac{S_n}{n}\right)\right] = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b(k,n,p)$$

$$= \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) f\left(\frac{k}{n}\right) p^k (1-p)^{n-k}$$

es un polinomio en la variable p de grado menor o igual que n. Se lo denomina el n-ésimo **polinomio de Bernstein**.

Observación 3.9.2 *En esta fórmula, debemos interpetar* $0^0 = 1$*, de acuerdo con la observación . Se deduce que* $B_n(0) = f(0), B_n(1) = f(1)$.

La demostración de S.N. Bernstein, consiste en probar que $B_n(p) \to f(p)$ uniformemente para $p \in [0,1]$ (Los argumentos anteriores no constituyen una prueba rigurosa, pero explican intuitivamente por qué esta afirmación es cierta).

De hecho, la demostración de esta afirmación se basa en argumentos muy similares a los que nos llevaron a la prueba del teorema de Bernoulli.

Para la prueba del teorema de Weierstrass utilizaremos, dos propiedades claves de las funciones continuas en un intervalo cerrado de la recta, a saber:

1. Una función continua en un intervalo cerrado de la recta, es acotada: existe una constante $M > 0$ tal que:
$$|f(p)| \leq M \ \forall \ p \in [0,1]$$

2. Una función continua en un intervalo cerrado de la recta, es uniformemente continua: dado $\varepsilon > 0$ existe $\delta > 0$ tal que si $x, y \in [0,1]$ y si $|x-y| \leq \delta$, entonces $|f(x)-f(y)| < \varepsilon$.

Necesitaremos una acotación de las colas de la distribución binomial: de acuerdo a la desigualdad (3.8):

$$P\left\{\left|\frac{S_n}{n} - p\right| > \delta\right\} \leq \frac{pq}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

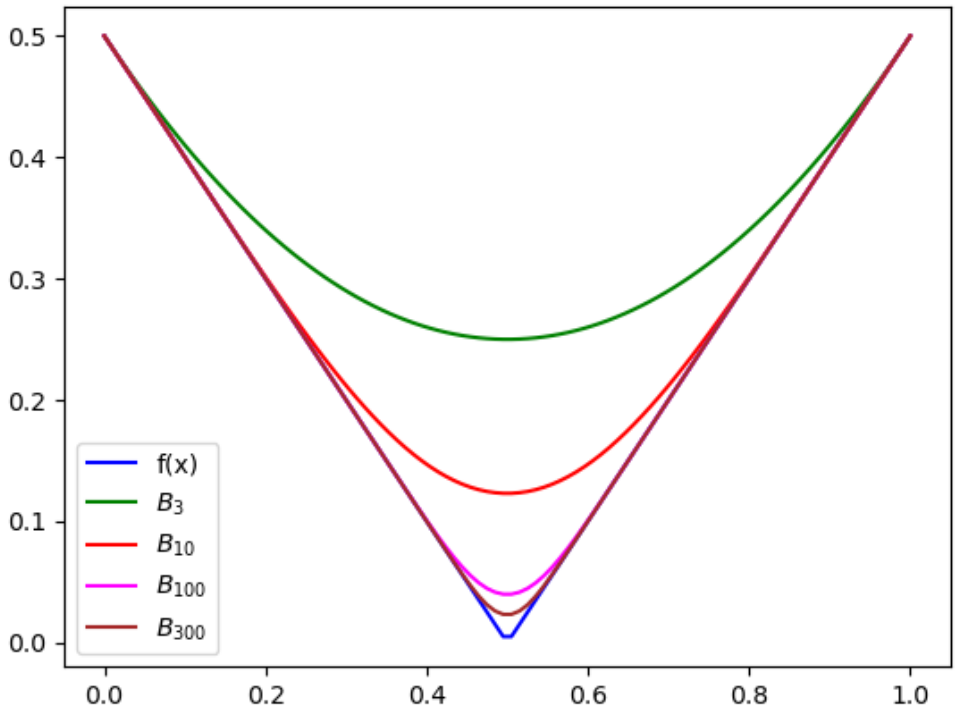


Figura 3.3: Aproximación a $f(x) = |\frac{1}{2} - x|$ mediante los polinomios de Bernstein.

ya que:

$$pq = p(1-p) \leq \frac{1}{4} \; \forall \; p \in [0,1]$$

Más explícitamente podemos escribir esto como:

$$\sum_{|k/n-p|>\delta} b(k,n,p) = \sum_{|k/n-p|>\delta} P\{S_n = k\} \leq \frac{1}{4n\delta^2}$$

Queremos acotar la diferencia:

$$B_n(p) - f(p) = \sum_{k=0}^{n} \left[f\left(\frac{k}{n}\right) b(k,n,p) \right] - f(p) = \sum_{k=0}^{n} \left[f\left(\frac{k}{n}\right) - f(p) \right] b(k,n,p)$$

pues

$$\sum_{k=0}^{n} b(k,n,p) = 1$$

(¡Es una distribución de probabilidades!). En consecuencia,

$$|B_n(p) - f(p)| \leq \sum_{k=0}^{n} \left| f\left(\frac{k}{n}\right) - f(p) \right| b(k,n,p)$$

En esta suma separamos dos partes, la suma sobre los k donde $|k/n - p| \leq \delta$ (con el δ dado por la continuidad uniforme), y la parte donde $|k/n - p| > \delta$.

La primer parte la acotamos, fácilmente:

$$\sum_{k:|k/n-p|\leq\delta} \left| f\left(\frac{k}{n}\right) - f(p) \right| b(k,n,p) \leq \sum_{k:|k/n-p|\leq\delta} \varepsilon\, b(k,n,p) \leq \varepsilon$$

pues los $b(k,n,p)$ suman 1.

La otra parte de la suma la acotamos usando nuestra estimación de las colas de la distribución binomial:[3]

$$\sum_{k:|k/n-p|>\delta} \left| f\left(\frac{k}{n}\right) - f(p) \right| b(k,n,p) \leq 2M \sum_{|k/n-p|>\delta} b(k,n,p) < \frac{2M}{4n\delta^2} < \varepsilon$$

si $n \geq n_0(\varepsilon)$. En consecuencia, $|B_n(p) - f(p)| < 2\varepsilon$ si $n \geq n_0(\varepsilon)$, para todo $p \in [0,1]$. Esto concluye la prueba del teorema de Weierstrass.

3.10. Otras distribuciones relacionadas con los ensayos de Bernoulli

Distribución Geométrica

Supongamos que realizamos una secuencia infinita de ensayos de Bernoulli, con probabilidad de éxito p. Sea T_1 la cantidad de ensayos que tenemos que realizar hasta obtener el primer éxito (esto generaliza el ejemplo de la página 35 que corresponde al caso $p = 1/2$.).

Entonces, si $T_1 = k$ significa que los primeros $k-1$ ensayos fueron fracasos y el k-ésimo fue un éxito, y como los ensayos son independientes obtenemos como antes que:

$$P\{T_1 = k\} = q^{k-1}p = (1-p)^{k-1}p$$

[3] Si en lugar de utilizar la desigualdad de Chebyshev, utilizamos otra herramienta de probabilidades conocida como la "teoría de grandes desviaciones", es posible obtener una acotación más precisa del error de aproximar f por B_n. Ver el artículo [GP97] citado en la bibliografía

(y $T_1 = +\infty$ con probabilidad cero). Esta distribución se conoce con el nombre de **distribución geométrica** de parámetro p.

Notación: $X \sim \mathcal{G}](p)$ significa que X se distribuye con la distribución geométrica de parámetro p.

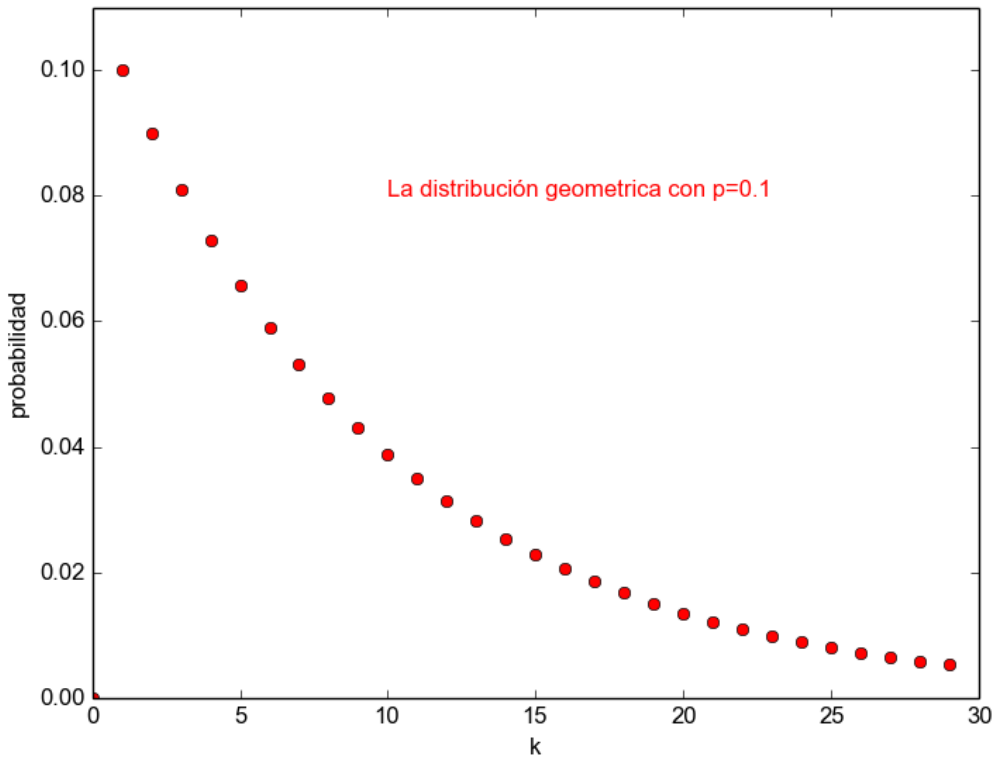


Figura 3.4: La distribución geométrica con $p = 0{,}1$.

Con una cuenta análoga a la que hicimos antes para el caso $p = 1/2$ podemos probar que $E[X] = \frac{1}{p}$.

La función generatriz de la distribución de probabilidades de X se obtiene justamente sumando una serie geométrica:

$$g_X(z) = \sum_{k=1}^{\infty} q^{k-1} p z^k = \frac{pz}{1-qz} \text{ si } |z| < \frac{1}{q} \tag{3.11}$$

Distribución binomial negativa

Más generalmente podemos considerar la variable T_r definida como el número de ensayos que tenemos que realizar hasta obtener r éxitos. Queremos calcular la distribución de T_r:

Para ello notamos que,

$$T_r = E_1 + E_2 + \ldots + E_r$$

donde $E_1 = T_1$ y $E_j =$ número de ensayos que debemos realizar después del éxito $j-1$ para obtener el siguiente éxito. Notamos que las variables E_j son independientes (ya que el tiempo que tenemos que esperar para obtener el siguiente éxito después de obtener $j-1$ éxitos no depende de cuánto tardamos en obtener j éxitos) y que por la discusión anterior, cada E_j tiene distribución geométrica de parámetro p.

Podemos entonces calcular la distribución de T_r utilizando el método de las funciones generatrices, ya que por la independencia de las E_j, la función generatriz de la distribución de probabilidades de T_r es:

$$g_{T_r}(z) = g_{E_1}(z)g_{E_2}(z)\cdots g_{E_r}(z) = \left(\frac{pz}{1-qz}\right)^r$$

Por lo tanto, utilizando el desarrollo del binomio $(1-qz)^{-r}$ y haciendo el cambio de índice $k = j + r$,

$$g_{T_r}(z) = (pz)^r \sum_{j=0}^{\infty} \binom{-r}{j} (-qz)^j = \sum_{k=r}^{\infty} \binom{-r}{k-r} p^r (-q)^{k-r} z^k$$

En consecuencia,

$$P\{T_r = k\} = \binom{-r}{k-r} p^r (-q)^{k-r} \ (k = r, r+1, \ldots)$$

Notamos que:

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

pues $(k-1)-(r-1) = k-r$.

Entonces alternativamente podemos escribir:

$$P\{T_r = k\} = \left(\begin{array}{c} k-1 \\ r-1 \end{array} \right) p^r q^{k-r} \ (k = r, r+1, \ldots)$$

Notación: $X \sim BN(r, p)$

Falta: distribucion hipergeometrica

Distribución Multinomial

Es una generalización de la distribución binomial donde consideramos experimentos con muchos varios posibles, en lugar de un experimento con sólo dos resultados.

Consideramos un experimento con N resultados posibles, y supongamos que la probabilidad de que ocurra el i-ésimo resultado en una realización del experimento es p_i, de modo que:

$$\sum_{i=1}^{N} p_i = 1$$

Supongamos que repetimos el experimento n veces en condiciones independientes, y llamemos X_i a la cantidad de veces que ocurre el i-ésimo resultado, de modo que:

$$X_1 + X_2 + \ldots + X_N = n$$

Entonces, la distribución de probabilidades conjunta de las X_i viene dada por:

$$P\{X_1 = k_1, X_2 = k_2, \ldots, X_N = k_N\} = \frac{n!}{k_1! k_2 \ldots k_N!} \, p_1^{k_1} p_2^{k_2} \ldots p_N^{k_N} \tag{3.12}$$

si $k_1 + k_2 + \ldots + k_N = N$ (y cero en caso contrario). Notamos que $X = (X_1, X_2, \ldots, X_N)$ es un vector aleatorio N-dimensional.

Notación: $X \sim \mathcal{M}(n, p_1, p_2, \ldots, p_N)$

Esta distribución recibe este nombre, debido a su relación con el desarrollo multinomial:

$$(x_1 + x_2 + \ldots + x_N)^n = \sum_{\substack{k_N : k_1 + k_2 + \ldots + k_N = n \\ 0 \leq k_i \leq n}} \frac{n!}{k_1! k_2 \ldots k_N!} \, x_1^{k_1} x_2^{k_2} \ldots x_N^{k_N}$$

(Tomando $x_i = p_i$ se ve que las probabilidades en (3.12) suman 1, por lo que se trata efectivamente de una distribución de probabilidades).

Una propiedad interesante de la distribución multinomial es que las distribuciones de cada una de las X_i por separado (distribuciones marginales) son binomiales:

Proposición 3.10.1 *Si $X \sim \mathcal{M}(n, p_1, p_2, \ldots, p_N)$, entonces*

$$X_i \sim Bi(n, p_i) \quad 0 \leq i \leq N$$

Prueba: Por simetría, basta verlo para la distribución de X_1. Si $0 \leq k_1 \leq n$,

$$\begin{aligned}
P\{X_1 = k_1\} &= \sum_{\substack{k_N : k_2 + \ldots + k_N = n - k_1 \\ 0 \leq k_i \leq n}} P\{X_1 = k_1, X_2 = k_2, \ldots, X_N = k_N\} \\
&= \sum_{\substack{k_N : k_2 + \ldots + k_N = n - k_1 \\ 0 \leq k_i \leq n}} \frac{n!}{k_1! k_2 \ldots k_N!} \, p_1^{k_1} p_2^{k_2} \ldots p_N^{k_N} \\
&= \frac{n!}{k_1!(n-k_1)!} \, p_1^{k_1} \sum_{\substack{k_N : k_2 + \ldots + k_N = n - k_1 \\ 0 \leq k_i \leq n}} \frac{(n-k_1)!}{k_2! \ldots k_N!} \, p_2^{k_2} \ldots p_N^{k_N} \\
&= \frac{n!}{k_1!(n-k_1)!} \, p_1^{k_1} (p_2 + p_3 + \ldots + p_N)^{n-k_1} \\
&= \frac{n!}{k_1!(n-k_1)!} \, p_1^{k_1} (1 - p_1)^{n-k_1}
\end{aligned}$$

luego

$$X_1 \sim \mathrm{Bi}(n, p_1)$$

□

Capítulo 4

Distribuciones Continuas

4.1. Variables aleatorias continuas

En este capítulo estudiaremos variables aleatorias no discretas, en particular variables continuas. La idea básica es la misma que antes: una variable aleatoria es un número asociado al resultado de un experimento aleatorio, por lo que será una función X definida sobre el espacio muestral Ω. Nuevamente, hay un requerimiento técnico, derivado del hecho de que en general no resulta posible asignar probabilidades a todas las partes de Ω; a saber que podamos calcular las probabilidades asociadas a dicha función. En el caso de variables discretas, pedíamos que estuvieran definidas las probabilidades de que X tome un determinado valor. En el caso de variables no discretas, esto no será suficiente: requeriremos que podamos calcular la probabilidad de que el valor de X caiga en un intervalo dado de la recta.

Definición 4.1.1 *Sea $(\Omega, \mathcal{E}, P)$ un espacio de probabilidad. Una variable aleatoria será una función $X : \Omega \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, con la siguiente propiedad: para cualquier intervalo de la recta $(a, b]$ $(a, b \in \overline{\mathbb{R}})$ la preimagen $X^{-1}(a, b] = \{\omega \in \Omega : a < X(\omega) \leq b\}$ pertenece a $\mathcal{E}$, es decir está definida la probabilidad $P(X^{-1}(a, b]) = P\{a < X \leq b\}$ de que X tome un valor entre a y b.*

Observación: En análisis real, el concepto análogo es el de *función medible* (ver apéndice D).

Definición 4.1.2 *Diremos que la variable X es (absolutamente) continua si existe una función integrable*[1] *no negativa $f : \mathbb{R} \to R_{\geq 0}$ tal que*

$$P\{a < X \leq b\} = \int_a^b f(x)\, dx$$

[1] Quiere decir que en algún sentido sea posible calcular la integral de f sobre un intervalo de la recta. Los que no conozcan la teoría de la integral de Lebesgue pueden pensar integrable Riemann, los que cursaron análisis real pueden pensar que es integrable Lebesgue

La función f debe verificar que:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Se dice que f se distribuye según la densidad de probabilidades $f(x)$ (o que f es la densidad de probabilidad de X). A veces se nota, $X \sim f(x)$.

Definición 4.1.3 *Si $X : \Omega \to \overline{\mathbb{R}}$ es una variable aleatoria, su función de distribución*[2] *será la función $F : \mathbb{R} \to \mathbb{R}$ dada por:*

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

Si X es absolutamente continua, y se distribuye según la densidad $f(x)$ tendremos:

$$F_X(x) = \int_{-\infty}^{x} f(t)\,dt$$

Ejemplo 4.1.4 Variables aleatorias discretas: *Sea X una variable aleatoria discreta que toma una sucesión a lo sumo numerable de valores (x_i). Entonces, X es una variable aleatoria de acuerdo a nuestra nueva definición (es decir, realmente estamos extendiendo el concepto) ya que:*

$$\{\omega \in \Omega : a < X(\omega) \leq b\} = \bigcup_{a<x_i\leq b} \{\omega \in \Omega : X(\omega) = x_i\}$$

Por definición de variable aleatoria discreta, $\{\omega \in \Omega : X(\omega) = x_i\} \in \mathcal{E}$, y como siendo la clase $\mathcal{E}$ una σ-álgebra, es cerrada por uniones numerables, deducimos que $\{\omega \in \Omega : a < X(\omega) \leq b\} \in \mathcal{E}$.

La función de distribución de X viene dada por la función "en escalera"

$$F_X(x) = \sum_{x_i<x} P\{X = x_i\}$$

que tiene un salto de magnitud $p_i = P\{X = x_i\}$ en el punto x_i (y que es constante en cada intervalo entre dos x_i).

Ejemplo 4.1.5 *Volvamos a considerar el experimento de elegir un número real en el intervalo $[0, 1]$ con distribución uniforme. Sea X el número obtenido.*

Que lo elegimos con distribución uniforme significa que para cualquier intervalo $I \subset [0, 1]$, postulamos que

$$P\{X \in I\} = |I|$$

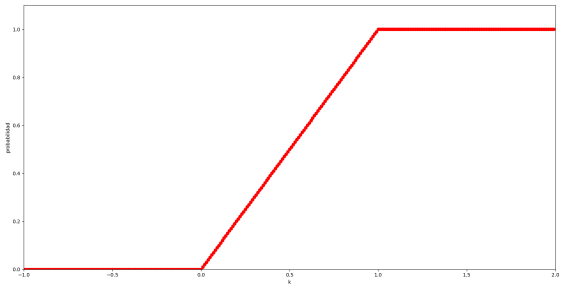


Figura 4.1: La función de distribución de una variable aleatoria con distribución uniforme en el intervalor $[0, 1]$.

donde $|I|$ representa la medida del intervalo.

Entonces la función de distribución de X viene dada por:

$$F_X(x) = \begin{cases} 0 & si \quad x < 0 \\ x & si \quad 0 \leq x \leq 1 \\ 1 & si \quad x > 1 \end{cases}$$

X es una variable absolutamente continua con densidad,

$$f_X(x) = \begin{cases} 1 & si \quad x \in [0,1] \\ 0 & si \quad x \notin [0,1] \end{cases}$$

Notación: *Notamos X se distribuye uniformemente en el intervalo $[0, 1]$ del siguiente modo: $X \sim \mathcal{U}(0, 1)$.*

Más generalmente si $[a, b]$ es un intervalo de la recta, decimos que X tiene distribución uniforme en el intervalo $[a, b]$ (Notación: $X \sim \mathcal{U}(a, b)$) si para cualquier intervalo $I \subset [a, b]$ la probabilidad de que X pertenezca a I es proporcional a la medida de I, es decir:

$$P\{X \in I\} = \frac{|I|}{b-a}$$

En este caso, la función de distribución es:

$$F_X(x) = \begin{cases} 0 & si \quad x < a \\ (x-a)/(b-a) & si \quad a \leq x \leq b \\ 1 & si \quad x > b \end{cases}$$

[2] También llamada a veces función de distribución acumulada en la literatura.

y la función de densidad es,

$$f_X(x) = \begin{cases} \frac{1}{b-a} & si \quad x \in [a,b] \\ 0 & si \quad x \notin [a,b] \end{cases}$$

Ejemplo 4.1.6 *Decimos que X tiene* **distribución normal**, *y lo notaremos $X \sim N(\mu, \sigma^2)$, si su función de densidad de probabilidad viene dada por:*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)}$$

donde μ, σ son dos parámetros reales con $\sigma > 0$. El caso $\mu = 0$, $\sigma = 1$, es decir $N(0,1)$, se conoce como **distribución normal estándar**.

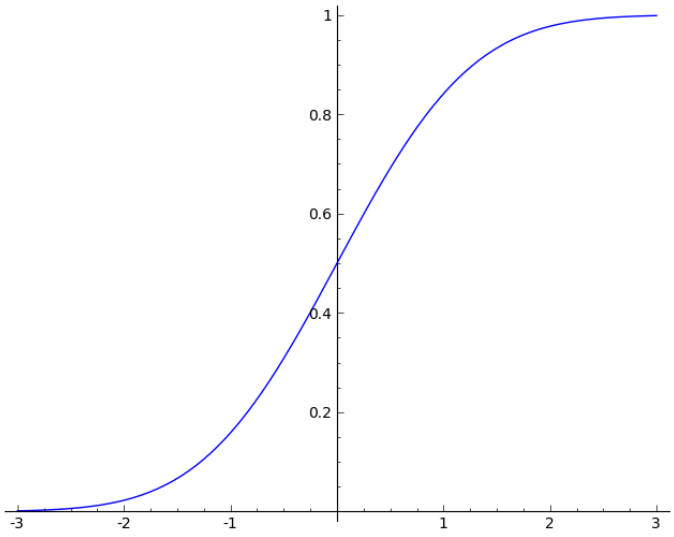


Figura 4.2: La densidad normal estándar

Si $X \sim N(0,1)$, la función de distribución de X será la función:

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/(2\sigma^2)}\, dt \tag{4.1}$$

Veremos en el capítulo 11 que la distribución normal resulta útil por ejemplo para aproximar la distribución binomial, del número S_n de éxitos en n ensayos de Bernoulli, cuando el número de ensayos es grande. Más generalmente, se puede usar para aproximar la suma de muchas variables aleatorias independientes cada una de las cuáles hace una pequeña contribución a la varianza de la suma (Este es el contenido del Teorema del Límite Central

que veremos en dicho capítulo). Como consecuencia, esta distribución juega un papel central en estadística. Se conoce también como distribución de Laplace o de Gauss.

Muchas ejemplos de datos reales se ajustan muy bien a esta distribución. Un ejemplo clásico es la altura en la población humana [MRR13].

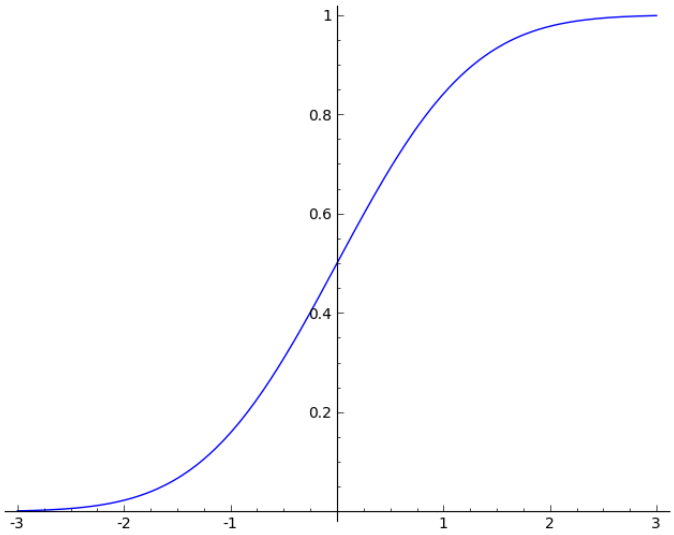


Figura 4.3: La función de distribución de una variable con distribución normal estándar

4.1.1. Propiedades de las funciones de distibución

El siguiente lema nos dice que propiedades tienen las funciones de distribución:

Lema 4.1.7 *Sea $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria y $F = F_X$ su función de distribución. Entonces F tiene las siguientes propiedades:*

i) $0 \leq F(x) \leq 1$ *y F es creciente.*

ii) F es continua por la derecha.

iii) $F(x_0) - \lim_{x \to x_0^-} F(x) = P\{X = x_0\}$ *En particular, F es continua en $x = x_0$ si y sólo si* $P\{X = x_0\} = 0$.

iv) Si X es finita con probabilidad 1 (o sea $P\{X = \pm\infty\} = 0$*), entonces:*

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to +\infty} F(x) = 1$$

Observación 4.1.8 *La propiedad iii) significa que los saltos de una función de distribución nos indican cuando la probabilidad se concentra en un punto dado x_0, y que la magnitud del salto nos dice cuanta probabilidad se concentra en ese punto x_0.*

Prueba: i) Que $0 \leq F(x) \leq 1$ es obvio por ser $F(x)$ una probablidad. Si $x_1 \leq x_2$ tenemos que: $\{X \leq x_1\} \subset \{X \leq x_2\}$, y en consecuencia $F(x_1) \leq F(x_2)$.

ii) Sea $x_0 \in \mathbb{R}$ y consideremos una sucesión decreciente $(x_n)_{n\in N} > x_0$ que converja a x_0. Entonces,

$$\{X \leq x_0\} = \bigcap_{n\in N} \{X \leq x_n\}$$

Es la intersección de una familia decreciente numerable de eventos. Entonces, por las propiedades de continuidad de la probabilidad:

$$P\{X \leq x_0\} = \lim_{n\to+\infty} P\{X \leq x_n\}$$

Es decir que:

$$F(x_0) = \lim_{n\to+\infty} F(x_n)$$

Y como esto vale para toda sucesión $(x_n) > x_0$ decreciente, que converja a x_0 deducimos que:

$$F(x_0) = \lim_{x\to x_0^+} F(x)$$

Es decir, que F es continua por la derecha.

iii) Análogamente, sea $x_0 \in \mathbb{R}$ y tomemos una sucesión creciente $(x_n)_{n\in N} < x_0$ que converja a x_0. Ahora tenemos que,

$$\{X < x_0\} = \bigcup_{n\in N} \{X \leq x_n\}$$

Entonces, aplicando nuevamente las propiedades de continuidad de la probabilidad:

$$P\{X < x_0\} = \lim_{n\to+\infty} P\{X \leq x_n\}$$

Es decir que:

$$P\{x < x_0\} = \lim_{n\to+\infty} F(x_n)$$

Como esto valle para toda sucesión $(x_n)_{n\in N} < x_0$ que converja a x_0, deducimos que:

$$\lim_{x\to x_0^-} F(x) = P\{X < x_0\}$$

En consecuencia,

$$F(x_0) - \lim_{x\to x_0^-} F(x) = P\{X \leq x_0\} - P\{X < x_0\} = P\{X = x_0\}$$

En particular, F será continua por la izquierda en x_0 (y por lo tanto continua en x_0) si y sólo si $P\{X = x_0\} = 0$.

iv) Es análoga tomando sucesiones crecientes (decrecientes) tales que $x_n \to \pm\infty$. □

Observación 4.1.9 *Es posible probar que estas propiedades caracterizan a las funciones de distribución, en el sentido de que cualquier función F con estas propiedades será la función de distribución de alguna variable aleatoria X. (ver la observación 4.4.4)*

Observación 4.1.10 *Es útil observar que como consecuencia de estas propiedades, los puntos de discontinuidad de una función de distribución son a lo sumo numerables. (Esto se prueba observando que para cada k, sólo puede haber a lo sumo k puntos donde el salto de la función de distribución sea mayor que $1/k$).*

4.2. La integral de Riemann-Stieltjes y la definición de esperanza

La integral de Riemann-Stieltjes es una generalización de la integral de Riemann. Stieltjes observó que cualquier función creciente $F : \mathbb{R} \to \mathbb{R}$ origina una noción de medida de intervalos,

$$m_F((a, b]) = F(b) - F(a)$$

Para las aplicaciones a la teoría de probabilidades, nos interesa el caso en que F es la función de distribución de una variable aleatoria.

Stieltjes definió la integral

$$\int_a^b \varphi(x)\, dF(x) \tag{4.2}$$

generalizando la definición de la integral de Riemann de la siguiente manera: sea

$$\pi : a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

una partición del intervalo $(a, b]$ (Dar una partición no es otra cosa que elegir finitos puntos del intervalo en orden creciente) y elijamos puntos intermedios $\xi_i \in (x_i, x_{i+1}]$ en cada intervalito de la partición (En realidad, estamos trabajando con particiones con puntos marcados, pero no lo haremos explícito en la notación). Consideramos entonces las sumas de Riemann-Stieltjes

$$S_\pi(\varphi, F) = \sum_{i=0}^{n-1} \varphi(\xi)(F(x_{i+1}) - F(x_i))$$

Definición 4.2.1 *Diremos que la integral (4.2) existe y toma el valor $I \in \mathbb{R}$ si las sumas $S_\pi(\varphi, F)$ tienden al valor I cuando la norma*

$$|\pi| = \max_{0 \leq i \leq n-1} |x_{i+1} - x_i|$$

de la partición π tiende a cero, es decir si dado $\varepsilon > 0$, existe $\delta > 0$ tal que $|I - S_\pi(\varphi, F)| < \varepsilon$ para toda partición π con $|\pi| < \delta$.

Observemos que si $F(x) = x$, la integral de Riemann-Stieltjes se reduce a la integral de Riemann usual.

Algunas propiedades de la integral que son consecuencias más o menos inmediatas de las definiciones:

Lema 4.2.2 (Linealidad) *1. Si $\int_a^b \varphi_1(x)dF(x)$ y $\int_a^b \varphi_2(x)dF(x)$ existen, y $\varphi = \lambda_1\varphi_1 + \lambda_2\varphi_2$ entonces, $\int_a^b \varphi(x)\, dF(x)$ también existe, y tenemos que:*

$$\int \varphi(x)\, dF(x) = \lambda_1 \int_a^b \varphi_1(x)\, dF(x) + \lambda_2 \int_a^b \varphi_2(x)\, dF(x)$$

2. Si $\int_a^b \varphi(x)\, dF_1(x)$ y $\int_a^b \varphi(x)\, dF_2(x)$ existen, y $F = \lambda_1 F_1 + \lambda_2 F_2$ con $\lambda_1, \lambda_2 \geq 0$, entonces $\int_a^b \varphi(x)\, dF$ existe, y vale que:

$$\int_a^b \varphi(x)\, dF(x) = \lambda_1 \int_a^b \varphi(x)\, dF_1(x) + \lambda_2 \int_a^b \varphi(x)\, dF_2(x)$$

Lema 4.2.3 (Aditividad respecto al intervalo) *Sea $c \in (a, b]$. Si $\int_a^b \varphi(x)\, dF(x)$ existe, entonces también existen $\int_a^c \varphi(x)\, dF(x)$ y $\int_c^b \varphi(x)\, dF(x)$ y se verifica:*

$$\int_a^b \varphi(x)\, dF(x) = \int_a^c \varphi(x)\, dF(x) + \int_c^b \varphi(x)\, dF(x)$$

El siguiente teorema nos da una condición que permite garantizar la existencia de integrales de Riemann-Stieltjes:

Teorema 4.2.4 *Si $\varphi : [a, b] \to \mathbb{R}$ es continua, y si $F : [a, b] \to \mathbb{R}$ es creciente, entonces la integral de Riemann-Stieltjes*

$$\int_a^b \varphi(x)\, dF(x)$$

existe

Para la prueba, veáse el apéndice F.

El siguiente lema, nos dice cómo acotar una integral de Stieltjes:

Lema 4.2.5 *Supongamos que $\int_a^b \varphi(x)\, dF(x)$ existe, siendo φ una función acotada en $[a,b]$ y F creciente en $[a,b]$. Entonces,*

$$\left| \int_a^b \varphi(x)\, dF(x) \right| \leq \left(\sup_{x\in[a,b]} |\varphi(x)| \right) (F(b) - F(a))$$

Obs: Más generalmente se puede demostrar que la integral de Riemann-Stieltjes

$$\int_a^b \varphi(x)\, dF(x)$$

existe si $\varphi(x)$ es continua en $[a,b]$ y F es de variación acotada (ya que toda función de variación acotada se puede escribir como diferencia de dos funciones crecientes). En este caso, la integral se acota del siguiente modo:

$$\left| \int_a^b \varphi(x)\, dF(x) \right| \leq \left(\sup_{x\in[a,b]} |\varphi(x)| \right) V_a^b(F)$$

4.3. La definición de Esperanza

Veamos como se aplican las integrales de Riemann-Stieltjes a la teoría de probabilidades. Para ello consideremos una variable aleatoria, $X : \Omega \to \mathbb{R}$ no discreta y veamos como podríamos definir la esperanza de X. Supongamos por simplicidad primero que X toma valores en un cierto intervalo $(a,b]$ de la recta.

Entonces, si tomamos una partición π del intervalo $(a,b]$ (con puntos marcados como antes), podemos considerar una variable aleatoria X_π que aproxima a X del siguiente modo:

$$X_\pi = \xi_i \text{ si } X \in (x_i, x_{i+1}]$$

Entonces:

$$E[X_\pi] = \sum_{i=0}^{n-1} \xi_i \cdot P\{X_\pi = \xi_i\} = \sum_{i=0}^{n-1} \xi_i \cdot P\{x_i < X \leq x_{i+1}\}$$

$$= \sum_{i=0}^{n-1} \xi_i \cdot (F(x_{i+1}) - F(x_i))$$

es exactamente la suma de Riemann-Stieltjes $S_\pi(\varphi, F)$ con $\varphi(x) = x$.

Entonces cuando la norma de la partición tiende a cero, $E[X_\pi]$ tiende a la integral

$$\int_a^b x\, dF(x)$$

(que de acuerdo al teorema anterior siempre existe), y podemos aceptar la siguiente definición:

Definición 4.3.1 *Sea X una variable aleatoria que tome valores en un intervalo $[a, b]$ de la recta, entonces la esperanza de X es la integral de Riemann-Stieltjes*

$$E[X] = \int_a^b x dF(x) \tag{4.3}$$

siendo $F = F_X$ su función de distribución.

Más generalmente podemos considerar la variable aleatoria $\varphi(x)$ siendo $\varphi : \mathbb{R} \to \mathbb{R}$ una función continua, entonces:

$$\begin{aligned} E[\varphi(X_\pi)] &= \sum_{i=0}^{n-1} \varphi(\xi_i) \cdot P\{X_\pi = \xi_i\} \\ &= \sum_{i=0}^{n-1} \varphi(\xi_i) \cdot P\{\xi_i < X \leq \xi_{i+1}\} \\ &= \sum_{i=0}^{n-1} \varphi(\xi_i) \cdot (F(x_{i+1}) - F(x_i)) \end{aligned}$$

Entonces, cuando la norma de la partición π tiende a cero, estas sumas convergen a la integral:

$$\int_a^b \varphi(x)\ dF(x)$$

y conjeturamos que

$$E[\varphi(X)] = \int_a^b \varphi(x)\ dF(x) \tag{4.4}$$

para toda función continua $\varphi \in C[a, b]$ (aunque demostrar esto directamente de la definición es bastante complicado).

En particular,

$$\mathrm{Var}(X) = E[(X - \mu)^2] = \int_a^b (x - \mu)^2\ dF(x)$$

siendo $\mu = E[X]$.

Veamos algunos ejemplos, para familiarizarnos con esta idea:

Ejemplo 1: Para $x_0 \in \mathbb{R}$, definimos la **función escalón de Heaviside**:

$$H_{x_0}(x) = \begin{cases} 0 & \text{si} \quad x < x_0 \\ 1 & \text{si} \quad x \geq x_0 \end{cases}$$

H_{x_0} es la función de distribución de una variable aleatoria X que toma el valor x_0 con probabilidad 1. Entonces tenemos:

Lema 4.3.2 *Si $x_0 \in [a,b]$ y $\varphi \in C[a,b]$, entones:*

$$\int_a^b \varphi(x) dH_{x_0} = \varphi(x_0)$$

Prueba: En $S_\pi(\varphi, F)$ el único término no nulo corresponde al intervalo $[x_i, x_{i+1}]$ que contiene a x_0, en consecuencia:

$$S_\pi(\varphi, F) = \varphi(\xi_i)$$

y cuando $|\pi| \to 0$, $\varphi(\xi_i) \to \varphi(x_0)$, por la continuidad de φ. □

Luego $E[\varphi(X)] = \varphi(x_0)$.

Ejemplo 2: Variables aleatorias discretas

Si X es una función de distribución de una variable discreta que toma finitos valores $x_1, x_2, \ldots, x_n$ con probabilidad $p_i = P\{X = x_i\}$, tenemos que:

$$F(x) = \sum_{i=1}^{n} p_i \; H_{x_i}(x)$$

En consecuencia, por la linealidad de la integral de Riemann-Stieltjes respecto a F:

$$E[\varphi(X)] = \int_a^b \varphi(x) dF(x) = \sum_{i=0}^{n} p_i \int_a^b \varphi(x) dH_{x_i} = \sum_{i=1}^{n} p_i \varphi(x_i)$$

(donde $a \leq x_i \leq b \; \forall \; i$). Este resultado coincide con la fórmula anteriormente vista para $E[\varphi(X)]$ para variables discretas.

Ejemplo 3: Variables aleatorias absolutamente continuas Supongamos que X es una variable aleatoria continua, que tiene la densidad $f(x)$. Queremos calcular $E[X]$. Para ello, resultará útil el siguiente lema:

Lema 4.3.3 *Supongamos que $F : [a,b] \to \mathbb{R}$ es una función creciente con derivada continua $F'(x) = f(x)$, entonces*

$$\int_a^b \varphi(x) \; dF(x) = \int_a^b \varphi(x) \; f(x) \; dx$$

para toda función $\varphi \in C[a,b]$.

Prueba: Por el teorema del valor medio, $F(x_{i+1}) - F(x_i) = f(\xi_i)(x_{i+1} - x_i)$ para cierto $\xi_i \in (x_i, x_{i+1})$. Entonces, con esta elección de los puntos intermedios, la suma S_π se puede escribir como

$$S_\pi = \sum_{i=0}^{n-1} \varphi(\xi_i) f(\xi_i)(x_{xi+1} - x_i)$$

y vemos que cuando la norma de la partición π tiende a cero, tiende a la integral de Riemann

$$\int_a^b \varphi(x)\ f(x)\ dx$$

□

En particular, podemos definir la esperanza de una variable aleatoria con densidad continua $f(x)$ por:

$$E[X] = \int_a^b x\ f(x)dx$$

y más generalmente,

$$E[\varphi(X)] = \int_a^b \varphi(x)\ f(x)\ dx$$

En particular:

$$\mathrm{Var}(X) = E[(x-\mu)^2] = \int_a^b (x-\mu)^2 dx$$

siendo $\mu = E[X]$.

Un ejemplo: Si consideramos X una variable con distribución uniforme en el intervalo $[a, b]$ entonces su densidad es:

$$f(x) = \frac{1}{b-a}$$

Con lo que

$$\mu = E(X) = \int_a^b xf(x)\ dx = \frac{a+b}{2}$$

y

$$\mathrm{Var}X = \int_a^b \left(x - \frac{a+b}{2}\right)^2\ f(x)\ dx = \frac{1}{12}(b-a)^2$$

¿Qué sucede si X no es una variable aleatoria acotada? En este caso debemos considerar integrales de Riemann-Stieltjes impropias, de la forma:

$$\int_{-\infty}^{\infty} \varphi(x)\ dF(x)$$

Naturalmente definimos esta integral, de la siguiente manera:

$$\int_{-\infty}^{\infty} \varphi(x)\ dF(x) = \lim_{a\to-\infty, b\to+\infty} \int_a^b \varphi(x)\ dF(x)$$

El problema es que este límite puede no existir. Si φ es no negativa, podemos decir que siempre existe, pero puede valer $+\infty$. Adoptaremos pues la siguiente definición.

Definición 4.3.4 *Sea $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria, y sea $F = F_X$ su función de distribución. Diremos que X tiene esperanza finita, o que X es integrable, si*

$$\int_{-\infty}^{\infty} |x| \; dF(x) < +\infty$$

En ese caso, definimos:

$$E[X] = \int_{-\infty}^{\infty} x \; dF(x) \tag{4.5}$$

Más generalmente, tenemos la fórmula[3]:

$$E[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(x) \; dF(x) \tag{4.6}$$

válida si

$$\int_{-\infty}^{\infty} |\varphi(x)| \; dF(x) < +\infty$$

análoga a la proposición 3.2.4. Y cuando X tiene una densidad continua,

$$E[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(x) \; f(x) \; dx$$

Ejemplo: Supongamos que X se distribuye según la densidad normal $N(\mu, \sigma^2)$. Entonces, haciendo el cambio de variable $y = \frac{x-\mu}{\sigma}$, econtramos que

$$\begin{aligned} E[X] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \; e^{-(x-\mu)^2/(2\sigma^2)} \; dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma y) \; e^{-y^2/2} \; dy \\ &= \mu \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \; dy \right] + \sigma \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y \; e^{-y^2/2} \; dy \right] = \mu \end{aligned}$$

[La segunda integral se anula, pues la densidad normal estándar es una función par]. Similarmente,

$$\mathrm{Var}(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 \; e^{-(x-\mu)^2/(2\sigma^2)} \; dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 y^2 \; e^{-y^2/2} \; dy$$

Para calcular esta integral, observamos que:

$$\left(e^{-y^2/2} \right)' = (-y) e^{-y^2/2}$$

[3] Sin embargo es complicado justificar esto directamente a partir de la definición (4.5), pues no es sencillo en general establecer cuál es la relación general entre las funciones de distribución $F_{\varphi(X)}$ y F_X. En la observación 4.4.3 consideraremos el caso de un cambio de variable estrictamente creciente y biyectivo. Una justificación rigurosa de su validez en general se da en el apéndice D, pero utilizando herramientas de la teoría de la integral de Lebesgue.

e integramos por partes, deducimos que:

$$\mathrm{Var}(X) = \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy = \sigma^2$$

Este ejemplo aclara el significado de los parámetros de la distribución normal.

Ejercicio: Se dice que la variable aleatoria tiene distribución exponencial $\mathrm{Exp}(\lambda)$ (donde $\lambda > 0$) cuando su densidad de probabilidad es

$$f_X(x) = \lambda e^{-\lambda x} I_{(0,+\infty)}(x) \tag{4.7}$$

Demostrar que entonces

$$E(X) = \frac{1}{\lambda} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Un ejemplo de una variable aleatoria que no es continua ni discreta: Sea X una variable aleatoria con distribución uniforme en el intervalo $[0, 1]$ y consideramos $Y = \mathrm{máx}(X, 1/2)$, entonces:

$$Y = \begin{cases} 1/2 & \text{si} \quad X \leq 1/2 \\ X & \text{si} \quad X > 1/2 \end{cases}$$

Calculemos la función de distribución de Y:

$$F_Y(x) = P\{Y \leq x\} = P\{X \leq x \ \wedge \ 1/2 \leq x\}$$

Deducimos que:

$$F_Y(x) = \begin{cases} P(\emptyset) = 0 & \text{si} \quad x < 1/2 \\ P\{X \leq x\} = x & \text{si} \quad 1/2 \leq x \leq 1 \\ 1 & \text{si} \quad x > 1 \end{cases}$$

Deducimos que Y no es una variable discreta ya que F_Y no es una función escalera, y que tampoco Y es una variable absolutamente continua ya que F_Y no es continua.

Calculemos la esperanza de Y, esto puede hacerse de varias formas, por ejemplo usando la aditividad con respecto al intervalo de integración:

$$E[Y] = \int_0^1 x\, dF(x) = \int_0^{1/2} x\, dF + \int_{1/2}^1 x\, dF$$

En el intervalo cerrado $[0, 1/2]$ la función F coincide con la función $\frac{1}{2} H_{1/2}$ en consecuencia:

$$\int_0^{1/2} x\, dF = \frac{1}{2} \int_0^{1/2} x\, dH_{1/2} = \frac{1}{4}$$

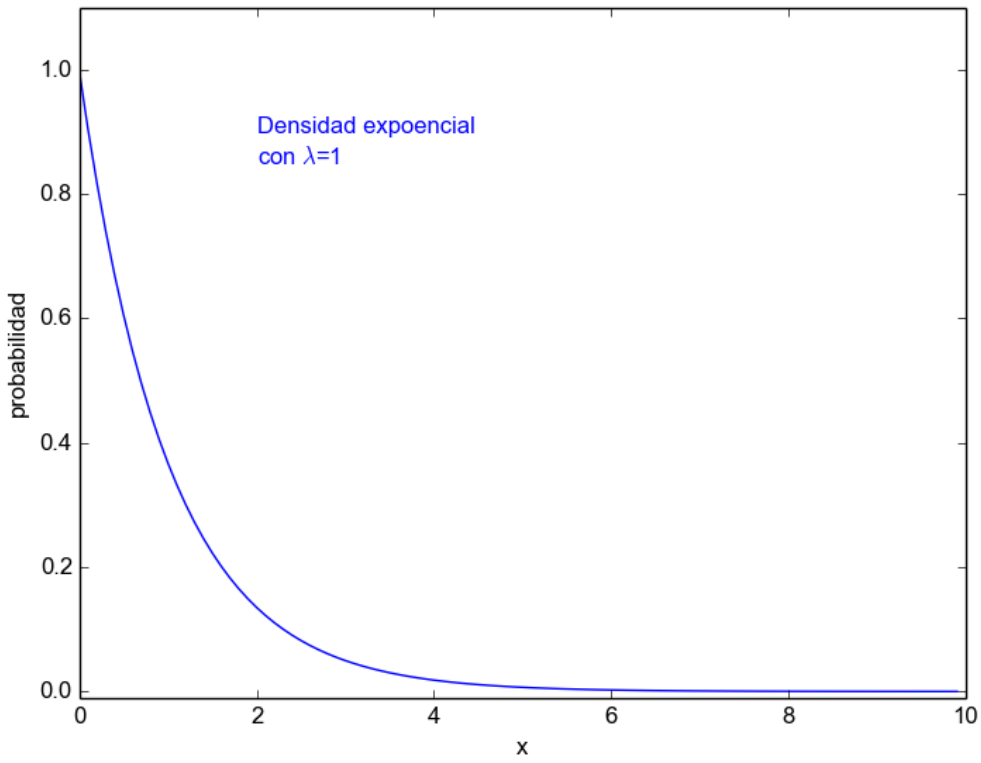


Figura 4.4: La densidad exponencial con $\lambda = 1$ (gráfico de la función exponencial).

mientras que:

$$\int_{1/2}^{1} x \, dF(x) = \int_{1/2}^{1} x dx = \frac{1}{2} - \frac{1}{8} = \frac{3}{8}$$

pues en $[1/2, 1]$ la función $F(x)$ tiene derivada continua $F'(x) = 1$. Concluimos que:

$$E[Y] = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$$

Otra manera de hacer la cuenta es considerar la función de variable real $\varphi(x) = \text{máx}(x, 1/2)$ y utilizar la fórmula para $E[\varphi(X)]$:

$$E[\varphi(X)] = \int_0^1 \text{máx}(x, 1/2) \, dx = \int_0^{1/2} 1/2 \, dx + \int_{1/2}^1 x \, dx = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$$

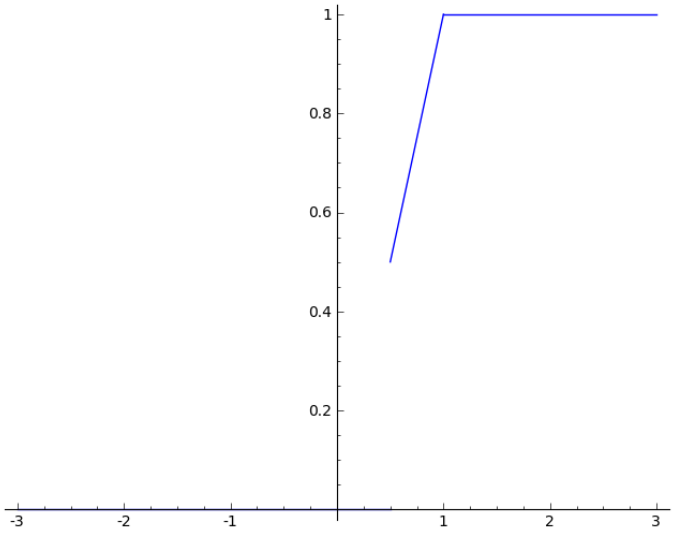


Figura 4.5: La función de distribución F_Y en este ejemplo

Ejercicio: Supongamos que $Z = \text{mín}(X, 1/2)$ donde X tiene distribución uniforme en $[0, 1]$. Determinar la función de distribución F_Z y la esperanza $E(Z)$.

4.4. Cambios de variables unidimensionales

Consideremos primero un cambio de variable de la forma $Y = \varphi(X)$ donde $\varphi : \mathbb{R} \to \mathbb{R}$ es una función biyectiva y estrictamente creciente.

Entonces podemos facilmente relacionar las funciones de distribución de X e Y

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\{\varphi(X) \leq y\} \\ &= P\{X \leq \varphi^{-1}(Y)(y)\} \\ &= F_X(\varphi^{-1}(y)) \end{aligned} \tag{4.8}$$

En particular (derivando con la regla de la cadena), se deduce que si X admite una densidad de probabilidad f_X de clase C^1, vemos que Y se distribuye según la densidad:

$$f_Y(y) = f_X(\varphi^{-1}(y))[\varphi^{-1}]'(y) \tag{4.9}$$

Ejemplo 4.4.1 *Supongamos que $X \sim N(\mu, \sigma^2)$ y hagamos un cambio de variable lineal, $Y = aX + b$ con $a > 0$. Esto corresponde a elegir*

$$\varphi(x) = ax + b \Rightarrow \varphi^{-1}(y) = \frac{y-b}{a}$$

Entonces según la fórmula (4.9), tenemos que

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

En particular en este ejemplo:

$$f_Y(y) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}\right\} = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-(a\mu+b))^2}{2(a\sigma)^2}\right\}$$

Concluimos que $Y \sim N(a\mu + b, a^2\sigma^2)$.

Ejemplo 4.4.2 (La distribución log-normal) *Supongamos que $X \sim N(\mu, \sigma^2)$. ¿Cuál es la distribución de $Y = e^X$?. Tomamos $\varphi(x) = e^x$, $\varphi : \mathbb{R} \to \mathbb{R}_{>0}$ es biyectiva y su inversa es $\varphi^{-1}(y) = \log y$. $\varphi^{-1} : \mathbb{R}_{>0} \to \mathbb{R}$. Recordamos que*

$$f_Y(y) = f_X(\varphi^{-1}(y)) \cdot (\varphi{-}1)'(y) = \frac{1}{\varphi'(x)} f_X(x) \text{ donde } x = \varphi^{-1}(y)$$

Como $\varphi(x) = e^{\log y} = y$, encontramos que

$$f_Y(y) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) \quad y > 0$$

Esta distribución se llama así porque si Y tiene distribución log-normal, entonces $\log Y = X$ tiene distribución normal.

Observación 4.4.3 *Como otra aplicación, podemos dar una justificación rigurosa de la fórmula (4.6) para el caso en que $\varphi : \mathbb{R} \to \mathbb{R}$ es biyectiva y estrictamente creciente. En efecto, en este caso, y llamamos $Y = \varphi(X)$, haciendo el cambio de variable $y = \varphi(x)$ en la integral de Stieltjes y teniendo en cuenta que entonces $F_Y(y) = F_X(x)$ por 4.8, obtenemos que:*

$$E[Y] = \int_{-\infty}^{\infty} y \, dF_Y(y) = \int_{-\infty}^{\infty} \varphi(x) \, dF_X(x)$$

La situación es bastante más compleja si admitimos cambios de variables que no son monótonos o biyectivos.

Consideremos por ejemplo el cambio de variable $Y = X^2$. Entonces para $z > 0$ tenemos que:

$$F_Y(y) = P\{X^2 \leq y\} = P\{|X| \leq \sqrt{y}\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = \\ P\{X \leq \sqrt{y}\} - P\{Z < -\sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y^-})$$

mientras que claramente $F_Y(y) = 0$ si $y < 0$.

En particular si X es una variable absolutamente continua con densidad f_X, encontramos (derivando como antes) que:

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] \quad (y > 0) \tag{4.10}$$

Observación 4.4.4 *Una aplicación importante de los cambios de variables es la* simulación *de distribuciones de probabilidad. En general, una computadora sabe generar números pseudo-aleatorios, que simulan la distribución uniforme en el intervalo* $[0, 1]$*. Si queremos generar a partir de ellos números pseudo-aleatorios que simulen la distribución* F*, se nos plantea el problema siguiente*

> *Dada una función de distribución* $F : \mathbb{R} \to [0, 1]$ *(con las propiedades del lema 4.1.7) y si* $X \sim \mathcal{U}(0, 1)$*, ¿cómo .podemos obtener otra variable aleatoria con distribución* F *?*

Conforme a la fórmula 4.8, si F *es continua (sin saltos) y estrictamente creciente, podemos tomar* $Y = F^{-1}(X)$ *donde* F^{-1} *denota la inversa de* F*. Cuando* F *no cumple estas hipótesis, es posible hacer lo mismo, pero considerando la inversa generalizada de* F *definida por*

$$F^{-1}(y) = \min\{x \in \mathbb{R} : F(x) \geq y\}$$

4.5. Suma de variables aleatorias independientes

Nuestro siguiente objetivo será extender a variables no discretas la noción de independencia:

Definición 4.5.1 *Dos variables aleatorias* X *e* Y *se dicen independientes, cuando para todo* $a < b$ *y todo* $c < d$ *los eventos* $\{X \in (a, b]\}$ *e* $\{Y \in (c, d]\}$ *son independientes. Es decir (en virtud de la definición de eventos independientes), si vale que:*

$$P\{a < X \leq b, c < Y \leq d\} = P\{a < X \leq b\} \cdot P\{c < Y \leq d\}$$

Lema 4.5.2 *Sean X e Y variables aleatorias independientes con funciones de distribución F_X y F_Y. Entonces $Z = X + Y$ tiene la función de distribución*

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z - x)\, dF_Y(x)$$

Prueba: Aproximamos X por una variable aleatoria discreta X_π. Suponemos primero que X está concentrada en un intervalo $(a, b]$ y y consideramos una partición $\pi : x_0 = a < x_1 < \ldots < x_n = b$ de $(a, b]$ con puntos marcados $\xi_k \in (x_{k-1}, x_k]$. Definimos

$$X_\pi = \xi_k \text{ si } X \in (x_k, x_{k+1}]$$

Sea $Z_\pi = X_\pi + Y$.

$$\begin{aligned} F_{Z_\pi}(z) &= P\{Z_\pi \leq z\} = P\{X_\pi + Y \leq z\} \\ &= \sum_k P\{X_\pi + Y \leq z / X_\pi = \xi_k\} \cdot P\{X_\pi = \xi_k\} \\ &= \sum_k P\{Y \leq z - \xi_k\} \cdot P\{X_\pi = \xi_k\} \\ &= \sum_k F_Y(z - \xi_k) \cdot P\{x_k < X \leq x_{k+1}\} \\ &= \sum_k F_Y(z - \xi_k) \cdot [F_X(x_{k+1}) - F_X(x_k)] \end{aligned}$$

Esta es una suma de Riemann-Stieltjes y en el límite se obtiene el enunciado. □

Derivando obtenemos:

Corolario 4.5.3 *Sean X e Y variables aleatorias independientes. Si X es una variable continua con densidad f_X y Y es una variable aleatoria cualquiera con función distribución F_Y entonces $Z = X + Y$ es una variable continua con densidad*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - x)\, dF_Y(x)$$

Definición 4.5.4 *Sean $f, g : \mathbb{R} \to \mathbb{R}$ funciones integrables. Definimos su* **convolución** *$f * g$ de la siguiente manera:*

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)\, g(x - t)\, dt$$

Algunas Observaciones sobre la convolución:

1. La convolución es conmutativa:

$$f * g = g * f$$

   También es posible probar que es asociativa:

$$(f * g) * h = f * (g * h)$$

2. Si f y g son densidades de probabilidad, entonces $f * g$ también lo es.

3. Si f y g están soportadas en la semirrecta $[0, +\infty)$ (es decir: $f(t) = g(t) = 0$ si $t < 0$), entonces:

$$(f * g)(x) = \int_0^x f(t)\, g(x-t)\, dt$$

Corolario 4.5.5 *Sean X e Y variables aleatorias independientes. Si X es una variable continua con densidad f_X y Y es una variable aleatoria continua con densidad f_Y entonces $Z = X + Y$ es una variable continua con densidad dada por la convolución $f_X * f_Y$:*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-x) \cdot f_Y(y)\, dy$$

4.5.1. Suma de variables normales independientes

Proposición 4.5.6 *Si $X \sim N(0, \sigma_1^2)$ e $Y \sim N(0, \sigma_2^2)$ son variables aleatorias independientes, entonces $X + Y \sim N(0, \sigma_1^2 + \sigma_2^2)$*

Prueba: Aplicamos el corolario 4.5.5 con

$$f(x) = \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-x^2/(2\sigma_1^2)}, \quad g(x) = \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-x^2/(2\sigma_2^2)}$$

Entonces $X \sim f * g$, donde

$$\begin{aligned}(f * g)(x) &= \int_{-\infty}^{\infty} \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-t^2/(2\sigma_1^2)} \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-(x-t)^2/(2\sigma_2^2)}\, dt \\ &= \frac{1}{\sigma_1\sigma_2 2\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}A(x,t)\right\}\, dt\end{aligned}$$

donde

$$A(x,t) := \frac{t^2}{\sigma_1^2} + \frac{(x-t)^2}{\sigma_2^2}$$

Trabajemos con esta expresión, buscando completar el cuadrado:

$$\begin{aligned} A(x,t) &= \frac{t^2}{\sigma_1^2} + \frac{x^2 - 2xt + t^2}{\sigma_2^2} \\ &= t^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) - \frac{2xt}{\sigma_2^2} + \frac{x^2}{\sigma_2^2} \\ &= t^2\frac{\sigma^2}{\sigma_1^2\sigma_2^2} - \frac{2xt}{\sigma_2^2} + \frac{x^2}{\sigma_2^2} \end{aligned}$$

siendo $\sigma^2 = \sigma_1^2 + \sigma_2^2$.Luego

$$A(x,t) = \frac{\sigma^2}{\sigma_1^2\sigma_2^2}\left[t^2 - 2xt\frac{\sigma_1^2}{\sigma^2}\right] + \frac{x^2}{\sigma_2^2}$$

Y completando entonces el cuadrado:

$$A(x,t) = \frac{\sigma^2}{\sigma_1^2\sigma_2^2}\left[\left(t - x\frac{\sigma_1^2}{\sigma^2}\right)^2 - x^2\frac{\sigma_1^4}{\sigma^4}\right] + \frac{x^2}{\sigma_2^2}$$

o sea:

$$A(x,t) = \frac{\sigma^2}{\sigma_1^2\sigma_2^2}\left(t - x\frac{\sigma_1^2}{\sigma^2}\right)^2 + \left(\frac{1}{\sigma_2^2} - \frac{\sigma_1^2}{\sigma^2\sigma_2^2}\right)x^2$$

Pero

$$\frac{1}{\sigma_2^2} - \frac{\sigma_1^2}{\sigma^2\sigma_2^2} = \frac{\sigma^2 - \sigma_1^2}{\sigma^2\sigma_2^2} = \frac{\sigma_2^2}{\sigma^2\sigma_2^2} = \frac{1}{\sigma^2}$$

Con lo que nos queda finalmente que

$$A(x,t) = \frac{\sigma^2}{\sigma_1^2\sigma_2^2}\left(t - x\frac{\sigma_1^2}{\sigma^2}\right)^2 + \frac{1}{\sigma^2}x^2$$

Sustituyendo

$$(f * g)(x) = \frac{1}{\sigma_1\sigma_2 2\pi}\exp\left(-\frac{x^2}{2\sigma^2}\right)\int_{-\infty}^{\infty}\exp\left\{-\frac{\sigma^2}{2\sigma_1^2\sigma_2^2}\left(t - x\frac{\sigma_1^2}{\sigma^2}\right)^2\right\}\,dt$$

Sólo nos falta pues calcular la integral,

$$I(x) = \int_{-\infty}^{\infty}\exp\left\{-\frac{\sigma^2}{2\sigma_1^2\sigma_2^2}\left(t - x\frac{\sigma_1^2}{\sigma^2}\right)^2\right\}\,dt$$

pero haciendo el cambio de variable

$$u = t - x\frac{\sigma_1^2}{\sigma^2}$$

vemos que no depende en realidad de x, y es

$$I(x) = \int_{-\infty}^{\infty} \exp\left\{-\frac{\sigma^2}{2\sigma_1^2\sigma_2^2}u^2\right\}\, du$$

Y haciendo un último cambio de variable

$$v = \frac{\sigma}{\sigma_1\sigma 2}u$$

nos queda que

$$I(x) = \frac{\sigma_1\sigma_2}{\sigma}\int_{-\infty}^{\infty} \exp\left\{-\frac{v^2}{2}\right\}\, dv = \sqrt{2\pi}\,\frac{\sigma_1\sigma_2}{\sigma}$$

Reemplazando nos queda que

$$X + Y \sim (f * g)(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Es decir, que $X + Y \sim N(0, \sigma^2)$. □

Nota: Otra manera de demostrar este resultado sin hacer tantas cuentas aparece en [Eis17]. Es una demostración muy corta y elegante, pero utiliza las ideas del capítulo siguiente.

4.6. Las Distribuciones Gama

Definición 4.6.1 *Definimos la función gama de Euler por*

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\, e^{-x}\, dx \ (\alpha > 0) \tag{4.11}$$

Introducimos también la función Beta de Euler (íntimamente relacionada con la función gama), definida para $\alpha_1, \alpha_2 > 0$ por

$$B(\alpha_1, \alpha_2) = \int_0^1 (1-u)^{\alpha_1-1}\, u^{\alpha_2-1}\, du$$

4.6.1. Análisis de la convergencia de la integral que define la función gama

Para analizar la convergencia de la integral, la partimos en el 1

$$\int_0^\infty x^{\alpha-1}\, e^{-x}\, dx = \int_0^1 x^{\alpha-1}\, e^{-x}\, dx + \int_1^\infty x^{\alpha-1}\, e^{-x}\, dx$$

Cuando $x \leq 1$, acotamos usando $e^{-x} \leq 1$,

$$\int_0^1 x^{\alpha-1}\, e^{-x}\, dx \leq \int_0^1 x^{\alpha-1} dx = \lim_{r\to 0} \int_r^1 x^{\alpha-1} dx = \frac{1}{\alpha} \text{ si } \alpha > 0$$

(esta integral es impropia cuando $\alpha < 1$ pero converge).

Por otra parte para $x \geq 0$ y cualquier $k \in \mathbb{N}$ vale la acotación,

$$e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!} \geq \frac{x^k}{k!} \Rightarrow e^{-x} \leq \frac{k!}{x^k}$$

Luego si $0 < \alpha < k$ tenemos

$$\int_1^\infty x^{\alpha-1}\, e^{-x}\, dx \leq \int_1^\infty x^{\alpha-1}\, \frac{k!}{x^k}\, dx = k! \int_1^\infty x^{\alpha-k-1} = \frac{k!}{k-\alpha}$$

Como para cada $\alpha > 0$ podemos elegir un k de modo que $\alpha < k$, deducimos que la integral converge para todo $\alpha > 0$.

Incluso sería posible considerar valores complejos de α, siempre que $\text{Re}(\alpha) > 0$, pero para los fines de esta materia nos bastará considerar valores reales de α.

4.6.2. Propiedades de la función gama

Proposición 4.6.2 *La función gamma tiene las siguientes propiedades:*

1. $\Gamma(1) = 1$
2. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
3. $\Gamma(k) = (k-1)!$ *(En consecuencia, la función gama puede pensarse como una generalización del factorial a valores no enteros de la variable).*
4. $\Gamma(1/2) = \sqrt{\pi}$

Prueba:

La propiedad 1) es inmediata de la definición:

$$\Gamma(1) = \int_0^\infty e^{-x}\, dx = \lim_{R\to+\infty} \int_0^R e^{-x}\, dx = \lim_{R\to+\infty} 1 - e^{-R} = 1$$

La propiedad 2) se prueba integrando por partes:

$$\begin{aligned}
\Gamma(\alpha+1) &= \lim_{\substack{R\to+\infty\\ r\to 0+}} \int_r^R x^\alpha\, e^{-x}\, dx \\
&= \lim_{\substack{R\to+\infty\\ r\to 0+}} \int_r^R x^\alpha\, (-e^{-x})'\, dx \\
&= \lim_{\substack{R\to+\infty\\ r\to 0+}} -\, x^\alpha e^{-x}\big|_0^R + \int_r^R (x^\alpha)'\; e^{-x}\, dx \\
&= \lim_{\substack{R\to+\infty\\ r\to 0+}} -R^\alpha e^{-R} + r^\alpha e^{-r} + \int_r^R \alpha x^{\alpha-1}\, e^{-x}\, dx \\
&= \alpha \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx \\
&= \alpha\Gamma(\alpha)
\end{aligned}$$

La propiedad 3) $\Gamma(k) = (k-1)!$ se deduce entonces de las propiedades 1) y 2) por inducción

Si $k = 1$ ya vimos que vale. El paso inductivo es:

$$\Gamma(k+1) = k\Gamma(k) = k(k-1)! = k!$$

La propiedad 4) sale con un cambio de variable: $x = y^2 \Rightarrow y = \sqrt{x}, dx = 2ydy$

$$\Gamma(1/2) = \int_0^\infty x^{-1/2}\, e^{-x}\, dx \;= \int_0^\infty \frac{1}{y}\, e^{-y^2}\, 2ydy = 2\int_0^\infty e^{-y^2} dy = \sqrt{\pi}$$

□

4.6.3. Las distribuciones gama

La función gama nos será útil para definir una familia de distribuciones de probabilidad[4]:

Definición 4.6.3 *Decimos que X se distribuye según la distribución gama $\Gamma(\alpha,\lambda)$ (siendo $\alpha,\lambda > 0$) si su función de densidad de probabilidad es:*

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\lambda x} I_{(0,+\infty)}(x) \tag{4.12}$$

[4]También tiene importantes aplicaciones en otras ramas de la matemática como la teoría de números, y aparece en numerosas fórmulas como la del volumen de una bola n-dimensional.

Observación 4.6.4 *Haciendo el cambio de variable $y = \lambda x$ en (4.11), tenemos que*

$$\frac{\Gamma(\alpha)}{\lambda^\alpha} = \int_0^\infty y^{\alpha-1}\, e^{-\lambda y}\, dy \tag{4.13}$$

Se deduce que (4.12) es efectivamente una densidad de probabilidades. Más aún esta fórmula permite calcular fácilmente los momentos de las distribuciones gama: si $X \sim \Gamma(\alpha, \lambda)$, entonces

$$\begin{aligned}\mu_k(X) = E(X^k) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+k-1}\, e^{-\lambda x}\, dx \\ &= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\lambda^k} = \frac{\alpha(\alpha+1)\cdots(\alpha+k)}{\lambda^k}\end{aligned}$$

En particular, la esperanza y la variancia de la distribución gama son

$$E(X) = \mu_1(X) = \frac{\alpha}{\lambda} \tag{4.14}$$

y

$$Var(X) = E(X^2) - E(X)^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2} \tag{4.15}$$

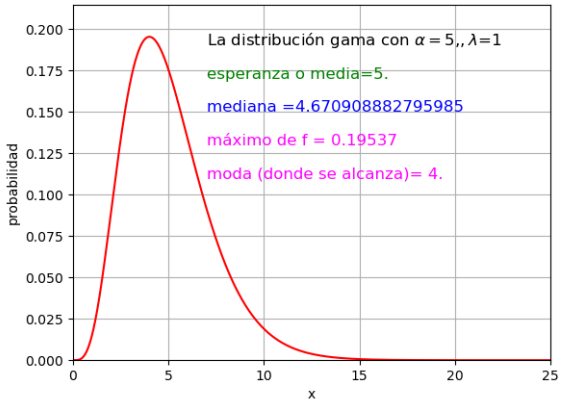


Figura 4.6: Un ejemplo de una densidad gama y algunos de sus parámetros.

Cálculo de los momentos de las distribuciones gama

$$E(X^n) = \int_0^\infty x^n \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+n)}{\lambda^{\alpha+n}} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\lambda^n}$$

En particular

$$E(X) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\lambda} = \frac{\alpha}{\lambda}$$

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}$$

Suma de variables independientes con distribución gama

Lema 4.6.5 *Si $X \sim \Gamma(\alpha_1, \lambda)$, $Y \sim \Gamma(\alpha_2, \lambda)$ y son independientes, entonces $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.*

Prueba: Según el corolario 4.5.5, $X+Y \sim f_{\alpha_1,\lambda} * f_{\alpha_2,\lambda}$. Hemos de calcular esta convolución:

$$\begin{aligned}(f_{\alpha_1,\lambda} * f_{\alpha_2,\lambda})(x) &= \int_0^x \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)}\,(x-t)^{\alpha_1-1}\,e^{-\lambda(x-t)}\frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)}\,t^{\alpha_2-1}\,e^{-\alpha t}\,dt \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\left(\int_0^x (x-t)^{\alpha_1-1}t^{\alpha_2-1}dt\right)e^{-\lambda x}\end{aligned}$$

En esta integral hacemos el cambio de variable $u = t/x$ $(0 \le x \le 1)$. Entonces:

$$\begin{aligned}(f_{\alpha_1,\lambda} * f_{\alpha_2,\lambda})(x) &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1))\Gamma(\alpha_2)}\,\left(\int_0^1 (x-xu)^{\alpha_1-1}\,(xu)^{\alpha_2-1}\,x\,du\right)\,e^{-\lambda x} \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}x^{\alpha_1+\alpha_2-1}\left(\int_0^1 (1-u)^{\alpha_1-1}\,u^{\alpha_2-1}\,du\right)e^{-\lambda x} \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\,B(\alpha_1,\alpha_2)\,x^{\alpha_1+\alpha_2-1}\,e^{-\lambda x}\end{aligned}$$

Notamos que esta es salvo la constante, la densidad gama $f_{\alpha_1+\alpha_2,\lambda}$, pero como la convolución de dos densidades de probabilidad es una densidad de probabilidad, y hay una única constante que hace que la integral sobre $(0, +\infty)$ dé 1 deducimos que:

$$f_{\alpha_1,\lambda} * f_{\alpha_2,\lambda} = f_{\alpha_1,\alpha_2,\lambda} \tag{4.16}$$

Como subproducto de la demostración obtenemos que:

$$\frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\,B(\alpha_1,\alpha_2) = \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}$$

o sea

$$B(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

□

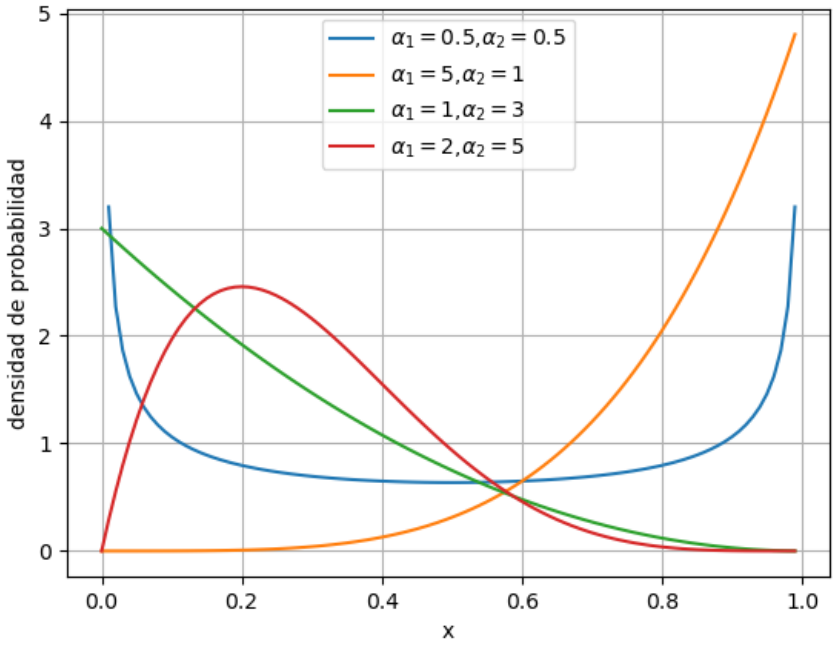


Figura 4.7: Las densidades beta

4.7. Las distribuciones Beta

La función beta también puede usarse para definir una familia de distribuciones: las *distribuciones beta*. Diremos que $X \sim \beta(\alpha_1, \alpha_2)$ si se distribuye según la densidad:

$$f_X(x) = \frac{1}{B(\alpha_1, \alpha_2)} \, x^{\alpha_1 - 1} \, (1-x)^{\alpha_2 - 1} \, I_{(0,1)}(x)$$

Podemos preuntarnos ¿cuánto valen la esperanza y la varianza de X? Más generalmente, podemos calcular los **momentos** de X.

$$\mu_k = E[X^k] = E[\varphi(X)] \; k \in \mathbb{N}$$

donde $\varphi(x) = x^k$. Entonces usando la ecuación funcional para la función gama:

$$\begin{aligned} E[X^k] &= \frac{1}{B(\alpha_1,\alpha_2)} \int_0^1 x^k \, x^{\alpha_1-1} \, (1-x)^{\alpha_2-1} \, dx \qquad = \frac{B(\alpha_1+k,\alpha_2)}{B(\alpha_1,\alpha_2)} \\ &= \frac{\Gamma(\alpha_1+k)\Gamma(\alpha_2)}{\Gamma(\alpha_1+k+\alpha_2)} \cdot \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\ &= \frac{\alpha^{(k)}}{(\alpha_1+\alpha_2)^{(k)}} = \prod_{r=0}^{k} \frac{\alpha_1+r}{\alpha_1+\alpha_2+r} \end{aligned}$$

donde

$$\alpha^{(k)} = \alpha \cdot (\alpha+1) \cdot \ldots \cdot (\alpha+k-1)$$

se llama el **símbolo de Pochhammer**.

En particular, para $k = 1$ vemos que la esperanza vale,

$$E[X] = \frac{\alpha_1}{\alpha_1+\alpha_2}$$

y el momento de segundo orden:

$$E[X^2] = \frac{\alpha_1(\alpha+1)}{(\alpha_1+\alpha_2)\cdot(\alpha_1+\alpha_2+1)}$$

Finalmente

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}$$

4.8. La Distribución Exponencial y la propiedad de Falta de Memoria

La distribución exponencial (4.7) es un modelo muy útil para distintos procesos: llamadas que llegan a una central telefónica, tiempo de duración de una lámpara, desintegración radiactiva, etc.

Por ejemplo, para fijar ideas, consideremos la desintegración radiactiva de un átomo. La hipótesis fundamental que haremos para describir este fenómeno, es la propiedad de "falta de memoria" que establece que la probabilidad de que un átomo se desintegre en un intervalo de tiempo de longitud Δt sólo depende de la longitud del intervalo y es independiente de la historia anterior del material.

Podemos describir con más precisión esta propiedad de la siguiente manera: Si llamamos T al tiempo en el que el átomo se desintegra, T es una variable aleatoria. La probabilidad condicional de que el átomo se desintegre en el intervalo $(t_0, t_0+\Delta t]$ sabiendo que no se ha desintegrado aún en tiempo $t = t_0$, es igual a la probabilidad de que se desintegre en el intervalo $(0, \Delta t]$:

$$P\{T > t_0 + \Delta t / T > t_0\} = P\{T > \Delta t\}$$

Por definición de probabilidad condicional, esto significa que:

$$\frac{P\{t < T \leq t + \Delta t\}}{P\{T > t\}} = P\{T > \Delta t\}$$

Llammemos F a la función de distribución de T, y sea $G(t) = 1 - F(t)$. Entonces, esta igualdad establece que:

$$G(t + \Delta t) = G(t)G(\Delta t)$$

Necesitaremos el siguiente lema:

Lema 4.8.1 *Sea $G : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ una función continua que satisface que:*

$$G(t + s) = G(t)G(s)$$

Entonces: $G(t) = G(0)a^t$, siendo $a = G(1)$.

Volviendo a nuestro problema de la desintegración radiactiva, si ponemos $G(1) = e^{-\lambda}$ (suponiendo $G(0) \neq 0$), y observamos que $G(0) = 1$ pues $T > 0$ (El átomo no se desintegró aún en $t = 0$), obtenemos que:

$$G(t) = e^{-\lambda t}$$

Por consiguiente la función de distribución de T es:

$$F(t) = 1 - e^{-\lambda t}$$

y derivando vemos que su densidad es

$$f(t) = \lambda\, e^{-\lambda t}\ (t > 0)$$

Decimos que la variable continua T se distribuye según la densidad exponencial de parámetro $\lambda > 0$, Exp(λ), que introdujimos en (4.7).

Supongamos ahora que tenemos un material radiactivo formado inicialmente por un grán número de átomos N_0, y llamemos $N(t)$ a la cantidad de átomos no desintegrados hasta el instante t. Hagamos la hipótesis de que las desintegraciones de los distintos átomos son independientes. Podemos pensar que son ensayos de Bernoullí, entonces por la ley de los grandes números

$$\frac{N(t)}{N_0} \approx P\{T > t_0\}$$

y deducimos que:

$$N(t) = N_0\, e^{-\lambda t} \tag{4.17}$$

Esta expresión se conoce como la ley de desintegración radiactiva de Rutherford-Soddy (1902). El valor de la constante λ depende de la sustancia.

Se define semivida o período de semi-desintegración $T_{1/2}$ el tiempo en que una muestra de material radiactivo tarda en reducirse a la mitad. De la fórmula (4.17), se deduce que

$$T_{1/2} = \frac{\log 2}{\lambda}$$

La siguiente tabla muestra por ejemplo los períodos de semi-desintegración de algunos isótopos radiactivos:

Isótopo	$T_{1/2}$
Berilio-8	$10^{-16}s$
Polonio-213	$4x10^{-6}s$
Aluminio-28	2.25 min
Yodo-131	8 días
Estroncio-90	28 años
Radio-226	1600 años
Carbono-14	5730 años
Rubidio-87	$5{,}7 \times 10^{10}$ años

Observación 4.8.2 *Entre las distribuciones discretas, la propiedad de falta de memoria es característica de la distribución geométrica, que puede entonces considerarse como el análogo discreto de la distribución exponencial.*

4.8.1. Tiempos de espera y procesos de Poisson

Llamemos T_i al tiempo en que ocurre la iésima densintegración radiactiva, de modo que:

$$T_1 < T_2 < \ldots < T_n$$

(Podemos suponer para simplificar que no hay dos desintegraciones simultáneas, ya que la probabilidad de que ello ocurra es despreciable). Notemos que:

$$T_n = T_1 + (T_2 - T_1) + (T_3 - T_2) + \ldots + (T_n - T_{n-1})$$

Las variables $T_k - T_{k-1}$ representan el tiempo entre la $(k-1)$-ésima desintegración y la k-ésima desintegración. Por la discusión anterior (y la propiedad de falta de memoria), $T_k - T_{k-1}$ tiene distribución exponencial de parámetro $\lambda > 0$ (donde $\lambda > 0$ es una constante que depende del material que estamos considerando).

Por otra parte, si suponemos que el tiempo que un átomo tarda en desintegrarse es independiente de lo que tardan los demás, las $T_{k+1} - T_k$ serán variables aleatorias independientes. Entonces la variable T_n será dada por una suma de n variables aleatorias independientes, todas con distribución exponencial de parámetro λ.

Como $\text{Exp}(\lambda) = \Gamma(1, \lambda)$, deducimos que T_n tiene distribución $\Gamma(n, \lambda)$, es decir que se distribuye según la densidad $g_n(t)$ dada por:

$$g_n(t) = \begin{cases} \frac{\lambda^n}{(n-1)!}\, t^{n-1}\, e^{-\lambda t} & \text{si} \quad t > 0 \\ 0 & \text{si} \quad t \leq 0 \end{cases}$$

Llamemos $D(t)$ al número de desintegraciones en el intervalo $[0, t]$. Entonces

$$D(t_0) = n \text{ si y sólo si } T_n \leq t_0 < T_{n+1}$$

Deducimos que:

$$\{D(t_0) = n\} = \{T_n \leq t_0\} - \{T_{n+1} \leq t_0\}$$

En consecuencia,

$$P\{D(t_0) = n\} = P\{T_n \leq t_0\} - P\{T_{n+1} \leq t_0\} = \int_0^{t_0} g_n(t)\, dt - \int_0^{t_0} g_{n+1}(t)\, dt$$

Integrando por partes, tenemos que:

$$\int_0^{t_0} g_{n+1}(t)\, dt = \int_0^{t_0} \frac{\lambda^{n+1}}{n!}\, t^n e^{-\lambda t}\, dt$$

$$= \frac{\lambda^{n+1}}{n!} \left[t^n\, \frac{e^{-\lambda t}}{(-\lambda)} \Big|_0^{t_0} - \int_0^{t_0} n\, t^{n-1}\, \frac{e^{-\lambda t}}{(-\lambda)}\, dt \right]$$

$$= \frac{\lambda^{n+1}}{n!}\, t_0^n \frac{e^{-\lambda t_0}}{(-\lambda)} - 0 - \int_0^{t_0} \frac{\lambda^{n+1}}{n!}\, n\, t^{n-1}\, \frac{e^{-\lambda t}}{(-\lambda)}\, dt$$

$$= -\frac{\lambda^n}{n!}\, t_0^n\, e^{-\lambda t_0} + \int_0^{t_0} \frac{\lambda^n}{(n-1)!}\, t^{n-1}\, e^{-\lambda t}\, dt$$

$$= -\frac{\lambda^n}{n!}\, t_0^n\, e^{-\lambda t_0} + \int_0^{t_0} g_n(t)\, dt$$

En definitiva concluimos que la distribución del número de desintegraciones viene dada por una distribución de Poisson (proceso de Poisson):

$$P\{D(t_0) = n\} = \frac{(\lambda t_0)^n}{n!}\, e^{-\lambda t_0}$$

Como dijimos al comienzo de la sección, aunque hemos presentado la distribución exponencial y este cálculo de los tiempos de espera como modelo de la desintegración radiactiva, este mismo modelo se puede aplicar a otros procesos donde la hipótesis de falta de memoria resulte razonable como por ejemplo la llegada de eventos a un servidor informático, o los siniestros en una compañía de seguros. Esto explica la utilización de las distribuciones exponencial y de Poisson en muchas aplicaciones de las probabilides.

4.9. Algunas densidades útiles en estadística

4.9.1. Las densidades χ^2

En esta sección veremos algunas densidades que resultan especialmente útiles en estadística. Nos proporcionarán ejemplos interesantes de las técnicas de cambio de variables.

Sea $X \sim N(0,1)$ una variable aleatoria con distribución normal estándar. Utilizando la fórmula (4.10), encontramos que $Y = X^2$ se distribuye según la densidad

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] = \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}}e^{-y/2} + \frac{1}{\sqrt{2\pi}}e^{-y/2}\right]$$

o sea

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}y^{-1/2}e^{-y/2} \quad (y > 0)$$

Esta densidad se conoce como la densidad χ^2 ("ji-cuadrado"] con un grado de libertad [abreviada χ^2_1]. Comparando con (4.12), y utilizando que $\Gamma(1/2) = \sqrt{\pi}$, vemos que coincide con la densidad $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$.

Sean ahora $X_1, X_2, \ldots, X_n$ variables aleatorias independientes con distribución normal estándar, y consideremos la variable aleatoria

$$Z_n = X_1^2 + X_2^2 + \ldots + X_n^2$$

¿cuál es la distribución de Z_n ? Por lo anterior cada una de las X_i se distribuye según la densidad $\chi^2_1 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, y la densidad de Z será (por la independencia) la convolución de la densidad $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$ n veces con sigo misma, que por el lema 4.6.5 da la densidad $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$. Es decir, que la densidad de Z_n será

$$f_{Z_n}(z) = \frac{(1/2)^{n/2}}{\Gamma(n/2)}x^{n/2-1}e^{-x/2} \quad (x > 0) \tag{4.18}$$

Esta densidad se conoce como densidad χ^2 con n grados de libertad [abreviada χ^2_n]. Las fórmulas (4.14) y (4.15) nos dicen que si $Z \sim \chi^2_n$, entonces

$$E[Z_n] = n, \quad \text{Var}[Z_n] = 2n$$

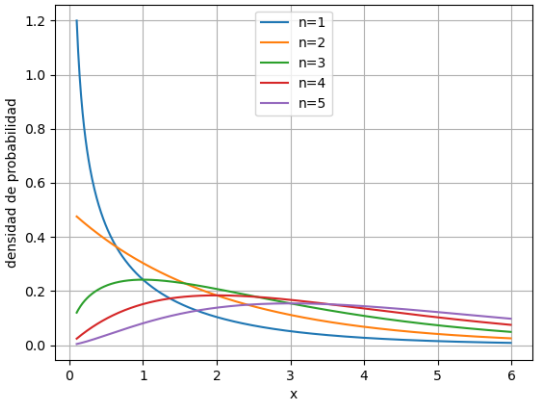


Figura 4.8: Gráfico de la densidad χ^2_n

4.9.2. Las densidades χ_n

Si consideramos el vector aleatorio $X = (X_1, X_2, \ldots, X_n)$ con las $X_i \sim N(0,1)$ independientes,

$$Z_n = \|X\| = \sqrt{Y_n}$$

entonces

$$\begin{aligned} f_{Z_n}(z) &= 2z\frac{(1/2)^{n/2}}{\Gamma(n/2)}\ (z^2)^{n/2-1} \cdot e^{-z^2/2} \\ &= \frac{2(1/2)^{n/2}}{\Gamma(n/2)} z^{n-1} \cdot e^{-z^2/2} \quad (z > 0) \end{aligned}$$

Esta distribución se lllama χ_n. Con $n = 3$ esta distribución aparece en física, como la **distribución de Maxwell-Boltzmann**, que es la distribución de probabilidad de las velocidades de un gas asociada a la estadística de Maxwell-Boltzmann para dicho sistema.

$$f(v) = 4\pi \left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} v^2 e^{\frac{-mv^2}{2kT}}$$

donde m es la masa de la partícula, T es la temperatura absoluta y k es una constante (constante de Boltzmann).

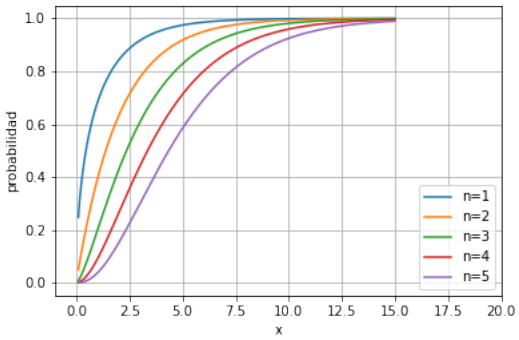


Figura 4.9: Gráfico de la distribución acumulada de una χ^2_n

Capítulo 5

Vectores Aleatorios

5.1. Vectores Aleatorios

Las ideas anteriores sobre variables aleatorias continuas, pueden generalizarse para considerar vectores aleatorios.

Definición 5.1.1 *Sea $(\Omega, \mathcal{E}, P)$ un espacio de probabilidad. Un* **vector aleatorio** *n-diemensional es una función $X : \Omega \to \mathbb{R}^n$ con la propiedad de que si $I = (a_1, b_1] \times (a_2, b_2] \times \ldots \times (a_n, b_n]$ es un intervalo de $\mathbb{R}^n$ entonces $X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in\} \in \mathcal{E}$, es decir está definida la probabilidad $P\{X \in I\}$ de que X pertenezca a I.*

Obsevación: Dar un vector aleatorio n-dimensional es equivalente a dar n variables aleatorias $X_1, X_2, \ldots, X_n$.

Ejemplos de vectores aleatorios:

1. Un ejemplo de vector aleatorio discreto es el que consideramos al describir la distribución multinomial (ver página 68).

2. Distribución uniforme en un conjunto $A \subset \mathbb{R}^n$ de medida positiva: si A es un conjunto de $\mathbb{R}^n$ de medida positiva y X es un vector aleatorio n-dimensional, decimos que X se distribuye uniformemente en A si X pertenece a A con probabilidad 1, y si

$$P\{X \in B\} = \frac{m(B)}{m(A)} \quad \forall\, B \subset A$$

   En esta definición A y B pueden ser conjuntos medibles Lebesgue cualesquiera, y $m(A)$ denota la medida de Lebesgue de A (Quienes no hayan cursado análisis real, pueden pensar que A y B son conjuntos para los que tenga sentido calcular la medida de A, por ejemplo que A y B son abiertos de $\mathbb{R}^2$ y $m(A)$ representa el área de A).

3. Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función integrable tal que $0 \leq f(x) \leq 1$, y

$$\int_{\mathbb{R}^n} f(x)\, dx = 1$$

Decimos que el vector X se distribuye según la **densidad conjunta** $f(x)$ si para cualquier conjunto medible $A \subset \mathbb{R}^n$, tenemos que:

$$P\{X \in A\} = \int_A f(x)\, dx$$

(De nuevo, quienes no hayan cursado análisis real pueden pensar que f es integrable en el sentido de Riemann, y A es cualquier abierto de $\mathbb{R}^n$).

4. Por ejemplo, una posible generalización de la distribución normal a dos dimensiones (normal bi-variada), se obtiene especificando que el vector (X, Y) se distribuye según la densidad conjunta:

$$f(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \tag{5.1}$$

Veremos más adelante que esta densidad corresponde al caso especial de dos variables aleatorias independientes con esperanza 0 y esperanza 1. Más generalmente, decimos que el vector aleatorio X tiene **distribución normal multivariada** si se distribuye según una densidad de la forma:

$$f(x) = ce^{-q(x)}$$

donde $q(x) = x^t Ax$ es una forma cuadrática definida positiva, y c es una constante elegida de modo que la integral de f sobre todo $\mathbb{R}^n$ dé 1. Más adelante volveremos sobre este concepto.

La noción de función de distribución puede generalizarse a vectores aleatorios.

Definición 5.1.2 *Si $X : \Omega \to \mathbb{R}^n$ es un vector aleatorio, su* **función de distribución conjunta** *es la función $F : \mathbb{R}^n \to \mathbb{R}$ dada por:*

$$F(x_1, x_2, \ldots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$$

Por ejemplo, si X es un vector aleatorio bidimensional que se distribuye según la densidad conjunta $f(x)$, entonces su función de distribución conjunta es:

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_n} f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)\, d\tilde{x}_1\, d\tilde{x}_2\, \ldots\, d\tilde{x}_n$$

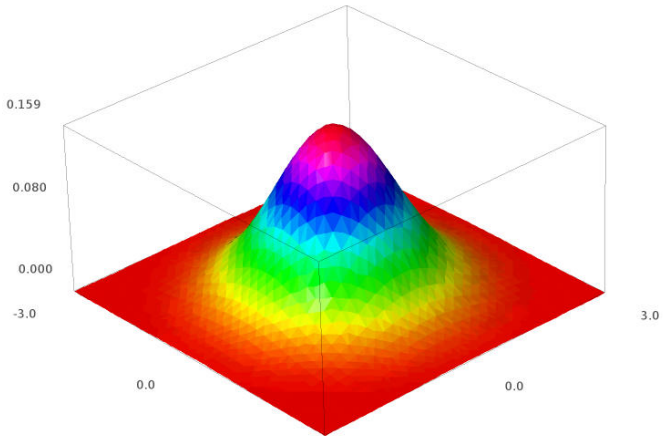


Figura 5.1: La función de densidad normal bivariada (para dos variables independientes con esperanza 0 y esperanza 1) dada por la ecuación (5.1).

La noción de función de distribución resulta más complicada que en el caso de variables aleatorias unidimensionales. En el caso unidimensional, la probabilidad de que la variable X tome un valor en el intervalo $(a, b]$ viene dada, en términos de la función de distribución F_X, por:

$$P\{X \in (a, b]\} = P\{X \leq b\} - P\{X \leq a\} = F_X(b) - F_X(a)$$

En cambio si (X, Y) es un vector aleatorio con función de distribución conjunta F, y $R = (a, b] \times (c, d]$ es un rectángulo (semiabierto) en $\mathbb{R}^2$, la probabilidad de que (X, Y) tome un valor en R es (por la fórmula de inclusiones y exclusiones):

$$\begin{aligned} P\{(X, Y) \in R\} = P\{X \leq b, Y \leq d\} - P\{X \leq a, Y \leq d\} \\ - P\{X \leq b, Y \leq c\} + P\{X \leq a, Y \leq c\} \end{aligned}$$

Es decir que:

$$P\{(X, Y) \in R\} = F(b, d) - F(a, d) - F(b, c) + F(a, c) := \Delta F(R) \tag{5.2}$$

(Esta cantidad es necesariamente no negativa, esta es la generalización bidimensional del hecho de que en el caso unidimensional la función de distribución es creciente.)

Una fórmula análoga (¡pero más complicada!) es cierta para vectores aleatorios en más dimensiones. Por ello, la noción de función de distribución no resultará tan útil como lo era en el caso unidimensional (y con frecuencia resulta más cómodo pensar directamente en términos de probabilidades asignadas a rectángulos, o subconjuntos más generales de $\mathbb{R}^n$).

5.2. Densidades y distribuciones marginales

Consideramos para simplificar la notación, un vector aleatorio bidimensional (X, Y). Investiguemos qué relación existe entre la función de distribución conjunta F del vector (X, Y) y las funciones de distribución F_X y F_Y de cada variable por separado:

Notemos que:

$$F_X(x) = P\{X \leq x\} = P\{X \leq x, Y \leq +\infty\} = F(x, +\infty) = \lim_{y \to +\infty} F(x, y)$$

Similarmente,

$$F_Y(y) = \lim_{x \to +\infty} F(x, y)$$

F_X y F_Y se conocen como las **funciones de distribución marginales** del vector aleatorio (X, Y).

Consideremos ahora el caso particular, en que el vector aleatorio (X, Y) se distribuye según la densidad conjunta $f(x, y)$, su función de distribución será entonces:

$$F(x_0, y_0) = P\{X \leq x_0, Y \leq y_0\} = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f(x, y)\, dx\, dy$$

y en consecuencia sus funciones de distribución marginales vendrán dadas por:

$$F_X(x_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy$$

$$F_Y(y_0) = \int_{-\infty}^{+\infty} \int_{-\infty}^{y_0} f(x, y)\, dx\, dy$$

Utilizando el teorema de Fubini, podemos escribir F_X como una integral reiterada:

$$F_X(x_0) = \int_{-\infty}^{x_0} \left(\int_{-\infty}^{\infty} f(x, y)\, dy \right) dx$$

Esta igualdad significa que el vector aleatorio X se distribuye según la densidad:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \tag{5.3}$$

Similarmente, el vector aleatorio Y se distribuye según la densidad:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \tag{5.4}$$

f_X y f_Y se conocen como las **densidades marginales** de probabilidad del vector aleatorio (X, Y).

Ejemplo 5.2.1 *Antes consideramos un vector aleatorio (X, Y) que se distribuía según la densidad conjunta (5.1). Entonces, en este caso*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\, dy \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \end{aligned}$$

Luego $X \sim N(0,1)$. Similarmente, por simetría,

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

y también $Y \sim N(0,1)$.

5.3. Esperanza de funciones de vectores aleatorios. Covariancia

Sea (X, Y) un vector aleatorio bidimensional, y $\varphi : \mathbb{R}^2 \to \mathbb{R}$ una función continua. La fórmula (4.6) para la esperanza de una función de una variable aleatoria puede generalizarse a vectores aleatorios:

$$E[\varphi(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x,y)\, dF(x,y) \tag{5.5}$$

donde la integral que aparece en el segundo miembro es una integral doble de Riemann-Stieltjes.

Para definir este concepto puede procederse como en análisis II, considerando primero la integral

$$\int_a^b \int_c^d \varphi(x,y)\, dF(x,y) \tag{5.6}$$

en un rectángulo $R = (a, b] \times (c, d]$ de $\mathbb{R}^2$. Consideramos una partición π del rectángulo R en rectángulos más pequeños $R_{ij} = (x_i, x_{i+1}] \times (y_j, y_{j+1}]$, definida por una partición π_x del intervalo $[a, b]$:

$$a = x_0 < x_1 < \ldots < x_M = b$$

y otra partición π_y del intervalo $[c, d]$:

$$a = y_0 < y_1 < \ldots < y_N = b$$

Elegimos puntos intermedios $\xi_i \in [x_i, x_{i+1}]$ y $\eta_j \in [y_j, y_{j+1}]$, y consideramos sumas de Riemann-Stieltjes dobles:

$$S_\pi(\varphi, F) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \varphi(\xi_i, \eta_j) \Delta F(R_{ij})$$

siendo

$$\Delta F(R_{ij}) = F(x_{i+1}, y_{j+1}) - F(x_i, y_{j+1}) - F(x_{i+1}, y_j) + F(x_i, y_j)$$

que de acuerdo a la fórmula (5.2), representa la probabilidad de que el vector (X, Y) tome un valor en el rectángulo R_{ij}.

Definamos la norma $|\pi|$ de la partición π como el máximo de las normas de las particiones π_x y π_y. Entonces si, cuando la norma de la partición π tiende a cero, las sumas $S(\pi, F)$ convergen a un número I, diremos que la integral (5.6) existe, y que toma el valor I. Análogamente a lo que sucede en el caso unidimensional, podemos demostrar que esto sucede si F es la función de distribución de un vector aleatorio, y φ es continua.

La intergral impropia, sobre todo el plano, que aparece en la fórmula (5.5) puede definirse como el límite de integrales sobre rectángulos:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y)\; dF(x,y) = \lim_{\substack{a,c\to-\infty;b\\ d\to+\infty}} \int_a^b \int_c^d \varphi(x,y)\; dF(x,y)$$

Para justificar intuitivamente la fórmula (5.5) podemos proceder como en el caso discreto, definiendo variables aleatorias discretas X_π e Y_π que aproximan a X e Y por:

$$X_\pi = \xi_i \text{ si } X \in (x_i, x_{i+1}]$$

$$Y_\pi = \eta_j \text{ si } Y \in (y_j, y_{j+1}]$$

y observando que:

$$E[\varphi(X_\pi, Y_\pi)] = S_\pi(\varphi, F)$$

Por lo que cuando la norma de la partición π tiende a cero, obtenemos formalmente la fórmula (5.5).

El caso que más nos va a interesar, es cuando el vector aleatorio (X, Y) se distribuye según una densidad conjunta $f(x, y)$. En este caso, como ocurría en el caso unidimensional, la esperanza de $\varphi(X, Y)$ puede calcularse mediante una integral de Riemann ordinaria, en lugar de una integral de Riemann-Stieltjes:

$$E[\varphi(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y)\; f(x,y)\; dx\; dy \tag{5.7}$$

Un caso importante de aplicación de las fórmulas anteriores es cuando queremos calcular la covarianza de dos variables aleatorias en el caso continuo. Recordamos que por definición:

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

siendo $\mu_X = E[X]$, $\mu_Y = E[Y]$. Entonces tomando $\varphi(x, y) = (x - \mu_X)(y - \mu_Y)$ en las fórmulas anteriores, tenemos que:

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\,dF(x,y)$$

en el caso , y

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\,f(x,y)\,dx\,dy$$

si el vector (X,Y) admite una densidad conjunta.

Ejemplo 5.3.1 *Volvamos a considerar el ejemplo de un vector aleatorio (X,Y) que se distribuía según la densidad conjunta (5.1). Ya vimos que $X,Y \sim N(0,1)$ por lo que $\mu_X = \mu_Y = 0$. Calculemos*

$$\begin{aligned}
Cov(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,f(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot y\cdot\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dx\,dy \\
&= \left(\int_{-\infty}^{\infty} x\cdot\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx\right)\cdot\left(\int_{-\infty}^{\infty} y\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,dy\right) \\
&= E[X]\cdot E[Y] = 0
\end{aligned}$$

Observación 5.3.2 *Una de las propiedade más básicas de la esperanza es su linealidad. Sin embargo, es difícil justificar su validez en general partiendo de la definición 4.5, ya que la función de distribución F_X no depende linealmente de la variable X. Utilizando la fómula (5.7), podríamos sin embargo dar una justificación de que $E[X+Y] = E[X]+E[Y]$ para el caso en que X e Y tienen una densidad conjunta continua y esperanza finita*[1]*. En efecto, en este caso, tomando $\varphi(x,y) = x+y$. vemos que*

$$\begin{aligned}
E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)\cdot f(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot f(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\cdot f(x,y)\,dx\,dy \\
&= \int_{\infty}^{\infty} x\cdot f_X(x)\,dx + \int_{\infty}^{\infty} y\cdot f_Y(y)\,dy \\
&= E[X]+E[Y].
\end{aligned}$$

[1]Esta propiedad es válida en general, como se deduce inmediatamente de la interpretación de la esperanza como una integral de Lebesgue, ver apéndice D.

5.4. Cambios de variable n-dimensionales

Proposición 5.4.1 *Supongamos que X es una vector que se distribuye según una densidad $f(x)$ con soporte en $\overline{U}$ siendo U un abierto $\mathbb{R}^n$, y que $\varphi : U \to V$ es un difeomorfismo C^1, donde V es otro abierto de $\mathbb{R}^n$ entonces, si consideramos el vector aleatorio $Y = \varphi(X)$, Y se distribuye en V según la densidad*

$$f(\varphi^{-1}(y))|det(D\varphi^{-1})(y)|$$

Prueba: Sea $W \subset V$ un abierto cualquiera, entonces

$$P\{Y \in W\} = P\{X \in \varphi^{-1}(W)\} = \int_{\varphi^{-1}(W)} f(x)dx$$

En esta integral, hagamos el cambio de variable $y = \varphi(x)$, $x = \varphi^{-1}(y)$. Entonces, según el teorema de cambio de variable

$$P\{Y \in W\} = \int_W f(\varphi^{-1}(y))|\text{det}D(\varphi^{-1})(y)|dy$$

Como esto vale para todo $W \subset V$, concluimos que Y se distribuye en V según la densidad $f(\varphi^{-1}(y))|\text{det}(D\varphi^{-1})(y))|$. □

5.5. Independencia

En el capítulo anterior anterior (definicion 4.5.1) introdujimos la noción de variables aleatorias independientes en el caso continuo. Vamos a dar una caracterización de la independencia en términos de la función de densidad de probabilidad conjunta. Para probarlo, necesitaremos un lema de análisis, que generaliza el teorema fundamental del cálculo para integrales bidimensionales:

Lema 5.5.1 (Teorema de diferenciación para integrales) *Supongamos que f es continua en (x_0, y_0) y consideramos la integral*

$$I_{hk} = \frac{1}{hk} \int \int_{R_{hk}} f(x, y)\ dx\ dy$$

siendo R_{hk} el rectángulo

$$R_{hk} = (x_0, x_0 + h] \times (y_0, y0 + k]$$

donde $h, k > 0$ entonces

$$I_{hk} \to f(x_0, y_0) \text{ cuando } (h, k) \to 0$$

Prueba: Como f es **continua** en (x_0, y_0), dado $\varepsilon > 0$ podemos elegir $\delta > 0$ tal que

$$|f(x,y) - f(x_0,y_0)| < \varepsilon$$

si $\|(x - x_0, y - y_0)\|_\infty = \max(x - x_0, y - y_0) < \delta$. Entonces

$$\begin{aligned}
|I_{hk} - f(x_0,y_0)| &= \left| \frac{1}{hk} \int\int_{R_{hk}} f(x,y)\, dx\, dy - f(x_0,y_0) \right| \\
&= \left| \frac{1}{hk} \int\int_{R_{hk}} f(x,y)\, dx\, dy - \frac{1}{hk} \int\int_{R_{hk}} f(x_0,y_0)\, dx\, dy \right| \\
&\leq \frac{1}{hk} \int\int_{R_{hk}} |f(x,y) - f(x_0,y_0)|\, dx\, dy \\
&\leq \frac{1}{hk} \int\int_{R_{hk}} \varepsilon\, dx\, dy = \varepsilon
\end{aligned}$$

si $\|(h,k)\|_\infty = \max(|h|,|k|) < \delta$. □

Teorema 5.5.2 *Supongamos que el vector (X,Y) admite una densidad conjunta continua $f(x,y)$. Entonces las variables X e Y son independientes, si y sólo si f se factoriza en la forma:*

$$f(x,y) = f_X(x) f_Y(y)$$

siendo f_X y f_Y las densidades marginales de probabilidad.

Prueba: Supongamos primero que X e Y son independientes, y que el vector (X,Y) se distribuye según la densidad conjunta $f(x,y)$. Entonces X se distribuye según la densidad marginal f_X dada por (5.3), y similarmente Y se distribuye según la densidad marginal dada por (5.4).

Entonces dado $(x_0, y_0) \in \mathbb{R}^2$ y $h, k > 0$, tenemos que:

$$P\{x_0 < X \leq x_0 + h, y_0 < Y \leq y_0 + k\} = \int_{x_0}^{x_0+h} \int_{y_0}^{y_0+k} f(x,y)\, dx\, dy \tag{5.8}$$

$$P\{x_0 < X \leq x_0 + h\} = \int_{x_0}^{x_0+h} f_X(x)\, dx \tag{5.9}$$

$$P\{y_0 < Y \leq y_0 + k\} = \int_{y_0}^{y_0+k} f_X(y)\, dy \tag{5.10}$$

En virtud de la definición (4.5.1), vemos que:

$$\frac{P\{x_0 < X \leq x_0 + h, y_0 < Y \leq y_0 + k\}}{hk} = \frac{P\{x_0 < X \leq x_0 + h\}}{h} \cdot \frac{P\{x_0 < X \leq x_0 + k\}}{k} \tag{5.11}$$

De la expresión (5.9) cuando $h \to 0$, deducimos que:

$$\frac{P\{x_0 < X \leq x_0 + h\}}{h} = \frac{F_X(x_0 + h) - F_X(x_0)}{h} \to f_X(x_0)$$

por el teorema fundamental del cálculo (siendo f_X continua en x_0).

Similarmente, cuando $k \to 0$, (5.10) y el teorema fundamental del cálculo nos dicen que:

$$\frac{P\{y_0 < Y \leq y_0 + k\}}{k} = \frac{F_Y(y_0 + k) - F_Y(y_0)}{k} \to f_Y(y_0)$$

Finalmente, de la expresión (5.8), por el teorema de diferenciación para integrales (generalización del teorema fundamental del cálculo), deducimos que:

$$\frac{P\{x_0 < X \leq x_0 + h, y_0 < Y \leq y_0 + k\}}{hk} \to f(x_0, y_0)$$

cuando $h, k \to 0$, siempre que f sea continua en el punto (x_0, y_0).

En consecuencia, cuando $h, k \to 0$, a partir de la relación (5.11), obtenemos que:

$$f(x_0, y_0) = f_X(x_0) f_Y(y_0) \tag{5.12}$$

Esto prueba una de las implicaciones del teorema[2]

Para probar la afirmación recíproca, supongamos que la densidad conjunta f puede expresarse en la forma:

$$f(x, y) = f_X(x) f_Y(y)$$

siendo f_X y f_Y dos densidades de probabilidad (Notemos que entonces, f_X y f_Y deben ser entonces necesariamente las densidades marginales dadas por (5.3 - 5.4), como se deduce integrando respecto de x y de y).

Entonces, en virtud del teorema de Fubini,

$$P\{a < X \leq b, c < Y \leq d\} = \int_a^b \int_c^d f(x, y)\, dx\, dy =$$

$$= \left(\int_a^b f_X(x)\, dx \right) \left(\int_c^d f_X(y)\, dy \right) = P\{a < X \leq b\} \cdot P\{c < Y \leq d\}$$

[2]Para evitar complicaciones técnicas, hemos supuesto que la densidad conjunta f es continua. No obstante, si f fuera solamente integrable, repitiendo el mismo argumento y usando el teorema de diferenciación de integrales que se ve en análisis real, obtendríamos que la relación (5.12) se verifica en casi todo punto.

por lo que se deduce que X e Y son variables aleatorias independientes. □

Notemos, que el significado de esta demostración, es que la relación (5.12), es una "expresión infinitesimal" de la definición de independencia.

Ejemplo 5.5.3 *Volvamos a considerar el ejemplo de un vector aleatorio (X, Y) que se distribuía según la densidad conjunta (5.1). Como*

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

donde

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

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

vemos que esta densidad describe dos variables con distribución normal estándar $N(0, 1)$ independientes.

Como corolario obtenemos el análogo de la proposición 3.2.9 para variables continuas[3].

Corolario 5.5.4 *Si X e Y son variables aleatorias independientes con esperanza finita, que se distribuyen según una densidad conjunta continua $f(x, y)$ entonces XY tiene esperanza finita y se tiene que*

$$E[XY] = E[X]E[Y]$$

Prueba: Nuevamente usamos la fórmula (5.7), para obtener que[4]

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (xy)\ f(x, y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (xy)\ f_X(x) f_y(Y)\ dx\ dy \\ &= \left(\int_{-\infty}^{\infty} x f_X(x) \right) \left(\int_{-\infty}^{\infty} y f_X(y) \right) \\ &= E[X]E[Y] \end{aligned}$$

□

[3] La propiedad vale aunque X e Y no admitan una densidad conjunta continua. Una demostración se da en el apéndice E pero utilizando la integral de Lebesgue.

[4] Para justificar rigurosamente este cálculo, hay que hacerlo primero con $|xy|$ en lugar de xy, lo que conduce a $E(|XY|) = E(|X|)E(|Y|)$, con lo que se establece que la integral doble es absolutamente convergente y se justifica la aplicación del teorema de Fubini.

5.6. Suma de variables aleatorias independientes

Como aplicación podemos volver a demostrar la siguiente proposición:

Proposición 5.6.1 *Supongamos que X e Y son variables aleatorias independientes, que se distribuyen en $\mathbb{R}$ según las densidades $f(x)$ y $g(x)$ respectivamente, entonces $X+Y$ se distribuye según la densidad $f * g(x)$.*

Prueba: Como X e Y son independientes,

$$(X, Y) \sim f(x)g(y)$$

Hacemos el cambio de variable lineal $(U, V) = \varphi(X, Y) = (X + Y, Y)$. Entonces $(X, Y) = \varphi^{-1}(U, V) = (U - V, V)$. Como φ es una transformación lineal, su diferencial coincide con ella misma. Para calcular el determinante de φ observamos que su matriz en la base canónica de $\mathbb{R}^2$ es:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

En consecuencia, el determinante de φ es 1. Por el teorema anterior, tenemos que (U, V) que:

$$(U, V) \sim f(u - v)g(v) \text{ (densidad conjunta)}$$

Para recuperar la densidad de U (densidad marginal) debemos integrar en la variable v:

$$U \sim \int_{-\infty}^{\infty} f(u - v)g(v)\, dv$$

□

5.6.1. Vectores aleatorios n-dimensionales

Las ideas anteriores se generalizan sin dificultad a vectores aleatorios multidimensionales, pero la notación resulta más complicada. Así pues si $X : \Omega \to \mathbb{R}^n$ es un vector aleatorio n-dimensional, que se distribuye según una densidad conjunta $f(x) = f(x_1, x_2, \ldots, x_n)$ que supongremos por simplicidad continua, tendremos que:

- La esperanza de una función $\varphi(X)$ del vector X, donde $\varphi : X \to \mathbb{R}$ es una función continua, se puede calcular mediante la fórmula:

$$E[\varphi(X)] = \int_{\mathbb{R}^n} \varphi(x)f(x)\, dx$$

- La k-ésima componente X_k del vector X $(1 \leq k \leq n)$ se distribuye según la densidad marginal:

$$f_{X_k}(x) = \int_{\mathbb{R}^{n-1}} f(x_1, x_2, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)\, dx_1\, dx_2\, \ldots\, dx_{k-1}\, dx_{k+1} \ldots dx_n$$

- Las componentes X_1, X_2, ...X_n del vector X se dirán mutuamente independientes si para cualquier rectángulo n-dimensional (producto de intervalos)

$$I = \prod_{k=1}^{n} (a_k, b_n]$$

se verifica que:

$$P\{X \in I\} = \prod_{k=1}^{n} P\{a_k < X_k \leq b_k\}$$

En términos de la función de distribución conjunta, $X_1, X_2, \ldots, X_n$ son mutuamente independientes si y sólo si $f(x)$ se factoriza en la forma:

$$f(x) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

5.7. Estadísticos de orden

Ejercicio 5.7.1 (práctica 6, item a)) *Dadas $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas con función de distribución acumulada F, se definen sus estadísticos de orden $X^{(1)}, \ldots, X^{(n)}$ como aquellas variables aleatorias que se obtienen ordenando las X_i de manera creciente. En particular, tenemos que*

$$X^{(1)} = \min_{1 \leq i \leq n} X_i$$

$$X^{(n)} = \max_{1 \leq i \leq n} X_i.$$

Hallar para cada $k = 1, \ldots, n$ la función de distribución acumulada de $X^{(k)}$ en términos de F.

En estadística, cuando $X_1, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas con función de distribución acumulada F, decimos que tenemos una **muestra aleatoria** de tamaño n de la distribución F (con reposición).

5.7.1. Distribución del máximo

Empezemos mirando el máximo $X^{(n)}$. Dado $x \in \mathbb{R}$, será $X^{(n)} \leq x$ si $X_i \leq x$ para todo i. De modo que

$$P\{X^{(n)} \leq x\} = \prod_{i=1}^{n} P\{X_i \leq x\} \textbf{ por independencia}$$

o sea

$$F_{X^{(n)}}(x) = \prod_{i=1}^{n} F_{X_i}(x) = F^n(x)$$

al ser las X_i idénticamente distribuidas.

5.7.2. Distribución del mínimo

Similarmente miremos el máximo $X^{(1)}$. Dado $x \in \mathbb{R}$, será $X^{(1)} \leq x$ si y $X_i \leq x$ para algún i.

Queremos hallar $F_{X^{(1)}}(x) = P\{X^{(1)} \leq x\}$. Es más fácil mirar la probabilidad complementaria: Nuevamente como las variables son independientes,

$$\begin{aligned} F_{X^{(1)}}(x) &= 1 - P\{X^{(1)} > x\} \\ &= 1 - P\{X_i > x \text{ para todo } i\} \\ &= 1 - \prod_{i=1}^{n} P\{X_i > x\} \quad \textbf{por independencia} \\ &= 1 - \prod_{i=1}^{n} [1 - F_{X_i}(x)] \\ &= 1 - [1 - F(x)]^n \end{aligned}$$

5.7.3. Distribución de los estadísticos de orden

Consideremos ahora uno cualquiera de los estadísticos de orden $X^{(k)}$ y, dado un $x \in \mathbb{R}$, preguntémosnos cuando $X^{(k)}$ eso significa que tenemos k observaciones que son menores o iguales que x.

Definimos las variables

$$Z_i = \begin{cases} 1 & \text{si} \quad X_i \leq x \\ 0 & \text{si} \quad X_i > x \end{cases}$$

Vemos que son variables de Bernoulli con probabilidad de éxito $p = F(x)$. Son independientes porque las X_i lo eran.

La variable aleatoria

$$N = \sum_{i=1}^{n} Z_i$$

representa el número total de observaciones X_i que son menores o iguales que x. Notamos que

$$N \sim \text{Bi}(n, p)$$

Entonces

$$P\{X^{(k)} \leq x\} = P\{N \geq k\} = \sum_{j=k}^{n} b(j, n, p)$$

donde

$$b(j, n, p) = \binom{n}{j} p^j q^{n-j} \quad q = 1 - p$$

O sea:

$$F_{X^{(k)}}(x) = \sum_{j=k}^{n} \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}$$

5.7.4. Un ejemplo

Ejercicio 5.7.2 (Ejercicio 9, práctica 6) *Sean $X_1, \ldots, X_n$ variables aleatorias independientes con distribución exponencial de parámetros $\alpha_1, \ldots, \alpha_n$ respectivamente. Mostrar que la distribución de $X^{(1)}$ es exponencial. ¿De qué parámetro?*

Solución: Recordamos que para una distribución exponencial $\text{Exp}(\alpha)$

$$F(x) = \int_0^x \alpha e^{-\alpha x}\, dx = 1 - e^{-\alpha x}$$

Entonces

$$\begin{aligned} F_{X^{(1)}} = 1 - P\{X^{(1)} > x\} &= 1 - P\{X_i > x \text{ para todo } i\} \\ &= 1 - \prod_{i=1}^{n} P\{X_i > x\} = 1 - \prod_{i=1}^{n} [1 - F_{X_i}(x)] \\ &= 1 - \prod_{i=1}^{n} e^{-\alpha_i x} = 1 - e^{-sx} \end{aligned}$$

donde $s = \alpha_1 + \alpha_2 + \ldots + \alpha_n$. Luego $X^{(1)} \sim \text{Exp}(s)$.

5.7.5. Densidad de los estadísticos de orden

Teorema 5.7.3 *Si $X_1, X_2, \ldots, X_n$ son variables continuas independientes identicamente distribuidas con densidad f y función de distribución acumulada*

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

entonces los estadísticos de orden $X^{(k)}$ también son variables continuas con la densidad

$$f_{X^{(k)}}(x) = c_k\, [F(k)]^{k-1}\, (1 - F(x))^{n-k}\, f(x)$$

donde

$$c_k = \frac{m!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1} = k\binom{n}{k}$$

Idea de la Demostración: Antes vimos que

$$F_{X^{(k)}}(x) = \sum_{j=k}^{n} \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}$$

Derivando

$$f_{X^{(k)}} = \sum_{j=k}^{n} \binom{n}{j} \left[jF(x)^{j-1}[1 - F(x)]^{n-j} - (n-j)F(x)^j[1 - F(x)]^{n-j-1} \right] f(x)$$

Pero

$$j\binom{n}{j} = j \cdot \frac{n!}{j!(n-j)!} = n \cdot \frac{(n-1)!}{(j-1)!(n-j)!} = n\binom{n-1}{j-1} = (n-j)\binom{n-1}{j}$$

¡Entonces la suma es telescópica y sólo sobrevive un término! (les dejo terminar la cuenta como ejercicio)

5.8. Las densidades beta como estadísticos de orden de la uniforme

Ejercicio 5.8.1 (Ejercicio 8, item d)) *Probar que si las X_i tienen distribución uniforme en el intervalo $[0, 1]$ entonces para cada $k = 1, \ldots, n$ la variable aleatoria $X^{(k)}$ tiene distribución $\beta(k, n - k + 1)$.*

Solución: Antes vimos que

$$f_{X^{(k)}}(x) = c_k\, [F(k)]^{k-1}\, (1 - F(x))^{n-k}\, f(x)$$

Para la distribución uniforme si $x \in (0, 1)$, $f(x) = 1$, $F(x) = x$, entonces

$$f_{X^{(k)}}(x) = c_k\, x^{k-1}\, (1 - x)^{n-k}$$

Por lo que vemos que $X^{(k)} \sim \beta(k, n - k + 1)$.

5.9. Otro ejercicio sobre estadísticos de orden, para comparar

Ejercicio 5.9.1 (Ejercicio 15 de la práctica 6, item a)) *Sean $X_1, \ldots, X_n$ variables aleatorias absolutamente continuas, independientes e idénticamente distribuidas con función de densidad f y consideremos el vector aleatorio $\overline{X} = (X^{(1)}, \ldots, X^{(n)})$ conformado por sus estadísticos de orden. Mostrar que $\overline{X}$ es absolutamente continuo y que su función de densidad viene dada por*

$$f_{\overline{X}}(x) = n! \prod_{i=1}^{n} f(x_i) I_{\{x : x_1 < \ldots < x_n\}}(x).$$

5.10. Un ejercicio de cambio de variable

Ejercicio 5.10.1 *Se tienen dos variables aleatorias independientes $U, V \sim \mathcal{U}(0,1)$. A partir de ellas se definen las variables aleatorias R y W :*

$$R = \sqrt{-2 \log U}, \quad W = 2\pi V$$

y

$$X = R \cdot \cos W, \quad Y = R \cdot \operatorname{sen} W$$

Caracterizar la distribución del vector (X, Y).

Notamos que R toma valores en $(0, +\infty)$ y W en $(0, 2\pi)$
Para la primera parte consideramos el cambio de variable

$$(R, W) = \varphi_1(U, V) \text{ donde } \varphi_1 : \Omega_1 = (0,1) \times (0,1) \to \Omega_2 = (0, +\infty) \times (0, 2\pi)$$

dado por $\varphi_1(u, v) = (\sqrt{-2 \log r}, 2\pi v)$. Este cambio de variable es biyectivo y su inversa $\varphi_1^{-1} : \Omega_2 \to \Omega_1$ es

$$\varphi_1^{-1}(r, w) = \left(e^{-r^2/2}, \frac{w}{2\pi} \right)$$

Para encontrarla, obervé que:

$$r = \sqrt{-2 \log u} \Leftrightarrow r^2 = -2 \ \log r \Leftrightarrow -\frac{r^2}{2} = \log u \Leftrightarrow u = e^{-r^2/2}$$

$$w = 2\pi v \Leftrightarrow v = \frac{w}{2\pi}$$

Además observamos que

$$r \in (0,1) \Leftrightarrow w \in (0, \infty), v \in (0,1) \Leftrightarrow w \in (0, 2\pi)$$

¡Esta cuenta es fácil porque las varialbes no se mezclan!

Entonces según el **teorema de cambio de variable**

$$f_{(R,W)}(r,w) = f_{(U,V)}(\varphi^{-1}(r,w)) \cdot |\det D(\varphi_1^{-1})(r,w)|$$

Pero

$$f_{(U,V)}(\varphi_1^{-1}(r,w)) = I_{\Omega_2}(z,w) = I_{(0,\infty)}(r) \cdot I_{(0,2\pi)}(w)$$

El jacobiano es:

$$|\det D(\varphi^{-1})(z,w)| = \left|\det \begin{pmatrix} re^{-r^2} & 0 \\ 0 & \frac{1}{2\pi} \end{pmatrix}\right| = \frac{1}{2\pi} z e^{-r^2/2}$$

Luego

$$f_{(R,W)}(r,w) = \frac{1}{2\pi} r e^{-r^2/2} I_{(0,\infty)}(r) \cdot I_{(0,2\pi)}(w)$$

Notamos que R y W son independientes. $W \sim \mathcal{U}(0,2\pi)$ mientras que $R \sim \chi_2$ (una de las distribuciones que introdujimos en la clase 11).

Ahora hacemos un nuevo cambio de variable $\varphi_2 : \Omega_2 \to \Omega_3 = \mathbb{R}^2$ dado por

$$(x,y) = \varphi_2(r,w) = (r\cos w, r \operatorname{sen} w)$$

Este cambio de variable lo conocemos bien: es el cambio de variables polares. Sabemos que su jacobiano es r, y que podemos hacerlo biyectivo quitando un conjunto de área cero. Entonces, el teorema de cambio de variable se aplica también. Además

$$r^2 = x^2 + y^2$$

Y como $\det(D\varphi)(r,w) = r \Rightarrow \det(D\varphi^{-1}) = \frac{1}{r}$. Encontramos que:

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi} r e^{-r^2/2} \cdot \frac{1}{r} = e^{-r^2/2} = \frac{1}{2\pi} e^{(x^2+y^2)/2} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

Como ya vimos, esto significa que X e Y con variables con distribución normal estándar independientes.

- Este ejercicio proporciona un método para simular en la computadora la distribución normal, a partir de un generador de números pseudo-aleatorios que simula la distribución uniforme.

- La cuenta del ejercicio es la misma que la que se hace en análisis 2 para calcular el área bajo la curva normal.

5.10.1. Densidad del cociente de dos variables aleatorias independientes

Supongamos que X e Y son variables aleatorias continuas independientes, con densidades f_X y f_Y respectivamente. Supongamos además que Y está concentrada en la semirrecta positiva $(0, +\infty)$. Queremos calcular la densidad del cociente $U = X/Y$.

La densidad conjunta del vector aleatorio (X, Y) será $f_X(x)f_Y(y)$ como consecuencia de independencia de las variables X e Y.

Consideramos ahora el cambio de variable $(U, V) = \varphi(X, Y)$ donde donde

$$(u, v) = \varphi(x, y) = (x/y, y)$$

entonces la función inversa será

$$(x, y) = \varphi^{-1}(u, v) = (uv, v)$$

Y la diferencial de φ^{-1} es

$$D\varphi^{-1}(u, v) = \begin{pmatrix} v & u \\ 0 & 1 \end{pmatrix}$$

de modo que el Jacobiano es v. De acuerdo a la proposición 5.4.1, encontramos que el vector (U, V) se distribuye según la densidad conjunta

$$f_X(tv)f_Y(v)v$$

e integrando respecto la variable v podemos recuperar la densidad (marginal) de U que resulta ser:

$$f_U(t) = \int_0^\infty f_X(tv)f_Y(v)v \, dv \tag{5.13}$$

5.10.2. La densidad t de Student

Sea X una variable aleatoria con distribución χ^2 con n grados de libertad, Y una variable aleatoria con distribución normal estándar y supongamos que X e Y son independientes. Queremos calcular la densidad de la variable aleatoria

$$T = \frac{\sqrt{\frac{X}{n}}}{Y}$$

[El porqué esta variable aleatoria es interesante, lo veremos más adelante al desarrollar conceptos de estadística]

Ya vimos que la densidad de X viene dada por (4.18) Consideramos $\varphi : (0, +\infty) \to (0, +\infty)$ dada por

$$\varphi(x) = \sqrt{\frac{x}{n}}$$

es un difeomorfismo cuya inversa es $\varphi^{-1}(y) = ny^2$.

Aplicando la fórmula de cambio de variables, encontramos que la densidad de $U = \sqrt{\frac{X}{n}}$ es

$$\begin{aligned} f_Y(y) &= \frac{(1/2)^{n/2}}{\Gamma(n/2)}\ (ny^2)^{n/2-1} e^{-ny^2/2} 2ny\ I_{(0,+\infty)}(y) \\ &= \frac{2n^{n/2}}{2^{n/2}\Gamma(n/2)}\ y^{n-1}\ e^{-ny^2/2}\ I_{(0,+\infty)}(y) \end{aligned}$$

Utilizando la fórmula (5.13), vemos que T se distribuye según la densidad

$$\begin{aligned} f_T(t) &= \int_0^\infty f_X(tv) f_Y(v) v\ dv = \frac{2n^{n/2}}{2^{n/2}\Gamma(n/2)\sqrt{2\pi}}\ \int_0^\infty e^{-t^2v^2/2}\ v^{n-1}\ e^{-nv^2/2}\ v\ dv \\ &= \frac{2^{(1-n)/2} n^{n/2}}{\Gamma(n/2)\sqrt{\pi}}\ \int_0^\infty e^{-(t^2+n)v^2/2}\ v^n\ dv \quad (t>0) \end{aligned}$$

Hacemos el cambio de variable $x = \frac{v^2}{2}(t^2+n)$, entonces esta integral se transforma en

$$\begin{aligned} f_T(t) &= \frac{2^{(1-n)/2} n^{n/2}}{\Gamma(n/2)\sqrt{\pi}} \frac{1}{n+t^2}\ \int_0^\infty e^{-x}\ \left(\frac{2x}{n+t^2}\right)^{(n-1)/2}\ dx \\ &= \frac{n^{n/2}}{\Gamma(n/2)\sqrt{\pi}} \frac{1}{(n+t^2)^{(n+1)/2}}\ \int_0^\infty e^{-x}\ x^{(n-1)/2}\ dx \\ &= \frac{n^{n/2}}{\Gamma(n/2)\sqrt{\pi}}\ \Gamma\left(\frac{n+1}{2}\right)\ \frac{1}{(n+t^2)^{(n+1)/2}} \\ &= \frac{1}{\Gamma(n/2)\sqrt{n\pi}}\ \Gamma\left(\frac{n+1}{2}\right)\ \frac{n^{(n+1)/2}}{(n+t^2)^{(n+1)/2}} \end{aligned}$$

Finalmente obtenemos

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)\sqrt{n\pi}}\ \left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad (t>0) \tag{5.14}$$

Esta distribución se conoce como distribución t de Student con n grados de libertad. Surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño y la desviación estándar poblacional es desconocida.

Un dato curioso: La distribución de Student fue descripta en el año 1908 por William Sealy Gosset. Gosset trabajaba en una fábrica de cerveza, Guinness, que prohibía a sus empleados la publicación de artículos científicos debido a una difusión previa de secretos industriales. De ahí que Gosset publicase sus resultados bajo el pseudónimo de "Student".

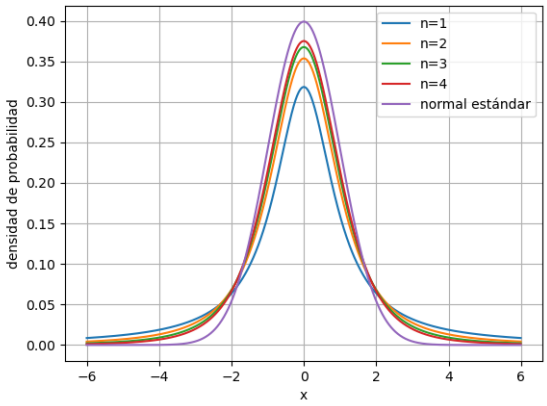


Figura 5.2: Gráfico de la densidad t de Student. Cuando $n \to +\infty$, estas curvas convergen a la densidad normal estándar (¡ejercicio fácil de límites!).

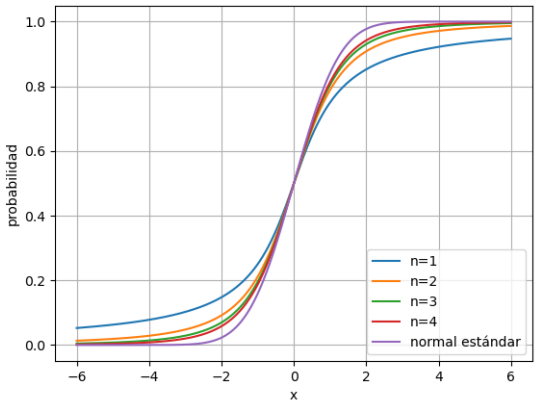


Figura 5.3: Gráfico de la distribución acumulada de una t de Student. Cuando $n \to +\infty$, estas curvas convergen a la distribución acumulada de una normal estándar.

Capítulo 6

Distribución normal multivariada

6.1. Un repaso de algunas nociones de Álgebra Lineal

6.1.1. Transpuesta de una matriz

Dada una matriz $A \in \mathbb{R}^{m \times n}$, su **matriz transpuesta** $A^t \in \mathbb{R}^{n \times m}$ se obtiene intercambiando las filas y las columnas.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \in \mathbb{R}^{2 \times 3} \Rightarrow A^t \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \in \mathbb{R}^{3 \times 2}$$

La operación de transponer tiene algunas propiedades interesantes:

$$(A + B)^t = A^t + B^t, \quad (A \cdot B)^t = B^t \cdot A^t, \quad \det(A^t) = \det(A)$$

Vamos a escribir los vectores como columnas. El producto escalar lo podemos escribir así:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} \in \mathbb{R}^n \Rightarrow \langle x, y \rangle = x^t \cdot y$$

6.1.2. Matrices Simétricas y Ortogonales

- $A \subset \mathbb{R}^{n \times n}$ se dice simétrica si $A^t = A$.
- $P \subset \mathbb{R}^{n \times n}$ se dice ortogonal si $P^t \cdot P = P \cdot P^t = I$.

Teorema 6.1.1 *Si $A \subset \mathbb{R}^{n\times n}$ es simétrica, entonces existe P ortogonal tal que*

$$D = P^t AP = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

es diagonal, siendo los $\lambda_k \in \mathbb{R}$ los ***autovalores*** *de la matriz A.*

6.1.3. Formas Cuadráticas

Una **forma cuadrática** en las variables $x_1, x_2, \dots x_n$ es un polinomio homogéneo de segundo grado en ellas, por ejemplo

$$q_2(x_1, x_2) = x_1^2 + 2x_2^2 - 6x_1x_2$$

$$q_3(x_1, x_2, x_3) = x_1^2 + 4x_2^2 - x_3^2 - 6x_1x_2 - 8x_1x_3$$

son formas cuadráticas en 2 y 3 variables respectivamente.

Dada una matriz simétrica $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, podemos asociarle la forma cuadrática en n variables

$$q_A(x) = \langle Ax, x\rangle = x^t \cdot A \cdot x = \sum_{i,j=1}^{n} a_{ij}x_ix_j$$

Recíprocamente, cada forma cuadrática está asociada a una única matriz simétrica. Veamos cómo:

$$\begin{aligned} q_2(x_1, x_2) &= x_1^2 + 2x_2^2 - 6x_1x_2 \\ &= x_1^2 + 2x_2^2 - 3x_1x_2 - 3x_2x_1 \end{aligned}$$

$$\Rightarrow q_2 = q_A \text{ con } A = \begin{pmatrix} 1 & -3 \\ -3 & 2 \end{pmatrix}$$

Similarmente

$$q_3(x_1, x_2, x_3) = x_1^2 + 4x_2^2 - x_3^2 - 6x_1x_2 - 8x_1x_3 = q_B(x)$$

con

$$B = \begin{pmatrix} 1 & -3 & -4 \\ -3 & 4 & 0 \\ -4 & 0 & -1 \end{pmatrix}$$

- Una **forma cuadrática** $q_A(x)$ y la correspondiente matriz simétrica A se dicen semi-definidas positivas si
$$q_A(x) \geq 0 \text{ para todo } x \in \mathbb{R}^n$$
Ejemplo:
$$q_A(x) = x_1^2 - 2x_1x_2 + x_2^2 = (x_1 - x_2)^2, \quad A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$
es semifinida positiva.

- Una **forma cuadrática** $q_A(x)$ y la correspondiente matriz simétrica A se dicen definidas positivas si
$$q_A(x) > 0 \text{ para todo } x \neq \vec{0} \in \mathbb{R}^n$$
Ejemplo
$$q_A(x) = x_1^2 - x_1x_2 + x_2^2 = \left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{4}x_2^2 \quad A = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}$$

Teorema 6.1.2 *Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica, (λ_k) sus autovalores, y q_A su forma cuadrática asociada.*

- *A es semi-definida positiva si y sólo si $\lambda_k \geq 0$ para todo k.*
- *A es definida positiva si y sólo si $\lambda_k > 0$ para todo k.*

Corolario 6.1.3

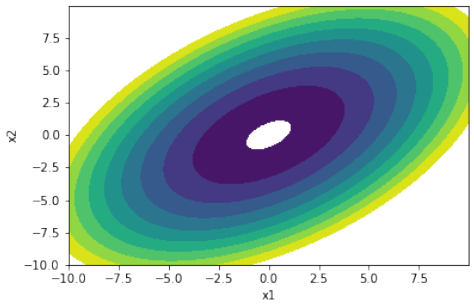
- *Si A es semi-definida positiva,* $\det(A) \geq 0$.
- *Si A es definida positiva,* $\det(A) > 0$.

Esto es inmediato, pues $\det(A) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$.

Curvas de nivel

En $n = 2$, las curvas de nivel de una forma cuadrática definida positiva son elipses. Veámoslo en el ejemplo

$$q_A(x) = x_1^2 - x_1x_2 + x_2^2 = \left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{4}x_2^2$$



En $n = 3$ las superficies de nivel de una forma cuadrática definida positiva serán elipsoides.

6.2. Espeanza de un vector aleatorio y Matriz de covariancias

Consideramos un vector aleatorio X. Su **esperanza** se define componente a componente, y es un nuevo vector (no aleatorio)

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_n \end{pmatrix} \in \mathbb{R}^n \Rightarrow \mu_X = E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \dots \\ E[X_n] \end{pmatrix} \in \mathbb{R}^n$$

Definimos su **matriz de covariancias** $\Sigma = \Sigma_X = \text{Cov}$ por $\Sigma_{i,j} = \text{Cov}(X_i, X_j)$.

$$\Sigma_X = \text{Cov}(X) = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(2_1, X_2) & \dots & \text{Cov}(X_2, X_n) \\ \dots & & & \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Cov}(X_n, X_n) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Notamos que es una **matriz simétrica**. También podemos escribir:

$$\text{Cov}(X) = E[(X - \mu_X) \cdot (X - \mu_X)^t]$$

Notamos que en la diagonal de la matriz de covariancias $\text{Cov}(X)$ aparecen las variancias

$$\sigma^2_{X_i} = \text{Cov}(X_i, X_i) = \text{Var}(X_i)$$

Otra observación interesante es que si las componentes del vector X son independientes, entonces serán no correlacionadas

$$\text{Cov}(X_i, X_j) = 0 \text{ si } i \neq j$$

por lo que la matriz $\text{Cov}(X)$ será diagonal.

Un ejemplo que ya vivmos: Distribución normal multivariada estándar

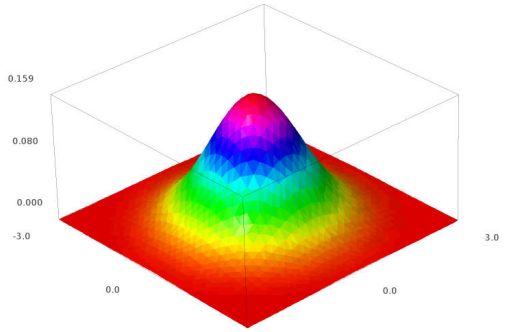
Si X es un vector con componentes $X_i \sim N(0,1)$ independientes, su densidad conjunta vendrá dada por

$$f(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|^2/2}$$

Tenemos

$$E[X] = \vec{0}, \quad \text{Cov}(X) = I \textbf{ (matriz identidad)}$$

Por ejemplo si $n = 2$ tenemos la distribución normal bivariada estándar



Efecto de un cambio lineal sobre la esperanza y la matriz de covariancias

Si hacemos un cambio lineal $Y = A \cdot X + b$ donde ahora $b \in \mathbb{R}^n$ es un vector no aleatorio, y $A \in \mathbb{R}^{n \times n}$ es una matriz no aleatoria, encontramos que:

$$E[Y] = AE[X] + E[b] = AE[X] + b = A \cdot \mu_X + b$$

mientras que:

$$\begin{aligned}
\text{Cov}[Y] &= E[(X - \mu_Y) \cdot (Y - \mu_Y)^t] \\
&= E[((A \cdot \mu_X + b)) - (A \cdot \mu_X + b)) \cdot (A \cdot X + b - (A \cdot \mu_X + b))^t] \\
&= E[(A \cdot (X - \mu)) \cdot (A \cdot (X - \mu_X))^t] \\
&= E[A \cdot (X - \mu)) \cdot (X - \mu_X)^t \cdot A^t] \\
&= A \cdot E[(X - \mu)) \cdot (X - \mu_X)^t] \cdot A^t \\
&= A \cdot \text{Cov}(X) \cdot A^t
\end{aligned}$$

o sea:

$$\Sigma_Y = A \cdot \Sigma_X \cdot A^t$$

La matriz de covariancias es siempre definida positiva

Teorema 6.2.1
- *Si X es un vector aleatorio n-dimensional, su matriz se covarinncias* $\mathrm{Cov}(X)$ *es una matriz simétrica* ***semi-definida positiva****.*
- *Además, es definida positiva, salvo en el caso en que la distribución del vector X está concentrada en un hiperplano afin H, es decir cuando existe un hiperplano afín*

$$H = \{x \in \mathbb{R}^n : \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \ldots + \alpha_n \cdot x_n = b\}$$

tal que

$$P\{X \in H\} = 1$$

Prueba: Sea $\mu_X = E[X]$. Entonces ya observamos que $\mathrm{Cov}(X - \mu_X) = \mathrm{Cov}(X)$. Por lo que podemos suponer sin pérdida de generalidad que $\mu_X = \vec{0}$.

Entonces $\mathrm{Cov}(X) = E[X \cdot X^t]$. Consideremos la expresión

$$q(\alpha) = E[(\alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n \cdot X_n)^2] \quad \alpha \in \mathbb{R}^n$$

Notamos que $q(\alpha) \geq 0$ y que

$$q(\alpha) = E\left[\sum_{i,j=1}^{n} X_i X_j \alpha_i \alpha_j\right] = \sum_{i,j=1}^{n} E[X_i X_j]\alpha_i \alpha_j = \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j)\alpha_i \alpha_j$$

Entonces $q(\alpha)$ es la **forma cuadrática** asociada a la matriz $\mathrm{Cov}(X)$. Deducimos que $\mathrm{Cov}(X)$ es semidefinida positiva. Finalmente si para algún $\alpha \in \mathbb{R}^n$,

$$q(\alpha) = 0 \Rightarrow \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n \cdot X_n = 0 \text{ con probabilidad 1}$$

y esto dice que la distribución del vector X está concentrada en un hiperplano.

□

6.3. Distribución normal multivariada en general

Planteo del problema

En el ejercicio 25 de la práctica 7 se plantea el siguiente problema, supongamos que X es un vector con distribución normal multivariada estándar como vimos antes, hacemos un cambio lineal de variable

$$Y = A \cdot X + b$$

donde A es una matriz no singular. ¿cuál es la distribución de Y?

Esta distribución se llamará distribución **normal multivariada**, y es sumamente útil en las aplicaciones a la estadística. Generalizará a n dimensiones la distribución normal.

En las guías prácticas se considera el caso especial en que $n = 2$ (distribución normal bivariada), pero las cuentas son igualmente fáciles en general (con la notación adecuada).

Algunas observaciones

- Para simplificar, vamos a considerar primero el caso especial donde $b = \vec{0}$. (distribución normal multivariada centrada en el origen)
- Ya vimos que entonces la esperanza y varianza de Y serán

$$\mu_Y = A \cdot \mu_X = \vec{0}$$

$$\Sigma_Y = A \cdot \Sigma_X \cdot A^t = A \cdot A^t$$

[Ojo: ¡en esta expresión el orden importa, no siempre una matriz A conmuta con su transpuesta A^t !]

Esta es una matriz **matriz simétrica definida positiva** asociada a la forma cuadrática $q(x) = \|A^t \cdot x\|^2$ pues

$$q(x) = (A^t \cdot x)^t \cdot (A^t \cdot x) = x^t \cdot (A \cdot A^t) \cdot x$$

y como A es no singular, A^t también con lo que $q(x) = 0$ si y sólo si $x = 0$.

Fórmula de la densidad conjunta en la normal multivariada

Usando el *teorema de cambio de variable* que vimos en la clase 11 con $y = \varphi(x) = A \cdot x$, $\varphi^{-1}(y) = A^{-1} \cdot x$, tenemos que la densidad conjunta de Y se relaciona con la de X por

$$\begin{aligned} f_Y(y) &= f_X(A^{-1}y) \cdot |det(A^{-1})| \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\|A^{-1}y\|^2/2} \cdot |det(A^{-1})| \end{aligned}$$

Vamos a reescribir esta fórmula en términos de la matriz de covariancias

$$\Sigma = \Sigma_Y = A \cdot A^t$$

Notamos que

$$\|A^{-1}y\|^2 = (A^{-1}y)^t \cdot (A^{-1}y) = y^t \cdot (A^{-1})^t \cdot A^{-1} \cdot y$$

Como por otra parte:

$$\Sigma^{-1} = (A \cdot A^t)^{-1} = (A^t)^{-1} \cdot A^{-1} = (A^{-1})^t \cdot A^{-1}$$

vemos que esta expresión es la **forma cuadrática** $q_{\Sigma^{-1}}$ asociada a Σ^{-1}

Fórmula de la densidad conjunta en la normal multivariada

Hasta ahora vimos que

$$f_Y(y) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}q(y)} \cdot |det(A^{-1})|$$

donde

$$q(y) = q_{\Sigma^{-1}}(y) = y^t \Sigma^{-1} y$$

Finalmente, veamos que relación tiene del determinante de A^{-1} con el de Σ. Como $\Sigma = A \cdot A^t$, entonces

$$\det(\Sigma) = \det(A) \cdot \det(A^t) = \det(A)^2 \Rightarrow |det(A^{-1})| = \det(\Sigma)^{-1/2}$$

y obtenemos la fórmula de la *Densidad normal multivariada centrada en el origen*:

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}y^t \Sigma^{-1} y}$$

Distribuciones marginales de la normal multivariada

Como

$$Y_j = \sum_{i=1}^{n} A_{i,j} \cdot X_j$$

y las $X_j \sim N(0,1)$ independientes,

$$A_{i,j} \cdot X_j \sim N(0, A_{i,j}^2)$$

Usando el teorema que vimos en la clase 11 sobre la suma de variables normales independientes, obtenemos que:

$$Y_j \sim N(0, \sigma_j^2) \text{ donde } \sigma_j^2 = \sum_{i=1}^{n} A_{i,j}^2$$

Notamos que esto es consistente con la fórmula

$$\Sigma_Y = A^t \cdot A$$

que obtuvimos antes (Las σ_j^2 aparecen en la diagonal de la matriz Σ_Y).

Un caso especial

Un caso de especial interés es cuando la matriz A con la que hacemos el cambio de variable es ortogonal $Y = A \cdot X$, lo que siginifca que $\Sigma = A \cdot A^t = I$. También $\det(\Sigma) = 1$. Por lo que obtenemos que $f_Y = f_X$, o sea:

Proposición 6.3.1 *Si X tiene distribución normal multivariada estándar, y hacemos un cambio de variable $Y = A \cdot X$ con A una **matriz ortogonal**, Y también tiene distribución normal multivariada estándar.*

Caso general

Si $b \in \mathbb{R}^n$ es cualquiera y $A \in \mathbb{R}^{n\times n}$ es una matriz no singular, a partir de $Y = A\cdot X + b$ obtendríamos que

$$\mu_Y = E[Y] = b$$

y que

$$\Sigma = \mathrm{Cov}(Y) = A \cdot A^t$$

mientras que la densidad de Y será

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(y-\mu)^t \Sigma^{-1}(y-\mu)}$$

Algunas observaciones

Proposición 6.3.2 *Si X tiene distribución normal multivariada, son equivalentes:*

- *Las componentes X_j de X son independientes.*
- *Las X_j no están correlacionadas, o sea*

  $$\mathrm{Cov}(X_i, X_j) = 0 \text{ si } i \neq j$$

  o sea la matriz $\mathrm{Cov}(X)$ *es diagonal.*

Recordamos que esta propiedad NO es cierta para vectores aleatorios en general.

El caso especial $n = 2$, distribución normal bivariada

En el ejercicio 25 de la práctica 7 se considera el caso especial $n = 2$, con un ligero cambio de notación: ahora el vector aleatorio se denota (X, Y), no (Y_1, Y_2). La densidad conjunta es:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2-2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]\right\}}$$

donde

$$\mu = \begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix} \qquad \text{y} \qquad \Sigma = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}.$$

y ρ es el **coeficiente de correlación** entre X e Y.

Esta fórmula sale de que en este caso

$$\det(\Sigma) = \sigma_X^2\sigma_Y^2(1-\rho^2), \quad \mathrm{Adj}(\Sigma) = \begin{pmatrix}\sigma_Y^2 & -\rho\sigma_X\sigma_Y\\ -\rho\sigma_X\sigma_X & \sigma_Y^2\end{pmatrix}$$

$$\Sigma^{-1} = \frac{\mathrm{Adj}(A)^t}{\det(A)} = \frac{1}{1-\rho^2}\begin{pmatrix}\frac{1}{\sigma_X} & -\frac{\rho}{\sigma_X\sigma_Y}\\ -\frac{\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y}\end{pmatrix}$$

Capítulo 7

Teoría de la predicción

7.1. El contexto abstracto en el que vamos a trabajar

Consideramos un espacio de probabilidad $(\Omega, \mathcal{E}, P)$. Consideramos el **espacio vectorial** de las variables aleatorias con **segundo momento finito**

$$L^2(\Omega) = \{\text{variables aleatorias } X : \Omega \to \overline{\mathbb{R}} : E(X^2) < \infty\}$$

Recordamos que si $X \in L^2(\Omega)$,

$$E(|X|) \leq E(X^2)^{1/2} \text{ por la desigualdad de Jensen}$$

y

$$\text{Var}(X) = E(X^2) - E(X)^2$$

Por lo que las variables aleatorias en L^2 tienen esperanza y varianza finitas.

$L^2(\Omega)$ es un espacio normado con la norma

$$\|X\| = E(X^2)^{1/2}$$

que proviene del **producto interno**

$$\langle X, Y \rangle = E(X \cdot Y)$$

Es un **espacio con producto interno** o **espacio pre-Hilbert**.

Para que $L^2(\Omega)$ sea realmente un espacio vectorial normado, hay considerar iguales a las variables aleatorias X e Y tales que

$$P\{X = Y\} = 1$$

Con esta convención,

$$\|X\| = 0 \Rightarrow X = 0$$

7.2. Planteo del problema

Consideramos dentro de L^2 un subespacio S. Queremos aproximar una variable aleatoria Y por un elemento del subespacio $\widehat{Y} \in S$.

Particularmente, vamos a usar dos subespacios:

$$S_1 = \text{variables aleatorias constantes} = \langle 1 \rangle$$

y dada una variable aleatoria X vamos a considerar

$$S_2 = \{\alpha X + \beta : \alpha, \beta \in \mathbb{R}\} = \langle 1, X \rangle$$

La idea es que queremos usar $\widehat{Y}$ para predecir el valor de Y, por eso en la teoría de probabilidades se lo llama un **predictor** de Y.

¿Cuál es la mejor manera de elegir $\widehat{Y}$? Eso depende de cómo midamos el error en la aproximación. Vamos a usar el criterio del **error cuadrático medio**. Queremos minimizar

$$\text{ECM}(Y, \widehat{Y}) = E(|Y - \widehat{Y}|^2) = \|Y - \widehat{Y}\|^2$$

7.3. Un lema de álgebra lineal

Lema 7.3.1 *Sea V un espacio con producto interno y $S \subset V$ un subespacio. Consideramos $x_0 \in V$. Entonces $s_0 \in S$ es el elemento de S que minimiza la distancia a x_0*

$$d(x, s) = \|x - s\| \quad x \in S$$

*si y sólo si s_0 es la **proyección ortogonal** de x_0 sobre s es decir:*

$$x_0 - s_0 \in S^\perp$$

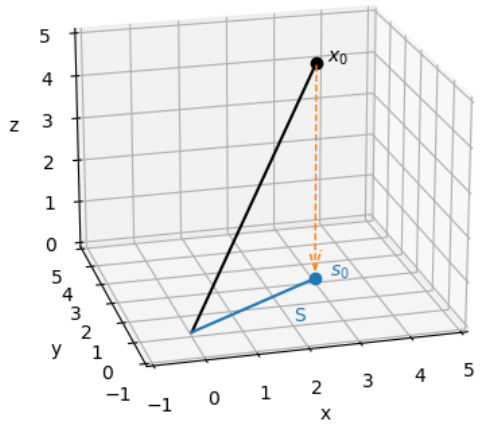
o sea

$$\langle x_0 - s_0, s \rangle = 0 \text{ para todo } s \in S \tag{7.1}$$

Nuestro $V = L^2(\Omega)$ es un espacio de dimensión infinita, pero este lema funciona exactamente igual que en dimensión finita (y con la misma prueba).

Si S fuera de diemnsión finita, es suficiente verificar la condición de ortogonalidad (7.1) para s en una base de S.

Figura 7.1: Ilustración gráfica de la proyección ortogonal



7.4. Predicción por variables aleatorias constantes

Apliquémoslo al primero de nuestros ejemplos

$$S_1 = \text{variables aleatorias constantes} = \langle 1 \rangle$$

(subespacio de dimensión 1).

Dada $Y \in L^2$, la condición para que $\widehat{Y} \in S_0$ sea el predictor constante que minimiza el error medio cuadrático es según el lema (con $S = S_0$, $x_0 = Y$,$s_0 = \widehat{Y}$):

$$\langle Y - \widehat{Y}, 1 \rangle = 0$$

o sea:

$$E[(Y - \widehat{Y}) \cdot 1] = 0$$

Como $\widehat{Y}$ es constante, esto nos dice que el mejor predictor de Y es:

$$\widehat{Y_0} = E[Y]$$

y entonces el error medio cuarático en esta aproximación será

$$\text{EMC}_1 = \min_{\widehat{Y} \in S_1} \|Y - \widehat{Y}\|^2 = \|Y - \widehat{Y}_0\|^2 = E[(Y - \widehat{Y_0})^2] = \text{Var}(Y)$$

7.5. Predicción por funciones lineales de X

Ahora dada otra variable aleatoria X, consideramos

$$S_2 = \{\alpha X + \beta : \alpha, \beta \in \mathbb{R}\} = \langle 1, X \rangle$$

(subespacio de dimensión 2). Según el lema, las condiciones de ortogonalidad que debe verificar el predictor óptimo son:

$$\langle Y - \widehat{Y}, 1 \rangle = 0$$

$$\langle Y - \widehat{Y}, X \rangle = 0$$

o sea:

$$E[(Y - \widehat{Y}) \cdot 1] = 0$$

$$E[(Y - \widehat{Y}) \cdot X] = 0$$

Entonces los coeficientes α, β para el predictor óptimo deben satisfacer que

$$E[(Y - \alpha X - \beta) \cdot 1] = 0$$

$$E[(Y - \alpha X - \beta) \cdot X] = 0$$

La primera condición dice que

$$E[Y] - \alpha E[X] - \beta = 0 \tag{7.2}$$

También multiplicándola por $E[X]$ obteneos que

$$E[(Y - \alpha X - \beta) \cdot E[X]] = 0$$

y entonces restándola de la segunda condición

$$E[(Y - \alpha X - \beta) \cdot (X - E(X))] = 0$$

Reemplazando el valor de β dado por (7.2),

$$E[(Y - \alpha X - E(Y) - \alpha E(X)) \cdot (X - E(X))] = 0$$

por lo tanto

$$E[(Y - E(Y)) - \alpha(X - E(X))] \cdot (X - E(X))] = 0$$

Entonces distribuyendo la esperanza, obtenemos

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E[X]) \cdot (Y - E(Y))] \\ &= \alpha \; E[(X - E(X))^2] \\ &= \alpha \text{Var}(X) \end{aligned}$$

En resumen, hemos demostrado

Teorema 7.5.1 *Sea $\widehat{Y}_0$ el predictor de menor error cuadrático medio en S_2. Vine dado por $\widehat{Y}_0 = \alpha X + \beta$ donde α y β se determinan por las ecuaciones:*

$$\alpha = \frac{Cov(X, Y)}{Var(X)}$$

$$\beta = E(Y) - \alpha E[X]$$

7.6. Cálculo del error cuadrático medio

Calculemos el error cuadrático medio óptimo al aproximar Y por una función lineal de X.

$$\text{EMC}_2 = \min_{\widehat{Y} \in S_2} \|Y - \widehat{Y}\|$$

Primero usamos que en el predictor óptimo $\beta = E(Y) - \alpha E[X]$

$$\begin{aligned} ECM_2 &= \|Y - \widehat{Y}_0\|^2 = E[(Y - \widehat{Y}_0)^2] = E[(Y - \alpha X - \beta)^2] \\ &= E[(Y - \alpha X - (E[Y] - \alpha E[X]))^2] \\ &= E[((Y - E(Y)) - \alpha(X - E(X)))^2] \\ &= E[(Y - E(Y))^2] + \alpha^2 E[(X - E(X))^2] - 2\alpha E[(Y - E(Y)) \cdot (X - E(X))] \\ &= \mathrm{Var}(Y) + \alpha^2 \mathrm{Var}(X) - 2\alpha \mathrm{Cov}(X, Y) \end{aligned}$$

Y usando que $\alpha = \frac{Cov(X,Y)}{\mathrm{Var}(X)}$ en el predictor óptimo,

$$\begin{aligned} ECM_2 &= \mathrm{Var}(Y) + \alpha^2 \mathrm{Var}(X) - 2\alpha \mathrm{Cov}(X, Y) \\ &= \mathrm{Var}(Y) + \frac{\mathrm{Cov}^2(X, Y)}{\mathrm{Var}(X)} - 2\frac{\mathrm{Cov}^2(X, Y)}{\mathrm{Var}(X)} \\ &= \mathrm{Var}(Y) - \frac{\mathrm{Cov}^2(X, Y)}{\mathrm{Var}(X)} \end{aligned}$$

7.7. Mejora en el error medio cuadático

Queremos comparar cuánto mejoró el error medio cuadrático al usar como predictor de Y una función lineal de X comparado con usar una variable aleatoria constante. Para ello consideramos el cociente

$$\begin{aligned} \frac{EMC_1 - EMC_2}{ECM_1} &= \frac{\mathrm{Var}(Y) + \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(X)} - \mathrm{Var}(Y)}{\mathrm{Var}(Y)} \\ &= \frac{\mathrm{Cov}^2(X, Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)} = \rho^2(X, Y) \end{aligned}$$

Esto permite interpretar el coeficiente $\rho^2(X, Y)$ como el decrecimiento relativo del error cuadrático medio cuando se usa un predictor lineal basado en X en vez de un predictor constante. Por lo tanto $\rho^2(X, Y)$ mide la utilidad de la variable X para predecir a Y por una función lineal.

7.8. Algunas observaciones

Notamos que como $S_1 \subset S_2$, $\mathrm{ECM}_2 \leq \mathrm{ECM}_1$. Esto nos dice nuevamente que $|\rho(X, Y)| \leq 1$, o sea nos proporciona otra prueba de la desigualdad de Cauchy-Schwarz.

¿Qué significaría $|\rho| = 1$?. Según la fórmula anterior, esto implica que

$$\text{ECM}_1 - \text{ECM}_2 = \text{ECM}_1 \Rightarrow \text{ECM}_2 = 0$$

o sea que Y es una función lineal de X.

Notamos también que como para el predictor óptimo

$$\alpha = \frac{Cov(X,Y)}{\text{Var}(X)}$$

el signo de $\rho(X,Y)$ coincide con el signo de α. [Si $\rho(X,Y) > 0$ el predictor óptimo será una función lineal creciente de X, mientras que si $\rho < 0$ será una función lineal decreciente]

7.9. Regresión lineal la computadora

Veamos un rogramita en Python, usando SciPy:

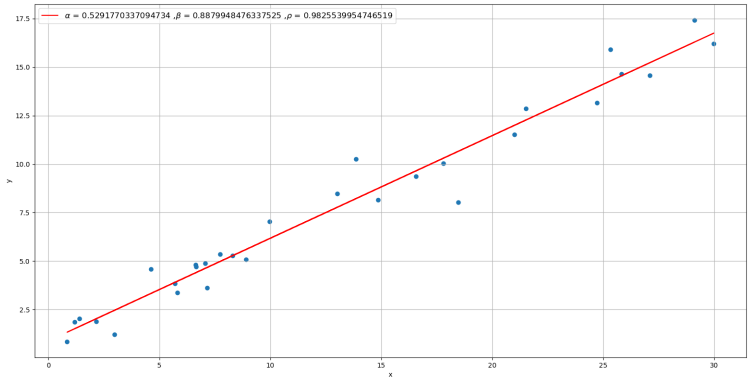
```
# x uniforme en (0,30)

x = np.random.uniform(size=30, low=0, high=30)

# y = a*x + b con ruido
y = 0.5 * x + 1.0 + np.random.normal(scale=1, size=x.shape)

regresion = scipy.stats.linregress(x, y)
alpha = regresion.slope
beta = regresion.intercept
rho = regresion.rvalue
y_predicho = alpha * x + beta
```

El gráfico de la recta resultante se muestra en el siguiente gráfico:



Nota: Este capítulo está basado en las notas de Victor Yohai [Yoh], aunque preferimos utilizar el lenguaje de álgebra lineal abstracta (análisis funcional) para hacer explícito el uso de la proyección ortogonal. Y hemos agregado este último ejemplo para ilustrar cómo realizar una regresión lineal en la computadora.

Capítulo 8

Convergencia de Variables Aleatorias, y Ley Fuerte de los Grandes Números

8.1. Convergencia en probabilidad

En la teoría de probabilidades se utilizan frecuentemente diferentes nociones de convergencia de una sucesión $(X_n)_{n\in\mathbb{N}}$ de variables aleatorias.

La primera noción que vamos a estudiar es la de convergencia en probabilidad, que aparece en el teorema de Bernoulli (ley débil de los grandes números).

Definición 8.1.1 *Sea $(X_n)_{n\in N}$ una sucesión de variables aleatorias, definidas sobre un mismo espacio de probabilidad $(\Omega, \mathcal{E}, P)$. Se dice que (X_n)* **converge en probabilidad** *a la variable X si para todo $\varepsilon > 0$, tenemos que*

$$P\{|X - X_n| > \varepsilon\} \to 0 \text{ cuando } n \to +\infty$$

Notación:

$$X_n \xrightarrow{P} X$$

Observación: Si (X_n) converge en probabilidad a X, cualquier subsucesión de (X_n) también converge en probabilidad a X.

Ejemplo 8.1.2 (Variables con distribución uniforme que se concentran) *Si $X_n \sim \mathcal{U}(-1/n, 1/n)$ y $X = 0$ con probabilidad 1. Entonces $X_n \xrightarrow{P} X$.*

Prueba:

$$\begin{aligned} P\{|X_n - X| > \delta\} &= P\{|X_n - X| > \delta / X = 0\} \cdot P\{X = 0\} \\ &+ P\{|X_n - X| > \delta / X \neq 0\} \cdot P\{X \neq 0\} \\ &= P\{|X_n| > \delta\} = 0 \end{aligned}$$

si $\frac{1}{n} < \delta$ o sea $n > \frac{1}{\delta}$, ya que $|X_n| \leq \frac{1}{n}$ con probabilidad 1. □

Ejemplo 8.1.3 (Variables con distribución normal que se concentran) *Si* $X_n \sim N(0, \sigma_n^2)$ *donde* $\sigma_n \to 0$ *y* $X = 0$ *con probabilidad* 1. *Entonces* $X_n \xrightarrow{P} X$.

Prueba:

$$\begin{aligned} P\{|X_n - X| > \delta\} &= P\{|X_n - X| > \delta / X = 0\} \cdot P\{X = 0\} \\ &+ P\{|X_n - X| > \delta / X \neq 0\} \cdot P\{X \neq 0\} \\ &= P\{|X_n| > \delta\} \leq \frac{1}{\delta^2} \ \mathrm{Var}(X_n) \\ &= \frac{\sigma_n^2}{\delta^2} \to 0 \end{aligned}$$

cuando $n \to +\infty$, por la desigualdad de de Chebyshev. □

Veamos algunas propiedades de la convergencia en probabilidad:

Proposición 8.1.4 (Unicidad del límite) *Si* $X_n \xrightarrow{P} X$ *y* $X_n \xrightarrow{P} Y$, *entonces* $X = Y$ *con probabilidad* 1.

Prueba: Por la desigualdad triangular,

$$|X - Y| \leq |X - X_n| + |X_n - Y|$$

Entonces

$$P\{|X - Y| > \varepsilon\} \leq P\{|X - X_n| > \varepsilon/2\} + P\{|X_n - Y| > \varepsilon/2\}$$

Deducimos que para todo $\varepsilon > 0$,

$$P\{|X - Y| > \varepsilon\} = 0$$

Como

$$\{X \neq Y\} = \bigcup_{n \in N} \left\{ |X - Y| > \frac{1}{n} \right\}$$

Por la σ-subaditividad de P, deducimos que:

$$P\{X \neq Y\} \leq \sum_{n=1}^{\infty} P\left\{ |X - Y| > \frac{1}{n} \right\} = 0$$

□

Proposición 8.1.5 *Si $X_n \xrightarrow{P} X$ y $c \in \mathbb{R}$, entonces $cX_n \xrightarrow{P} cX$.*

Prueba: Si $c \neq 0$, tenemos que

$$P\{|cX_n - cX| > \varepsilon\} = P\left\{|X_n - X| > \frac{\varepsilon}{|c|}\right\} \to 0 \text{ cuando } n \to +\infty$$

Si $c = 0$ es trivial. □

Proposición 8.1.6 *Si $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$, entonces $X_n + Y_n \xrightarrow{P} X + Y$.*

Prueba:

$$P\{|(X+Y) - (X_n + Y_n)| > \varepsilon\} \leq P\{|X - X_n| > \varepsilon/2\} + P\{|Y - Y_n| > \varepsilon/2\}$$

□

Observación 8.1.7 *Sea $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria finita en casi todo punto. Entonces X está acotada en probabilidad en el siguiente sentido, dado $\varepsilon > 0$ existe $k_0 > 0$ tal que*

$$P\{|X| \geq k_0\} < \frac{\varepsilon}{2}$$

Prueba: Notamos que

$$\sum_{k=1}^{\infty} P\{k - 1 \leq |X| < k\} = 1$$

es una serie convergente, por consiguiente dado $\varepsilon > 0$, existirá un k_0 tal que:

$$\sum_{k=k_0+1}^{\infty} P\{k - 1 \leq |X| \leq k\} < \frac{\varepsilon}{2}$$

Es decir que:

$$P\{|X| \geq k_0\} < \frac{\varepsilon}{2}$$

□

Lema 8.1.8 *Si $X_n \xrightarrow{P} X$, entonces (X_n) está acotada en probabilidad, en el siguiente sentido, dado $\varepsilon > 0$ existn $M = M_\varepsilon$*

$$\forall\, n \geq n_0(\varepsilon) :\ P\{|X_n| > M\} < \varepsilon$$

para todo n.

Prueba: Elegimos k_0 como en la observación anterior. De la desigualdad triangular,

$$|X_n| \leq |X_n - X| + |X|$$

Deducimos que:

$$P\{|X_n| > k_0 + \delta\} \leq P\{|X_n - X| > \delta\} + P\{|X| > k_0\}$$

y en consecuencia que

$$P\{|X_n| > k_0 + \delta\} \leq \varepsilon$$

si $n \geq n_0(\varepsilon)$. Como hay una cantidad finita de valores de $n < n_0$ combinando esto con la observación anterior, se obtiene el resultado. □

Lema 8.1.9 *Si $X_n \xrightarrow{P} 0$ e Y_n está acotada en probabilidad, entonces $X_n Y_n \xrightarrow{P} 0$.*

Prueba:

$$P\left\{|X_n Y_n| > \varepsilon\right\} = P\left\{|X_n| > \frac{\varepsilon}{|Y_n|}\right\}$$

$$\leq P\left\{|X_n| > \frac{\varepsilon}{|Y_n|} \ \wedge \ |Y_n| \leq M\right\} + P\left\{|X_n| > \frac{\varepsilon}{|Y_n|} \ \wedge \ |Y_n| > M\right\}$$

$$\leq P\left\{|X_n| > \frac{\varepsilon}{M}\right\} + P\left\{|Y_n| > M\right\} < \varepsilon$$

si $n \geq n_0(\varepsilon)$. □

Corolario 8.1.10 *Si $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$, entonces $X_n Y_n \xrightarrow{P} XY$.*

Prueba: Utilizamos el truco habitual de "sumar y restar":

$$XY - X_n Y_n = XY - X_n Y + X_n Y - X_n Y_n = (X - X_n)Y + X_n(Y_n - Y)$$

Entonces como $X - X_n \xrightarrow{P} 0$ e Y está acotada en probabilidad, deducimos que $(X - X_n)Y \xrightarrow{P} 0$. Similarmente, como $Y_n - Y \xrightarrow{P} 0$ y X_n está acotada en probabilidad (por la proposición 8.1.8, deducimos que $(X - X_n)Y \xrightarrow{P} 0$. Tenemos entonces que $X_n Y_n - XY \xrightarrow{P} 0$, y en consecuencia $X_n Y_n \xrightarrow{P} XY$ (por la proposición 8.1.6) □

8.2. Convergencia casi-segura

Definición 8.2.1 *Se dice que la sucesión (X_n) de variables aleatorias* **converge casi seguramente** *a la variable X si*

$$P\left\{\lim_{n\to+\infty} X_n = X\right\} = 1$$

Notación:

$$X_n \xrightarrow{c.s.} X$$

Proposición 8.2.2 *Si $X_n \xrightarrow{c.s.} X$, entonces $X_n \xrightarrow{P} X$.*

Prueba: Notamos que por la definición de límite,

$$X_n(\omega) \longrightarrow X(\omega) \Leftrightarrow \forall k \geq 1\ \exists n_0 \forall n \geq n_0 : |X_n(\omega) - X(\omega)| \leq \frac{1}{k}$$

Negándola, tenemos que:

$$X_n(\omega) \not\longrightarrow X(\omega) \Leftrightarrow \exists k \geq 1\ \forall n_0 \exists n \geq n_0 : |X_n(\omega) - X(\omega)| > \frac{1}{k}$$

Esto podemos traducirlo en una relación entre conjuntos:

$$\{\omega \in \Omega : X_n(\omega) \not\longrightarrow X(\omega)\} = \bigcup_{k=1}^{\infty} \bigcap_{n_0=1}^{\infty} \bigcup_{n\geq n_0} \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \frac{1}{k}\right\}$$

Como $X_n \xrightarrow{c.s.} X$, este conjunto tiene probabilidad 0. En consecuencia, también tienen probabilidad cero los eventos (más perqueños)

$$A_k = \bigcap_{n_0=1}^{\infty} \bigcup_{n\geq n_0} \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \frac{1}{k}\right\}$$

Como los eventos:

$$B_{k,n_0} = \bigcup_{n\geq n_0} \left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \frac{1}{k}\right\}$$

son decrecientes, deducimos (por la continuidad de la probabilidad) que:

$$A_k = \bigcap_{n_0=1}^{\infty} B_{k,no} \wedge P(A_k) = 0 \Rightarrow \lim_{n_0\to+\infty} P(B_{k,n_0}) = 0$$

Vale decir que si elegimos n_0 suficientemente grande, $P(B_{k,n_0}) < \delta$ En consecuencia,

$$P\left\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \frac{1}{k}\right\} < \delta$$

para todo $n \geq n_0$. Deducimos que X_n tiende en probabilidad a X. □

8.3. Un ejemplo para ver que convergencia en probabilidad no implica convergencia casi segura

Cuando trabajamos con la noción de convergencia casi-segura va a importar cuál es el espacio muestral $(\Omega, \mathcal{E}, P)$.

En este ejemplo, vamos a considerar, el espacio muestral correspondiente al experimento de elegir un número real con distribución uniforme en $[0, 1]$.

- $\Omega = [0, 1]$
- $P(E) = m(E)$. Comentamos que es una **medida σ-aditiva** que extiende la medida elemental de uniones finitas intervalos.
- $\mathcal{E} \subset \mathcal{P}([0, 1])$ será la σ-álgebra de Borel de $[0, 1]$, generada por los sub-intervalos de $[0, 1]$.

Recordamos que una forma de pensarlo es que elegimos los dígitos binarios de un número real en $[0, 1]$ tirando infinitas veces una moneda equilibrada (ensayos de Bernoulli con probabilidad de éxito $1/2$).

Para $n \in \mathbb{N}$, definimos los intervalos

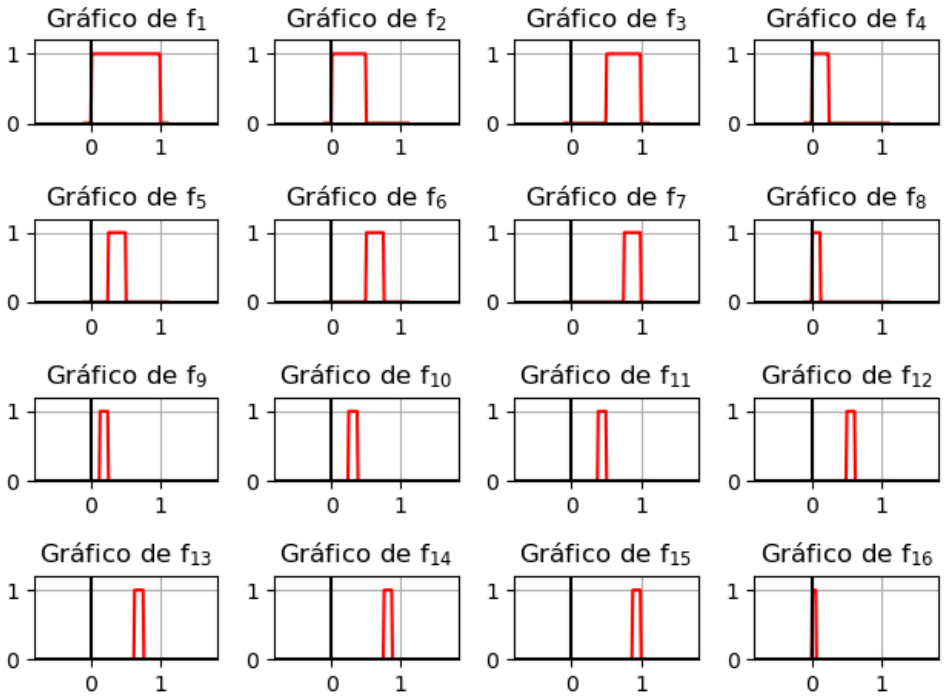
$$J_n = \left[\frac{j}{2^k}, \frac{j+1}{2^k}\right]$$

donde $k = k(n) = [\log_2(n)]$ y $j = j(n)$ cumple que $n = 2^k + j$ con $j \in \{0, 1, 2, \ldots, 2^{j-1}\}$.

Definimos $X_n : [0, 1] \to \mathbb{R}$ como la **función indicadora** del intervalo J_n.

$$X_n(\omega) = \begin{cases} 1 & \text{si } \omega \in J_n \\ 0 & \text{si } \omega \notin J_n \end{cases}$$

¡Cuando usamos el espacio muestral $\Omega = [0, 1]$ las funciones reales se vuelven variables aleatorias!



En este ejemplo, observamos que dado $0 < \delta < 1$,

$$P\{|X_n| > \delta\} = \frac{1}{2^k} \to 0 \text{ cuando } n \to +\infty$$

Entonces

$$X_n \xrightarrow{P} 0$$

Pero dado cualquier $\omega \in [0,1]$, hay infinitos n tales que $\omega \in J_n$, o sea $X_n(\omega) = 1$. Por lo que X_n no converge en forma casi segura a cero.

8.4. El lema de Borel-Cantelli

Lema 8.4.1 (de Borel-Cantelli [Bor09], [Fra17]) *Consideramos una sucesión $(A_n)_{n\in N}$ de eventos, y consideramos el el evento "ocurren infinitos A_n", es decir:*

$$A_\infty = \bigcap_{k\in N} \bigcup_{n\geq k} A_n$$

entonces

i) Si

$$\sum_{n=1}^{\infty} P(A_n) < +\infty \tag{8.1}$$

entonces, con probabilidad 1 *ocurre un número finito de tales sucesos. Es decir*

$$P(A_\infty) = 0$$

ii) Si los A_n *son eventos independientes, y*

$$\sum_{n=1}^{\infty} P(A_n) = +\infty \tag{8.2}$$

entonces, con probabilidad 1 *ocurren infinito s* A_n*. Es decir,*

$$P(A_\infty) = 1$$

Prueba: Demostración de i): Dado $\varepsilon > 0$, teniendo en cuenta la hipótesis (8.1), podemos elegir k tal que

$$\sum_{n=k}^{\infty} P(A_n) < \varepsilon$$

Entonces, por la σ-subaditividad de la probabilidad:

$$P\left(\bigcup_{n\geq k} A_n\right) \leq \sum_{n=k}^{\infty} P(A_n) < \varepsilon$$

y como la probabilidad es creciente:

$$P(A_\infty) \leq P\left(\bigcup_{n\geq k} A_n\right) < \varepsilon$$

Como, ε es arbitrario, deducimos que:

$$P(A_\infty) = 0$$

Demostración de ii): Miremos el complemento de A_∞, que es según las leyes de De Morgan:

$$A_\infty^c = \bigcup_{k\in N} \bigcap_{n\geq k} A_n^c$$

Entonces, tenemos que:

$$P\left(\bigcap_{n=k}^{l} A_n^c\right) = \prod_{n=k}^{l} P(A_n^c) = \prod_{n=k}^{l} P(A_n^c)$$

ya que como los eventos (A_n) son independientes, también lo son sus complementos. Ahora utilizando la desigualdad elemental

$$1 - x \leq e^{-x} \quad x \in [0, 1],$$

tenemos que:

$$P\left(\bigcap_{n=k}^{l} A_n^c\right) \leq \prod_{n=k}^{l} e^{-P(A_n)} = \exp\left(-\sum_{n=k}^{l} P(A_n)\right)$$

y en consecuencia utilizando que la probabilidad es creciente, y la hipótesis (8.2), deducimos que:

$$P\left(\bigcap_{n=k}^{\infty} A_n^c\right) = 0$$

(ya que el segundo miembro de la desigualdad anterior tiende a cero cuando $l \to \infty$). Entonces, por la σ-subaditividad de la probabilidad,

$$P(A_\infty^c) \leq \sum_{k=1}^{\infty} P\left(\bigcup_{k \in N} \bigcap_{n \geq k} A_n^c\right) = 0$$

deducimos que

$$P(A_\infty) = 1$$

□

8.4.1. Un ejemplo para el lema de Borel-Canteli

Ejemplo 8.4.2 *Un mono teclea al azar en una computadora. Supongamos que cada tecla tiene una probabilidad positiva (no necaraiamente todas la misma) de ser pulsada y que las distintas pulsaciones del mono son independientes. Demostrar que con probabilidad 1, el mono eventualmente tecleará el cuento* El Aleph *de Borges (o cualquier otra obra que queramos), infinitas veces.*

Solución: El mono teclea letras de un alfabeto con N caracteres. Cada caracter tiene probabilidad $p_k > 0$ de ser pulsado cada vez que el mono pulsa una tecla, de modo que

$$\sum_{k=1}^{N} p_k = 1, \quad p_k > 0$$

Supongamos que **El Aleph** tiene L caracteres correspondientes a los índices $k_1, k_2, \ldots, k_L > 0$. Dada una secuencia de L caracteres, la probabilidad de que coincida con los caracteres de *El Aleph* será el producto

$$p = p_{k_1} p_{k_2} \ldots p_{k_L} > 0$$

de las correspondientes probabilidades, por la independencia de las pulsaciones. En general, esta probabilidad será extremadamente pequeña. Pero esto vo va a afectar a nuestro argumento.

Ahora dividamos las pulsaciones del mono en bloques de L caracteres, y sea A_n el evento: "el mono teclea el Aleph en el n-ésimo bloque'´. Notmaos que los A_n son eventos independientes y tienen todos probabilidad p. Como $p > 0$ la serie

$$\sum_{n=1}^{\infty} P(A_n)$$

diverge. Entonces por el lema de Borel Cantelli (parte II), con probabilidad 1 ocurrirán infintos de los sucesos A_n, o sea el mono tecleará infinitas veces el Aleph.

Obviamente este ejemplo es una abstracción matemática: en la realidad no funciona, ¡porque la vida del mono no es infinita! ¡Y el lema de Borel-Cantelli no nos dice nada sobre cuánto tiempo tendremos que esperar hasta que el mono teclee por puro azar nuestra obra literaria favorita!.

8.5. Un Criterio para la convergencia casi segura

Como aplicación del lema de Borel-Cantelli, se tiene el siguiente criterio para la convergencia casi segura:

Proposición 8.5.1 *Sea $(X_n) : \Omega \to \overline{\mathbb{R}}$ una sucesión de variables aleatorias, y $X : \Omega \to \mathbb{R}$ otra variables aleatoria. Supongamos que para todo $\varepsilon > 0$,*

$$\sum_{n=1}^{\infty} P\{|X_n - X| > \varepsilon\} < +\infty$$

(o sea, esta serie converge). Entonces

$$X_n \xrightarrow{c.s.} X$$

Prueba: El lema de Borel Cantelli (parte I) implica que si llamamos $A_{n,\varepsilon}$ al evento

$$A_{n,\varepsilon} = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}$$

entonces, con probabilidad 1 ocurren sólo finitos de los sucesos $A_{n,\varepsilon}$, es decir que el evento

$$A_{\infty,\varepsilon} = \bigcap_{k\in N} \bigcup_{n\geq k} A_{n,\varepsilon}$$

tiene probabilidad cero. Tomando $\varepsilon = 1/m$, con $m \in \mathbb{N}$, y usando la σ sub-aditividad de la probabilidad, vemos que el evento:

$$\begin{aligned} B &= \{\omega \in \Omega : X_n(\omega) \not\to X(\omega)\} \\ &= \left\{\omega \in \Omega : \exists m \in \mathbb{N}\ \forall k \in \mathbb{N}\ \exists n \geq k : |X_n(\omega) - X(\omega)| > \frac{1}{m}\right\} \\ &= \bigcup_{m\in\mathbb{N}} A_{\infty,1/m} \end{aligned}$$

tiene probabilidad cero, ya que es la unión numerable de eventos de probabilidad cero. En consecuencia, $P(B^c) = 1$, es decir que

$$X_n \xrightarrow{c.s.} X.$$

□

Como aplicación del lema de Borel-Cantelli, se tiene el siguiente criterio para la convergencia casi segura:

Corolario 8.5.2 *Sea $(X_n) : \Omega \to \overline{\mathbb{R}}$ una sucesión de variables aleatorias, y $X : \Omega \to \mathbb{R}$ otra variables aleatoria. Supongamos que para algún $p > 0$,*

$$\sum_{n=1}^{\infty} E[|X_n - X|^p] < +\infty$$

(o sea, esta serie converge). Entonces

$$X_n \xrightarrow{c.s.} X$$

Prueba: Usando la desigualdad de Markov tenemos que

$$\sum_{n=1}^{\infty} P\{|X_n - X| > \varepsilon\} \leq \sum_{n=1}^{\infty} \frac{E[|X_n - X|^p]}{\varepsilon^p} < +\infty$$

por lo que se deduce del resultado anterior. □

8.6. Un caso especial de la desigualdad de Khinchine

El siguiente lema nos será de utilidad en la prueba de la ley fuerte de los grandes números (con $a_j = 1$), pero lo enuncio así porque nos puede ser útil en algún ejemplo más adelante.

Lema 8.6.1 (Un caso especial de la desigualdad de Khinchine) *Sean (X_k) una sucesión de variables aleatorias independientes con $E[X_k] = 0$ y cuarto momento acotado*

$$E[|X_k|^4] \leq M \text{ donde } M \in \mathbb{R}$$

Entonces si los (a_j) son reales,

$$E\left[\left(\sum_{i=1}^{n} a_i X_i\right)^4\right] \leq 3M\left(\sum_{i=1}^{n} a_i^2\right)^2$$

Prueba: Usando la linealidad de la esperanza, tenemos que

$$E\left[\left(\sum_{i=1}^{n} a_i X_i\right)^4\right] = \sum_{1\leq i_1,i_2,i_3,i_4\leq n} a_{i_1}a_{i_2}a_{i_3}a_{i_4}E[X_{i_1}X_{i_2}X_{i_3}X_{i_4}]$$

Como las X_i son independientes, notamos que

$$E[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] = 0$$

salvo en el caso en que los subíndices son todos iguales, o si son iguales por pares (utilizando que la esperanza del producto es el producto de las esperanzas cuando las variables son independientes, y que la esperanza de cada variable es cero). Notemos que cada término $E[X_i^2X_j^2]$ con $i < j$ aparece $\binom{4}{2} = 6$ veces en esta suma.

Nos queda:

$$E\left[\left(\sum_{j=1}^{n} a_j X_j\right)^4\right] = \sum_{i=1}^{n} a_i^4 E[X_i^4] + 6\sum_{i,j=1,i<j}^{n} a_i^2 a_j^2 E[X_i^2X_j^2]$$

Notamos que por la desigualdad de Jensen

$$E[X_i^2]^2 \leq E[(X_i^2)^2] = E[X_i^4] \leq M$$

Y por otra parte $i \neq j$, X_i^2 es independiente de X_j^2 en consecuencia:

$$E[X_i^2X_j^2] = E[X_i^2]E[X_j^2] \leq M$$

Nos queda:

$$E\left[\left(\sum_{i=1}^{n} a_i X_i\right)^4\right] \leq M\left[\sum_{i=1}^{n} a_i^4 + 6\sum_{i,j=1,i<j}^{n} a_i^2 a_j^2\right] \leq 3M\left[\sum_{i=1}^{n} a_i^2\right]^2$$

□

8.7. La ley fuerte de los grandes números

Teorema 8.7.1 *Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e identicamente distribuidas con $m_4 = E[X_n^4] < +\infty$. Sea $\mu = E[X_i]$ entonces*

$$\frac{X_1 + X_2 + \ldots X_n}{n} \xrightarrow{c.s.} \mu$$

cuando $n \to +\infty$.

Nota: La hipótesis de que el cuarto momento m_4 es finito no es necesaria para la validez de este teorema, pero facilitará enormemente la demostración. Una demostración del teorema sin esta hipótesis (ley fuerte de Kolmogorov) se da en el apéndice G.

Prueba: La idea de la demostración va a ser usar el **criterio para convergencia casi segura** que vimos antes con $p = 4$. Podemos suponer que $\mu = 0$, cambiando sino X_n por $Y_n = X_n - \mu$, ya que

$$\overline{Y}_n = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n} - \mu = \overline{X}_n - \mu$$

con lo que

$$\overline{X}_n \xrightarrow{c.s.} \mu \Leftrightarrow \overline{Y}_n \xrightarrow{c.s.} 0$$

Notamos $S_n = X_1 + X_2 + \ldots + X_n$.

Usando el lema con $a_j = 1$ para todo j, podemos estimar el cuarto momento de S_n:

$$E\left[S_n^4\right] \leq 3m_4\, n^2$$

Deducimos que:

$$E\left[\left(\frac{S_n}{n}\right)^4\right] \leq \frac{C}{n^2}$$

Como la serie

$$\sum_{n=1}^{\infty} \frac{C}{n^2}$$

converge, el **criterio para la convergencia casi segura** que vimos antes (con $p = 4$), implica que

$$\frac{S_n}{n} \xrightarrow{c.s.} 0$$

□

8.7.1. Un ejemplo: La ley fuerte de Borel para ensayos de Bernoulli

Un primer ejemplo que podemos considerar es el esquema de ensayos de Bernoulli, que consideramos en el capítulo 3. Recordamos que en este esquema, un experimento con dos posibles resultados (llamados convencionalmente éxito y fracaso) se repite infinitas veces en condiciones independientes. Llamamos p a la probabilidad del éxito.

Como antes, consideramos entonces las variables aleatorias de Bernoulli:

$$X_i = \begin{cases} 1 & \text{si el i-ésimo experimento fue un éxito} \\ 0 & \text{si el i-ésimo experimento fue un fracaso} \end{cases}$$

Entonces S_n representa la cantidad de éxitos en los n primeros ensayos, y

$$f_n = \frac{S_n}{n}$$

la frecuencia relativa de éxitos en los n primeros ensayos. La ley fuerte de los grandes números afirma entonces que

$$f_n \to p \text{ con probabilidad 1} \tag{8.3}$$

donde llamamos p a la probabilidad del éxito (Este enunciado que se conoce como *la ley fuerte de los grandes números de Borel*, es un caso particular del teorema anterior. Notamos que la hipótesis de que las X_i tengan cuarto momento finito, se satisface trivialmente).

¿Pero qué significa exactamente esto? ¿cuál es el espacio muestral para este experimento compuesto ?. Como dijimos anteriormente, el espacio muestra podemos representarlo como

$$\Omega = \{\omega = (x_1, x_2, \ldots, x_n, \ldots) : \omega_i = 0 \text{ o } \omega_i = 1\} = \{0,1\}^{\mathbb{N}}$$

donde ω_i representará el resultado del i-ésimo ensayo. Entonces, las variables aleatorias X_i se definen sencillamente por:

$$X_i(\omega) = \omega_i$$

Para poder darle sentido a la afirmación (8.3), debemos decir cómo asignamos probabilidades en el espacio Ω. El caso más sencillo es cuando $p = q = 1/2$ (éxito y fracaso equiprobables).

En se caso, definamos para ello la función

$$\phi : \Omega \to [0,1]$$

por

$$\phi(\omega) = \sum_{i=1}^{\infty} \frac{\omega_i}{2^i}$$

En otras palabras, para cada $\omega \in \Omega$, $\phi(\omega)$ será el número en $[0,1]$ cuyo desarrollo binario tiene por dígitos a los ω_i.

Podemos definir entonces la sigma-álgebra $\mathcal{E}$ como:

$$\mathcal{E} = \{E \subset \Omega : \phi(E) \text{es un subconjunto boreliano del intervalo } [0,1]\}$$

y la probabilidad P por

$$P(E) = m(\varphi(E))$$

donde m denota la medida de Lebesgue (ver la discusión en la sección 1.5).

Ejercicio: Comprobar que la función P así definida asigna correctamente las probabilidades, en el sentido de que

$$P(\{\omega \in \Omega : \omega_1 = x_1, \omega_2 = x_2, \ldots, \omega_n = x_n\}) = 2^{-n}$$

donde $k = S_n(\omega)$. En particular, las variables aleatorias X_1, X_2, ..., X_n resultan independientes. Ayuda: notar que $\phi(E)$ consta en este caso de una unión finita de intervalos.

Entonces, cuando $p = 1/2$, la afirmación (8.3) puede interpretarse equivalentemente, como la afirmación de que que para casi todo número en el intervalo $[0,1]$, si f_n designa la frecuencia de dígitos uno en los primeros n lugares de su desarrollo binario, se tiene que $f_n \to 1/2$. En esta afirmación, como es usual en la teoría de la medida, significa "salvo quizás para un conjunto de medida de Lebesgue cero".

8.7.2. Números Normales

Una generalización de la idea anterior es considerar desarrollos en otra base de numeración b, con $b \geq 2$. Entonces pensamos en un experimento cuyos posibles resultados son los dígitos $0, 1, \ldots, b-1$ de la base b, que consideramos equiprobables y lo repetimos infinitas veces.

$$\Omega = D^{\mathbb{N}} \text{ siendo } D = \{0, 1, \ldots, b-1\}$$

Ahora definimos la función

$$\phi : \Omega \to [0,1]$$

por

$$\phi(\omega) = \sum_{i=1}^{\infty} \frac{\omega_i}{b^i}$$

Fijamos un dígito $d \in D$ y nos preguntamos por la frecuencia relativa de ese dígito en los primeros n lugares del número real $x = \phi(\omega)$

$$f_n = \frac{\#\{i : 1 \leq i \leq n, \omega_i = d\}}{n}$$

que podremos escribir como antes en la forma

$$f_n = \frac{S_n}{n}$$

si definimos las variables X_i por

$$X_i = \begin{cases} 1 & \text{si } \omega_i = d \\ 0 & \text{si } \omega_i \neq d \end{cases}$$

Como antes, asignamos las probabilidades en Ω por:

$$P(E) = m(\varphi_b(E))$$

y resulta que

$$P(\{\omega \in \Omega : \omega_1 = d_1, \omega_2 = d_2, \ldots, \omega_n = d_n\}) = b^{-n}$$
$$P(\{\omega \in \Omega : X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}) = p^k q^{n-k}$$

donde $k = S_n(\omega)$, $p = 1/b$, $q = 1 - 1/b$. En particular las variables X_i son de nuevo independientes. Se deduce

$$f_n \to \frac{1}{b} \tag{8.4}$$

con probabilidad 1, o lo que es equivalente f_n tiende a cero para casi todo $x \in [0, 1]$ (o sea: salvo para los x en un conjunto de medida cero en el sentido de Lebesgue). Los números que verifican la relación (8.4) para todo dígito $d \in D$ fueron denominador por Borel *números (simplemente) normales* en la base b. Se deduce de lo demostrado que casi todo número es simplemente normal en la base b.

Más aún, Borel definió los *números absolutamente normales* como aquellos que son simplemente normales en cualquier base $b \geq 2$. Como la unión numerable de conjuntos de medida cero en el sentido de Lebesgue también tiene medida cero, se deduce el siguiente teorema:

Teorema 8.7.2 *(de Borel, [Bor09]) Casi todo número real del intervalo* $[0, 1]$ *es absolutamente normal.*

Nota: Aunque este teorema implica que existen números absolutamente normales, su prueba no es constructiva en el sentido que no nos provee ningún ejemplo de un número absolutamente normal. El primer ejemplo fue dado por Sierpinski en 1916 [Sie17]. Ver también [BF02] para una versión computable de la construcción de Sierpinski.

Capítulo 9

Convergencia en Distribución

Convergencia en Distribución

Definición 9.0.1 *Se dice que una sucesión de variables aleatorias X_n* **converge en distribución** *a la variable aleatoria X, si*

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$$

en cada x en el que F_X sea continua. **Notación:**

$$X_n \xrightarrow{D} X$$

Ejemplo 9.0.2 *Supongamos que $X_n \sim N(0, \sigma_n^2)$ donde $\sigma_n \to 0$. Entonces X_n converge en distribución a la variable aleatoria X con $P\{X = 0\} = 1$, cuya distribución F (que es la función escalón de Heavside) es discontinua en cero. Este ejemplo muestra porqué resulta natural pedir que haya convergencia sólo en los puntos de continuidad de F.*

Proposición 9.0.3 *Si $X_n \xrightarrow{D} X$ y $X_n \xrightarrow{D} Y$, entonces $F_X = F_Y$ (X e Y están idénticamente distribuidas)*

Prueba: $F_X(x) = F_Y(x)$ en cada x que sea simultáneamente punto de continuidad de F_X y F_Y. Pero F_X y F_Y son crecientes, y tienen por lo tanto a lo sumo una cantidad numerable de discontinuidades. Deducimos que $F_X(x) = F_Y(x)$ para los x en un subconjunto denso de $\mathbb{R}$, y entonces para todo x ya que ambas son continuas por la derecha. □

Proposición 9.0.4 *Si $X_n \xrightarrow{D} X$ y $c \in \mathbb{R}$ es una constante, entonces $cX_n \xrightarrow{D} cX$ y $X_n + c \xrightarrow{D} X + c$.*

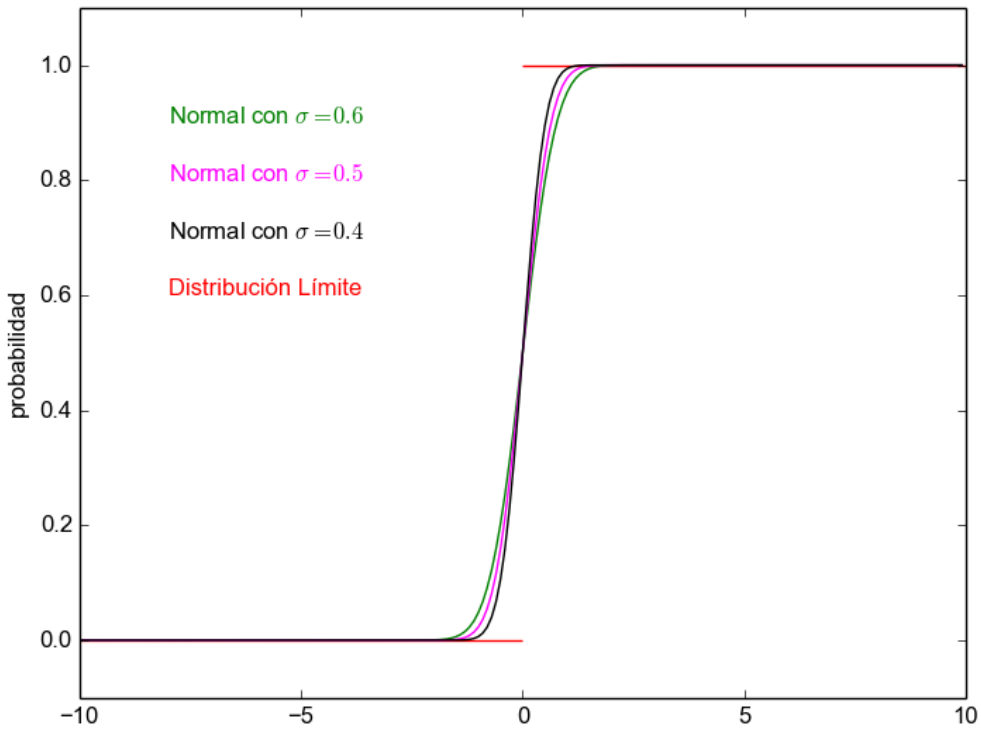


Figura 9.1: Convergencia en distribución de la densidad normal al escalón cuando $\sigma \to 0$.

Observación 9.0.5 *Sin embargo, no es cierto en general que si*

$$X_n \xrightarrow{D} X, Y_n \xrightarrow{D} Y \Rightarrow X_n + Y_n \xrightarrow{D} X + Y$$

Para comprobarlo basta tomar $X_n, Y_n \sim N(0, \sigma^2)$ *independientes,* $X \sim N(0,1)$ *y* $Y = -X$. *Entonces* $X_n + Y_n \sim N(0, 2\sigma^2)$ *que no converge en distribución a cero cuando* $\sigma \to 0$, *aunque* $X + Y = 0$. *Este ejemplo patológico se explica porque la convergencia en distribución se refiere más a las distribuciones de las variables, que a las variables en sí mismas.*

9.1. Relación entre los modos de convergencia

Proposición 9.1.1 *Sea* (X_n) *es una sucesión de variables aleatorias definidas sobre un mismo espacio de probabilidad* $(\Omega, \mathcal{E}, P)$ *y finita son probabilidad 1. Si* $X_n \xrightarrow{P} X$, *entonces*

$X_n \xrightarrow{D} X$.

Prueba: Sea $x_0 \in \mathbb{R}$ tal que F_X sea continua en x_0. Entonces

$$X - |X - X_n| \leq X_n$$

Si $X > x_0 + \varepsilon$ y $|X_n - X| < \varepsilon \Rightarrow X_n > x_0$. Lo podemos traducir en una inclusión de conjuntos:

$$\{X > x_0 + \varepsilon\} \cap \{|X_n - X| < \varepsilon\} \subset \{X_n > x_0\}$$

Tomamos complemento. La inclusión se da vuelta, y usamos las leyes de De Morgan.

$$\{X_n \leq x_0\} \subset \{X \leq x_0 + \varepsilon\} \cup \{|X_n - X| \geq \varepsilon\}$$

Tomamos probabilidad. Usamos que es creciente y subaditiva:

$$P\{X_n \leq x_0\} \leq P\{X \leq x_0 + \varepsilon\} + P\{|X_n - X| \geq \varepsilon\}$$

Esto establece la desigualdad:

$$F_{X_n}(x_0) \leq F_X(x_0 + \varepsilon) + P\{|X_n - X| \geq \varepsilon\}$$

Similarmente

$$X_n \leq X + |X_n - X|$$

Si $X \leq x_0 - \varepsilon$ y $|X_n - X| \leq \varepsilon \Rightarrow X_n \leq x_0$. Lo podemos traducir en una inclusión de conjuntos:

$$\{X \leq x_0 - \varepsilon\} \cap \{|X_n - X| \leq \varepsilon\} \subset \{X_n \leq x_0\}$$

Tomamos complemento. La inclusión se da vuelta, y usamos las leyes de De Morgan.

$$\{X_n > x_0\} \subset \{X > x_0 - \varepsilon\} \cup \{|X_n - X| > \varepsilon\}$$

Tomamos probabilidad. Usamos que es creciente y subaditiva:

$$P\{X_n > x_0\} \leq P\{X > x_0 - \varepsilon\} + P\{|X_n - X| > \varepsilon\}$$

Esto establece la desigualdad:

$$1 - F_{X_n}(x_0) \leq 1 - F_X(x_0 - \varepsilon) + P\{|X_n - X| > \varepsilon\}$$

o

$$F_X(x_0 - \varepsilon) - P\{|X_n - X| > \varepsilon\} \leq F_{X_n}(x_0)$$

Entonces juntando todo tenemos que

$$F_X(x_0-\varepsilon) - P\{|X_n - X| > \varepsilon\} \leq F_{X_n}(x_0) \leq F_X(x_0+\varepsilon) + P\{|X_n - X| \geq \varepsilon\}$$

Entonces como $X_n \xrightarrow{P} X$ por hipótesis,

$$F_X(x_0-\varepsilon) \leq \liminf_{n\to+\infty} F_{X_n}(x_0) \leq \limsup_{n\to+\infty} F_{X_n}(x_0) \leq F_X(x_0+\varepsilon)$$

Y cuanndo $\varepsilon \to 0$, como F_X es continua en x_0,

$$\liminf_{n\to+\infty} F_{X_n}(x_0) = \limsup_{n\to+\infty} F_{X_n}(x_0) = F_X(x_0)$$

Es decir que

$$\lim_{n\to+\infty} F_{X_n}(x_0) = F_X(x_0)$$

como queríamos probar. □

Proposición 9.1.2 *Si $X_n \xrightarrow{D} 0$, entonces $X_n \xrightarrow{P} 0$.*

Prueba: Fijemos $\delta > 0$.

$$\{|X_n| \geq \delta\} = \{X_n \leq -\delta\} \cup \{X_n \geq \delta\}$$

$$\begin{aligned} P\{|X_n| \geq \delta\} &= P\{X_n \leq -\delta\} + P\{X_n \geq \delta\} \\ &= P\{X_n \leq -\delta\} + 1 - P\{X_n < \delta\} \\ &\leq P\{X_n \leq -\delta\} + 1 - P\{X_n \leq \delta/2\} \\ &= F_{X_n}(-\delta) + 1 - F_{X_n}(\delta/2) \end{aligned}$$

Pero por la hipótesis

$$F_{X_n}(t) \to F_0(t) = \begin{cases} 1 & \text{si} \quad t > 0 \\ 0 & \text{si} \quad t < 0 \end{cases}$$

para todo $t \neq 0$. Luego,

$$P\{|X_n| > \delta\} \to 0$$

Como $\delta > 0$ es arbitrario, deducimos que $X_n \xrightarrow{P} 0$. □

9.2. El Teorema de Helly-Bray

Teorema 9.2.1 (Helly) *Supongamos que $F_n : [a,b] \to \mathbb{R}$ es una sucesión de funciones de distribución tales que $F_n(x) \to F(x)$ en cada punto de continuidad de $F(x)$, entonces:*

$$\int_a^b \varphi(x)\, dF_n(x) \to \int_a^b \varphi(x)\, dF(x) \tag{9.1}$$

para toda función continua $\varphi \in C[a,b]$.

Prueba: Dado $\varepsilon > 0$, por el corolario F.0.3 del apéndice F (teorema de existencia para la integral de Riemman-Stieltjes; corolario sobre la convergencia uniforme respecto de la función de distribución), existirá un $\delta > 0$ tal que:

$$\left| \int_a^b \varphi(x)\, dF_n(x) - S_\pi(\varphi, F_n) \right| < \varepsilon$$

para todo n, y también

$$\left| \int_a^b \varphi(x)\, dF(x) - S_\pi(\varphi, F) \right| < \varepsilon$$

para cualquier partición π de $[a,b]$ que verifique que $|\pi| < \delta$ (Pues $F_n(1) - F_n(0) \leq 1$).

Fijemos una partición cualquiera π de $[a,b]$ tal que $|\pi| < \delta$. Claramente podemos elegir los puntos de subdivisión de esta partición π para que sean puntos de continuidad de F (pues el conjunto de puntos de discontinuidad de F es a lo sumo numerable, y por lo tanto su conjunto de puntos de continuidad es denso en $[a,b]$).

Entonces notamos que como hay finitos puntos en la partición, claramente tendremos que:

$$\lim_{n \to +\infty} S_\pi(\varphi, F_n) = S_\pi(\varphi, F)$$

Es decir, que dado $\varepsilon > 0$, existirá un n_0, tal que si $n \geq n_0$,

$$|S_\pi(\varphi, F_n) - S_\pi(\varphi, F)| < \varepsilon$$

En consecuencia, si $n \geq n_0$,

$$\begin{aligned} \left| \int_a^b \varphi(x)\, dF_n(x) - \int_a^b \varphi(x)\, dF(x) \right| &\leq \left| \int_a^b \varphi(x)\, dF_n(x) - S_\pi(\varphi, F_n) \right| \\ &\quad + |S_\pi(\varphi, F_n) - S_\pi(\varphi, F)| \\ &\quad + \left| S_\pi(\varphi, F) - \int_a^b \varphi(x)\, dF(x) \right| \\ &< 3\varepsilon \end{aligned}$$

Como $\varepsilon > 0$ es arbitrario, esto prueba el teorema. □

Un resultado análogo se verifica para integrales en intervalos infinitos:

Teorema 9.2.2 *Supongamos que $F_n : \mathbb{R} \to [0,1]$ es una sucesión de funciones de distribución tales que $F_n(x) \to F(x)$ en cada punto de continuidad de $F(x)$, entonces:*

$$\int_{-\infty}^{\infty} \varphi(x)\ dF_n(x) \to \int_{-\infty}^{\infty} \varphi(x)\ dF(x) \tag{9.2}$$

para toda función continua acotada $\varphi : \mathbb{R} \to \mathbb{R}$.

Prueba: Supongamos que $|\varphi(x)| \leq M\ \forall\ x \in \mathbb{R}$. Dado $\varepsilon > 0$, podemos elegir $R > 0$ tal que:

$$1 - F(R) + F(-R) = \int_{x \leq -R \vee X > R} dF(x) < \frac{\varepsilon}{M}$$

y por lo tanto

$$\left| \int_{|x|>R} \varphi(x)\ dF_n(x) \right| < 2\varepsilon.$$

Además, podemos suponer que R y $-R$ son puntos de continuidad de F. Entonces, como $F_n(R) \to F(R)$ y $F_n(-R) \to F(-R)$ cuando $n \to +\infty$, podemos elegir n_1 tal que para $n \geq n_1$ se verifique

$$F_n(R) - F_n(-R) = \int_{|x|>R} dF_n(x) < \frac{2\varepsilon}{M}$$

y por lo tanto:

$$\left| \int_{|x|>R} \varphi(x)\ dF_n(x) \right| < 2\varepsilon$$

y en virtud del teorema anterior, podemos elegir un n_2 tal que si $n \geq n_2$ se verifica:

$$\left| \int_{-R}^{R} \varphi(x)\ dF_n(x) - \int_{-R}^{R} \varphi(x)\ dF(x) \right| < \varepsilon$$

Entonces, tendremos que:

$$\begin{aligned} \left| \int_{-\infty}^{\infty} \varphi(x)\ dF_n(x) - \int_{-\infty}^{\infty} \varphi(x)\ dF(x) \right| \leq & \left| \int_{-\infty}^{\infty} \varphi(x)\ dF_n(x) - \int_{-R}^{R} \varphi(x)\ dF(x) \right| \\ & + \left| \int_{-R}^{R} \varphi(x)\ dF_n(x) - \int_{-R}^{R} \varphi(x)\ dF(x) \right| \\ & + \left| \int_{-\infty}^{\infty} \varphi(x)\ dF(x) - \int_{-\infty}^{\infty} \varphi(x)\ dF(x) \right| < 4\varepsilon \end{aligned}$$

Como $\varepsilon > 0$ es arbitrario, esto prueba el teorema. □

Corolario 9.2.3 *Si (X_n) es una sucesión de variables aleatorias tales que $X_n \xrightarrow{D} X$, entonces $E[\varphi(X_n)] \to E[\varphi(X)]$ para toda función continua acotada.*

9.3. Un disgresión técnica: Funciones de prueba

Para el siguiente teorema, vamos a usar el **espacio de funciones de prueba**

$$\mathcal{D} = \mathcal{C}_c^\infty(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : f \text{ es } C^\infty \text{ y tiene soporte compacto}\}$$

La condición de que f es C^∞ dice que todas las derivadas $f^{(k)}$ de f existen y son continuas en todo $\mathbb{R}$.

La condición de que f tiene soporte compacto, dice que

$$\text{soporte}(f) = \overline{\{x \in \mathbb{R} : f(x) \neq 0\}}$$

es un conjunto compacto de $\mathbb{R}$, o equivalentemente: existe un intervalo $[a, b]$ tal que $f(x) = 0$ si $x \notin [a, b]$.

A primera vista, parece un espacio muy pequeño. Uno podría pensar que $\mathcal{D} = \{0\}$. ¡Sin embargo vamos a ver que esto no es así!

La función de Cauchy

Para construir una función de prueba no nula, comenzamos considerando la función $f : \mathbb{R} \to \mathbb{R}$ dada por

$$f(x) = \begin{cases} e^{-1/x} & \text{si} \quad x > 0 \\ 0 & \text{si} \quad x \leq 0 \end{cases}$$

Esta función es C^∞ (no tiene soporte compacto). Notamos que

$$f^{(k)}(0) = 0 \text{ para todo } k$$

por lo que el polinomio de Taylor de f de grado k en el origen es el polinomio nulo para todo k, aunque la función f no es idénticamente nula.

Construyendo una función de prueba

Consideremos ahora: $g : \mathbb{R} \to \mathbb{R}$ dada por

$$g(x) = \begin{cases} e^{-1/(1-x^2)} & \text{si} \quad |x| < 1 \\ 0 & \text{si} \quad |x| \geq 1 \end{cases}$$

Vemos que $g \in \mathcal{D}$ y soporte$(g) = [-1, 1]$. Además todas las derivadas de g se anulan en los puntos -1 y 1. Reescalando, dado un intervalo cualquiera $[a, b]$ podríamos construir $g \in \mathcal{D}$ tal que soporte$(g) = I = [a, b]$

Escalones suaves

Consideremos ahora la función $h : \mathbb{R} \to \mathbb{R}$ dada por

$$h(x) = \frac{1}{c} \int_{-1}^{x} g(x)dx \quad \text{para } -1 \leq x \leq 1$$

donde g es la del ejemplo anterior y

$$c = \int_{-1}^{1} g(x)dx$$

y donde $h(x) = 0$ si $x < -1$ y $h(x) = 1$ si $x > 1$. Resulta que h es C^∞ y $0 \leq h(x) \leq 1$. Notemos que $h'(x) = g(x)$ si $x \in (-1, 1)$ por el teorema fundamental del cálculo.

Finalmente, tomando

$$\varphi_\delta(x) = 1 - h\left(\frac{x - x_0}{\delta} - 1\right)$$

podemos probar el siguiente lema que afirma que podemos aproximar la función indicadora $I_{\leq x_0} = I_{(-\infty, x_0]}$ del la semirrecta $(-\infty, x_0]$ por la derecha, por funciones suaves.

Lema 9.3.1 *Para cada $x_0 \in \mathbb{R}$ y cada $\delta > 0$, existe φ_δ de clase C^∞ tal que:*

- $0 \leq \varphi_\delta(x) \leq 1$.
- $\varphi_\delta(x) = 1$ *si* $x \leq x_0$.
- $\varphi_\delta(x) = 0$ *si* $x \geq x_0 + \delta$.

En particular, $\varphi_\delta(x) \geq I_{\leq x_0}(x)$ para todo x.

Similarmente, podemos aproximar $I_{\leq x_0}$ por la izquierda, por funciones suaves, definiendo

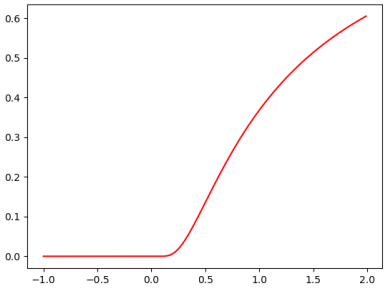
$$\varphi_{-\delta}(x) = 1 - h\left(\frac{x - x_0}{\delta} + 1\right)$$

Lema 9.3.2 *Para cada $x_0 \in \mathbb{R}$ y cada $\delta > 0$, existe $\varphi_{-\delta}$ de clase C^∞ tal que:*

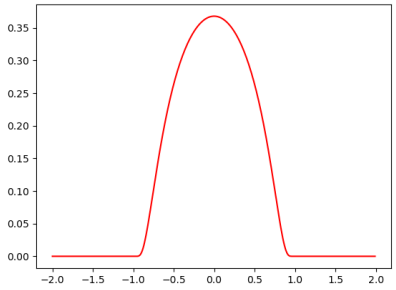
- $0 \leq \varphi_{-\delta}(x) \leq 1$.
- $\varphi_{-\delta}(x) = 1$ *si* $x \leq x_0 - \delta$.
- $\varphi_\delta(x) = 0$ *si* $x \geq x_0$.

En particular, $\varphi_\delta(x) \leq I_{\leq x_0}(x)$ para todo x.

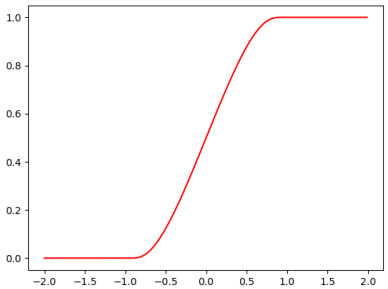
Esta construcción es ilustrada en las figuras siguientes:



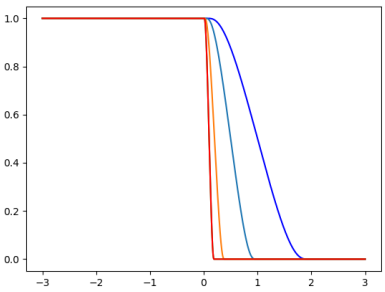
(a) La función de Cauchy f.



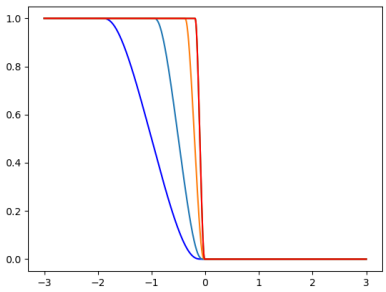
(b) La función g.



(c) La función h.



(d) Las funciones φ_δ.



(e) Las funciones $\varphi_{-\delta}$.

Figura 9.2: Etapas en la construcción de los escalones suaves, y su convergencia a la indicadora de la semirrecta.

9.4. El Recíproco del tereorema de Helly-Bray

Teorema 9.4.1 (Recíproco del teorema de Helly-Bray) *Si (X_n) es una sucesión de variables aleatorias tales que $E[\varphi(X_n)] \to E[\varphi(X)]$ para toda función C^∞ acotada, entonces $X_n \xrightarrow{D} X$.*

Prueba: Tenemos que probar que $F_{X_n}(x_0) \to F_X(x_0)$ cuando $n \to +\infty$, para cada punto de continuidad x_0 de F_X. Para ello, la idea es usar los escalones suaves φ_δ. Primero por la derecha

Dado $\varepsilon > 0$, afirmamos que si δ es suficientemente pequeño,

$$|E[\varphi_\delta(X)] - F_X(x_0)| < \frac{\varepsilon}{2} \tag{9.3}$$

Notamos que:

$$\begin{aligned} E[\varphi_\delta(X)] &= \int_{-\infty}^{\infty} \varphi_\delta(x)\, dF_X(x) \\ &= \int_{-\infty}^{x_0} \varphi_\delta(x)\, dF_X(x) + \int_{x_0}^{x_0+\delta} \varphi_\delta(x)\, dF_X(x) + \int_{x_0+\delta}^{\infty} \varphi_\delta(x)\, dF_X(x) \\ &= \int_{-\infty}^{x_0} 1\, dF_X(x) + \int_{x_0}^{x_0+\delta} \varphi_\delta(x)\, dF_X(x) + \int_{x_0+\delta}^{\infty} 0\, dF_X(x) \\ &= F_X(x_0) + \int_{x_0}^{x_0+\delta} \varphi_\delta(x)\, dF_X(x) \end{aligned}$$

Entonces

$$|E[\varphi_\delta(X)] - F_X(x_0)| = \left| \int_{x_0}^{x_0+\delta} \varphi_\delta(x)\, dF_X(x) \right| \leq F_X(x_0+\delta) - F_X(x_0)$$

acotando la integral de Stieltjes usando el lema 4.2.5) y que $0 \leq \varphi_\delta(x) \leq 1$.

Entonces, la afirmación (9.3) se deduce de la continuidad (por la derecha) de la función de distribución F_X.

Fijamos un $\delta = \delta(\varepsilon)$ tal que se verifique (9.3). Entonces, por la hipótesis, existirá un n_0 tal que si $n \geq n_0$ tenemos que,

$$|E[\varphi_\delta(X_n)] - E[\varphi_\delta(X)]| < \frac{\varepsilon}{2} \tag{9.4}$$

Como consecuencia, usando que $\varphi_\delta(x) \geq I_{\leq x_0}(x)$ deducimos que si $n \geq n_0$, tenemos

que:

$$\begin{aligned} F_{X_n}(x_0) &= P\{X \leq x_0\} = E[I_{\leq x_0}(X)] \\ &\leq E[\varphi_\delta(X_n)] \\ &\leq E[\varphi_\delta(X)] + \frac{\varepsilon}{2} \quad \text{por (9.4)} \\ &\leq F_X(x_0) + \varepsilon \quad \text{por (9.3)} \end{aligned}$$

si $n \geq n_0(\varepsilon)$. Como $\varepsilon > 0$ es arbitrario, hemos probado que

$$\limsup_{n \to +\infty} F_{X_n}(x_0) \leq F_X(x_0) \tag{9.5}$$

Para probar que $F_{X_n}(x_0) \to F_X(x_0)$, necesitamos demostrar también una desigualdad en el sentido contrario.

Para ello, aproximamos $I_{(-\infty, x_0]}$ por escalones suaves desde la izquierda.

El argumento entonces es similar. Usando la continuidad de F en x_0 por la izquierda tendremos que si δ es suficientemente pequeño,

$$|E[\varphi_{-\delta}(X)] - F_X(x_0)| < \frac{\varepsilon}{2} \tag{9.6}$$

ya que

$$|E[\varphi_{-\delta}(X)] - F_X(x_0)| = \left| \int_{x_0 - \delta}^{x_0} \varphi_{-\delta}(x)\, dF_X(x) \right| \leq F_X(x_0) - F_X(x_0 - \delta)$$

Fijamos un $\delta = \delta(\varepsilon)$ tal que se verifique (9.6). Usando la hipótesis, dado $\varepsilon > 0$, existirá un n_0 tal que si $n \geq n_0$ tenemos que,

$$|E[\varphi_{-\delta}(X_n)] - E[\varphi_{-\delta}(X)]| < \frac{\varepsilon}{2} \tag{9.7}$$

Ahora notamos que $\varphi_{-\delta} \leq I_{\leq x_0}$, luego

$$\begin{aligned} F_{X_n}(x_0) &= P\{X_n \leq x_0\} = E[I_{(-\infty, x_0 - \delta]}(X_n)] \\ &\geq E[\varphi_{-\delta}(X_n)] \\ &\geq E[\varphi_{-\delta}(X)] - \frac{\varepsilon}{2} \quad \text{por (9.7)} \\ &\leq F_X(x_0) - \varepsilon \quad \text{por (9.6)} \end{aligned}$$

si $n \geq n_0(\varepsilon)$. Como $\varepsilon > 0$ es arbitrario, hemos probado que

$$\liminf_{n \to +\infty} F_{X_n}(x_0) \geq F_X(x_0) \tag{9.8}$$

Juntando (9.5) y (9.8), hemos probado que:

$$F_{X_n}(x_0) \to F_X(x_0) \quad \text{cuando } n \to +\infty$$

como queríamos. □

9.4.1. Una versión más fuerte

Con un poco más de esfuerzo, vamos a probar una versión más precisa:

Teorema 9.4.2 (Recíproco fuerte del teorema de Helly) *Sean (X_n) es una sucesión de variables aleatorias finitas en casi todo punto, y X otra variable aleatoria finita en casi todo punto, tales que*

$$E[\psi(X_n)] \to E[\psi(X)]$$

para toda función $\psi : \mathbb{R} \to \mathbb{R}$ de clase C^∞ con soporte compacto, entonces $X_n \xrightarrow{D} X$.

La razón por la que será útil para nosotros considerar funciones de prueba C^∞ con soporte compacto en el enunciado de este teorema, es que las utilizaremos en la prueba del teorema 10.9.1.

Un lema previo

Lema 9.4.3 *Si una sucesión de variables aleatorias (X_n) verifica la hipótesis del recíproco fuerte del teorema de Hellly-Bray, entonces es acotada en probabilidad, o equivalentemente (F_{X_n}) es ajustada. Dado, $\varepsilon > 0$ existe N_ε tal que*

$$P\{|X| > N_\varepsilon\} < \varepsilon \text{ para todo } n \in \mathbb{N}$$

Prueba: Por la observación 8.1.7, sabemos que si X es una variable aleatoria finita en casi todo punto. está acotada en probabilidad: dado $\varepsilon > 0$ existe M_ε tal que

$$P\{|X| > M_\varepsilon\} < \varepsilon$$

Elegimos φ C^∞ con soporte en $K_\varepsilon = [-2M_\varepsilon, 2M_\varepsilon]$ tal que $\varphi \geq 1$ en $J_\varepsilon = [-M_\varepsilon, M_\varepsilon]$.

Entonces $I_{K_\varepsilon} \geq \varphi$, luego

$$P\{|X_n| \leq 2M_\varepsilon\} = E[I_{K_\varepsilon}(X_n)] \geq E[\varphi(X_n)]$$

Y por lo tanto si $n \geq n_0(\varepsilon)$ tendremos

$$P\{|X_n| \leq 2M_\varepsilon\} \geq E[\varphi(X)] - \varepsilon$$

Pero $\varphi \geq I_{J_\varepsilon}$, luego

$$E[\varphi(X)] \geq E[I_{J_\varepsilon}(X)] = P\{|X| \leq M_\varepsilon\} \geq 1 - \varepsilon$$

Entonces

$$P\{|X_n| \leq 2M_\varepsilon\} \geq 1 - 2\varepsilon$$

o sea:

$$P\{|X_n| > 2M_\varepsilon\} \leq \varepsilon$$

□

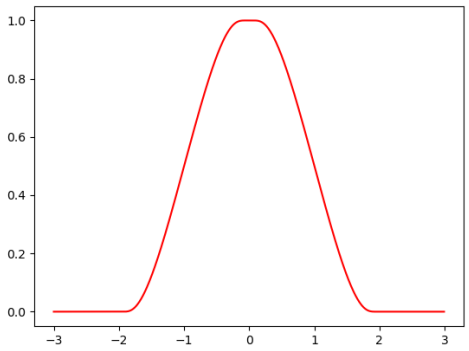


Figura 9.3: Por ejemplo, acá vemos el gráfico de φ si $M_\varepsilon = 1$, con lo que $K_\varepsilon = [-2, 2]$ y $J_\varepsilon = [-1, 1]$

Demostración de la versión fuerte del recíproco del teorema de Helly-Bray

Prueba: Sabemos que

$$E[\psi(X_n)] \to E[\psi(X)]$$

para toda función $\psi : \mathbb{R} \to \mathbb{R}$ de clase C^∞ con soporte compacto. Vamos a probar que

$$E[\varphi(X_n)] \to E[\varphi(X)]$$

para toda φ de clase C^∞ acotada. Con lo que la versión débil del teorema va implicar la versión fuerte.

Fijemos φ. Supongamos que $|\varphi(x)| \leq C$ para todo x.

Dado $\varepsilon > 0$, consideramos ρ de clase C^∞ tal que $\rho(x) = 1$ en $[-N_\varepsilon, N_\varepsilon]$ y $0 \leq \rho(x) \leq 1$ siempre, consideramos $\phi = \rho \cdot \varphi$ y escribimos

$$\begin{aligned} |E[\varphi(X)] - E[\varphi(X_n)]| \leq\ & |E[\varphi(X)] - E[\psi(X)]| \\ & + |E[\psi(X)] - E[\psi(X_n)]| + |E[\varphi(X_n)] - E[\psi(X_n)]| \end{aligned}$$

Acotemos:

$$\begin{aligned} |E[\varphi(X_n)] - E[\psi(X_n)]| &= \left| \int_{-\infty}^{\infty} [\varphi(x) - \psi(x)]\ dF_n(x) \right| \\ &\leq \int_{|x|>N_\varepsilon} (1 - \rho(x)) \cdot |\varphi(x)|\ dF_n(x) \\ &\leq \int_{|x|>N_\varepsilon} C\ dF_n(x) = CP\{|X_n| > N_\varepsilon\} < C\varepsilon \end{aligned}$$

para todo n por el lema (¡Es una cota uniforme!). Similarmente:

$$|E[\varphi(X)] - E[\psi(X)]| < C\varepsilon$$

y finalmente por la hipótesis

$$|E[\psi(X)] - E[\psi(X_n)]| < \varepsilon \text{ si } n \geq n_0(\varepsilon)$$

Luego reemplazando en la desigualdad triangular que teníamos antes

$$|E[\varphi(X)] - E[\varphi(X_n)]| < (2C+1)\varepsilon \text{ si} n \geq n_0(\varepsilon)$$

Entonces vemos que:

$$E[\varphi(X_n)] \to E[\varphi(X)]$$

Como esto vale para toda φ de clase C^∞ acotada, por la vesión débil del terema deducimos que:

$$X_n \xrightarrow{D} X$$

como queríamos. □

Un ejemplo de aplicación del teorema de Helly-Bray

Proposición 9.4.4 *Sea $D \subset \mathbb{R}$ un conjunto discreto (=sin puntos de acumulación) y sean $X_n, X : \Omega \to \mathbb{R}$ variables aleatorias concentradas en D. Llamemos*

$$p_n(k) = P\{X_n = k\}, \quad p(k) = P\{X = k\}$$

Entonces

$$X_n \xrightarrow{D} X \Leftrightarrow p_n(k) \to p(k) \text{ para todo } k \in D$$

Esta enunciado generaliza el ejercicio 2 de la práctica 8, que corresponde al caso especial $D = \mathbb{Z}$. Vamos a resolverlo usando el teorema de Helly-Bray (aunque podría resolverse usando la definición de convergencia en distribución).

Prueba: Supongamos primero que $X_n \xrightarrow{D} X$. Dado un $k \in D$ consideramos un entono abierto U de k donde k sea el único punto de D (existe por la hipótesis de que D es discreto).

Consideramos $\varphi : \mathbb{R} \to \mathbb{R}$ de clase C^∞ con soporte contenido en U tal que $\varphi(k) = 1$. Entonces por Helly-Bray

$$E[\varphi(X_n)] \to E[\varphi(X)]$$

pero $\varphi(X_n)$ es una variable aleatoria discreta, pues X_n está concentrada en D. Luego

$$E[\varphi(X_n)] = \sum_{d \in D} \varphi(d) \cdot p_n(d) = p_n(k)$$

Similarmente

$$E[\varphi(X)] = p(k)$$

Se deduce que $p_n(k) \to p(k)$. Esto vale para todo $k \in D$.

Recíprocamente, supongamos que $p_n(k) \to p(k)$ para todo $k \in D$. Queremos ver que $X_n \xrightarrow{D} X$. Para eso, vamos a usar el recíproco fuerte del teorema de Helly-Bray. Luego queremos probar que

$$E[\psi(X_n)] \to E[\psi(X)] \tag{9.9}$$

para toda $\psi : \mathbb{R} \to \mathbb{R}$ de clase C^∞ con soporte compacto. Llamemos K al soporte de ψ. Entonces $D \cap K$ es finito, y

$$E[\psi(X_n)] = \sum_{d \in D \cap K} \psi(d) \cdot p_n(d)$$

Similarmente

$$E[\psi(X)] = \sum_{d \in D \cap K} \psi(d) \cdot p(d)$$

Como son sumas finitas, es claro que (9.9) se va a cumplir, ya que el límite de una suma finita es igual a la suma de los límites. □

Otra aplicación del teorema de Helly-Bray

Corolario 9.4.5 *Si $X_n \xrightarrow{D} X$ y $g : \mathbb{R} \to \mathbb{R}$ es una función continua, entonces $g(X_n) \xrightarrow{D} g(X)$.*

Prueba: Por el recíproco del teorema de Helly-Bray nos bastará probar que

$$E[\psi(g(X_n))] \to E[\psi(g(X))] \text{ para toda } \psi \in C_c^\infty(R)$$

Esto es:

$$E[\varphi(X_n)] \to E[\varphi(X)]$$

donde $\varphi = \psi \circ g$. Pero notamos que ψ es continua y acotada, por ser composición de continuas y ψ acotada. Deducimos que esto es cierto, en virtud del teorema de Helly-Bray. □

Corolario 9.4.6 *Si $X_n \xrightarrow{D} X$, y a, b son constantes, entonces $aX_n + b \xrightarrow{D} aX + b$.*

Observación 9.4.7 *Sin embargo, no es cierto en general que si*

$$X_n \xrightarrow{D} X, Y_n \xrightarrow{D} Y \Rightarrow X_n + Y_n \xrightarrow{D} X + Y$$

9.5. El teorema de Slutsky

9.5.1. Una versión simple del teorema

Lema 9.5.1 *Sean (X_n) e (Y_n) dos sucesiones de variables aleatorias finitas con probabilidad 1. Supongamos que $X_n \xrightarrow{D} X$ e $Y_n \xrightarrow{P} 0$. Entonces $X_n + Y_n \xrightarrow{D} X$.*

En el apunte de Victor Yohai [Yoh] pueden ver una prueba usando directamente la definición de convergencia en distribución. Yo les voy a presentar una prueba alternativa usando la caracerización dada por el teorema de Helly-Bray.

Prueba: Usando la caracerización dada por el teorema de Helly-Bray, queremos probar que para toda $\psi \in C_c^\infty(\mathbb{R})$,

$$E[\psi(X_n + Y_n)] \to E[\psi(X)$$

y sabemos por hipótesis que:

$$E[\psi(X_n)] \to E[\psi(X)]$$

Luego nos bastará probar que para cada ψ fija,

$$E[\psi(X_n + Y_n)] - E[\psi(X_n)] \to 0$$

Notamos que como $\psi \in C_c^\infty(\mathbb{R})$, ψ será acotada

$$|\psi(x)| \leq C \quad \text{para todo } x, y \in \mathbb{R}$$

y cumplirá la **condición de Lipschitz**

$$|\psi(x) - \psi(y)| \leq M|x - y| \quad \text{para todo } x \in \mathbb{R}$$

donde M es cualquier cota de $|\psi'|$ (por el teorema del valor medio).

Usando las observaciones anteriores, tenemos que dado $\varepsilon > 0$,

$$|\psi(X_n + Y_n) - \psi(X_n)| \leq M|Y_n| < \frac{\varepsilon}{2}$$

si

$$|Y_n| < \delta = \frac{\varepsilon}{2M}$$

Entonces, introducimos los eventos:

$$A_{n,\delta} = \{\omega \in \Omega : |Y_n(\omega)| < \delta\}$$

y podemos estimar:

$$E[|\psi(X_n + Y_n)] - E[\psi(X_n)| \cdot I_{A_{n,\delta}}] \leq \frac{\varepsilon}{2}$$

Ahora vamos a necesitar mirar que pasa en

$$A_{n,\delta}^c = \{\omega \in \Omega : |Y_n(\omega)| \geq \delta\}$$

Ahí vamos a usar la estimación más bruta

$$|\psi(X + Y_n) - \psi(X_n)| \leq 2C$$

Entonces:

$$E[|\psi(X_n + Y_n) - \psi(X_n)| \cdot I_{A_{n,\delta}^c}] \leq 2C \cdot E[I_{A_{n,\delta}^c}] = 2C \cdot P(A_{n,C}) < \frac{\varepsilon}{2}$$

si $n \geq n_0(\varepsilon, \delta)$ pues $Y_n \xrightarrow{P} 0$. Pero $\delta = \delta(\varepsilon)$, así que en definita n_0 depende sólo de ε.

Finalmente acotamos

$$\begin{aligned} |E[\psi(X_n + Y_n)] - E[\psi(X_n)]| &\leq E[|\psi(X_n + Y_n)] - E[\psi(X_n)|] \\ &\leq E[|\psi(X_n + Y_n)] - E[\psi(X_n)|I_{A_{n,\delta}}] \\ &+ \leq E[|\psi(X_n + Y_n)] - E[\psi(X_n)|I_{A_{n,\delta}^c}] \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

si $n \geq n_0$. O sea, que efectivamente hemos probado que:

$$E[\psi(X_n + Y_n)] - E[\psi(X_n)] \to 0$$

y como observamos antes, esto implica la validez del lema. □

9.5.2. Un lema para el teorema de Slutsky

Lema 9.5.2 *Sea (X_n) una sucesión de variables aleatorias finitas con probabilidad 1, tales que $X_n \xrightarrow{P} c$ donde $c \in \mathbb{R}$ es una constante. Entonces si $g : \mathbb{R} \to \mathbb{R}$ es una función boreliana continua en c, entonces:*

$$Y_n = g(X_n) \xrightarrow{P} g(c)$$

Prueba: Dado $\varepsilon > 0$ por definición de continuidad, existirá un $\delta > 0$ tal que $|x - c| \leq \delta$ implica $|g(x) - g(c)| \leq \varepsilon$. Luego,

$$\{|g(x) - g(c)| \geq \varepsilon\} \subset \{|x - c| \geq \delta\}$$

En particular,

$$\{|g(X_n) - g(c)| \geq \varepsilon\} \subset \{|X_n - c| \geq \delta\}$$

tomando probabilidades:

$$0 \leq P\{|g(X_n) - g(c)| \geq \varepsilon\} \leq P\{|X_n - c| \geq \delta\}$$

por lo que si el lado derecho tiende a cero cuando $n \to +\infty$, también el término del medio. O sea que si $X_n \xrightarrow{P} c$, se deduce que $g(X_n) \xrightarrow{P} g(c)$.

□

9.5.3. El Teorema de Slutksky

Teorema 9.5.3 *Sean (X_n) e (Y_n) dos sucesiones de variables aleatorias finitas con probabilidad 1. Supongamos que $X_n \xrightarrow{D} X$ e $Y_n \xrightarrow{P} c$ donde X es otra variable aleatoria finita con probabilidad 1 y c una constante. Entonces,*

- $X_n + Y_n \xrightarrow{D} X + c$.
- $X_n Y_n \xrightarrow{D} cX$.
- *Si $c \neq 0$,*

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}$$

Prueba: Para probar que $X_n + Y_n \xrightarrow{D} X + c$, escribimos:

$$X_n + Y_n = (X_n + c) + (Y_n - c)$$

Comomo $X_n \xrightarrow{D} X$, tendremos que $X_n + c \xrightarrow{D} X + c$ por los resultados previos. También

$$Y_n \xrightarrow{P} c \Rightarrow Y_n - c \xrightarrow{P} 0$$

(esto sale directamente la definición).

El resultado se deduce entonces de la versión simple del teorema de Slutsky que probamos antes (lema 8.1.9).

Similarmente, para ver que $X_n Y_c \xrightarrow{D} cX$, escibimos

$$X_n Y_n = cX_n + (Y_n - c)X_n = U_n + Z_n$$

donde llammamos $U_n = cX_n$ y $Z_n = (Y_n - c)X_n$.

Como $X_n \xrightarrow{D} X \Rightarrow U_n \xrightarrow{D} cX$ por los resultados previos.

También sabemos que $Y_n - c \xrightarrow{P} 0$. Por otra parte, (X_n) está acotada en probabilidad, ya que converge en distribución (Esto se deduce del lema 8.1.9, aunque puede probarse directamente).

Pero entonces $Z_n \xrightarrow{P} 0$ ya que es el producto que una sucesión que tiende a cero en probabilidad por una que está acotada en probabilidad (lema 8.1.9), y por lo tanto $Z_n \xrightarrow{D} 0$.

Entonces usando la versión simple del teorema de Slutsky, concluimos que $X_n Y_n \xrightarrow{D} cX$.

Finalmente, para ver que si $c \neq 0$,

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}$$

escribimos

$$\frac{X_n}{Y_n} = X_n \cdot \frac{1}{Y_n}$$

y observamos que

$$Y_n \xrightarrow{P} c \Rightarrow \frac{1}{Y_n} \xrightarrow{P} \frac{1}{c}$$

por el lema previo aplicado a la función $g(y) = \frac{1}{y}$ que es continua en $y = c$ si $c \neq 0$. Entonce el resultado se deduce del ítem anterior. □

Capítulo 10

Funciones características

10.1. Esperanza de variables aleatorias con valores complejos

Notemos que que podemos considerar variables aleatorias con valores complejos $X : \Omega \to \mathbb{C}$, en lugar de con valores reales como hemos hecho hasta ahora. Escribiendo $X = A + Bi$ donde A y B son la parte real e imaginaria de X, no ofrece ninguna dificultad extender la definición de esperanza para ellas, escribiendo

$$E(X) = E(A) + iE(B)$$

Las propiedades de la esperanza se generalizan fácilmente para estas variables.

Lema 10.1.1 *Si $X : \Omega \to \mathbb{R}$ es una variable aleatoria con valores complejos tal que $E[|X|] < +\infty$, entonces $E[X]$ está bien definida y*

$$|E[X]| \leq E[|X|] \tag{10.1}$$

Prueba: Como $|A| \leq |X|, |B| \leq |X|$ se deduce que

$$E[|A|] \leq E[|X|] < \infty, E[|B|] \leq E[|X|] < \infty$$

luego $E[X]$ está bien definida. Probemos la desigualdad (10.1). Si $E[X]) = 0$ no hay nada que probar. Sino, la escribimos en forma polar

$$E[X] = r \cdot e^{i\theta} \quad \text{con } \theta \in \mathbb{R}$$

donde

$$r = |E(X)| = e^{-i\theta} E[X] \in \mathbb{R}_{>0}$$

Entonces

$$r = |\mathrm{Re}(E[e^{-i\theta} X])| = |E[\mathrm{Re}(e^{-i\theta} X)]| \leq E[|\mathrm{Re}(-e^{i\theta} X)|] \leq E[|-e^{i\theta} X|] = E[|X|]$$

usando que ya sabemos que (10.1) es válida para variables aleatorias reales, y que

$$|\mathrm{Re}(z)| \leq |z| \text{ para todo } z \in \mathbb{C}$$

□

Similarmente, si $f : [a, b] \to \mathbb{C}$ es una función continua, escribimos $f(t) = x(t) + iy(t)$ donde $x, y : [a, b] \to \mathbb{R}$. Y definimos

$$\int_a^b f(t)\, dt = \int_a^b x(t)\, dt + i \int_a^b y(t)\, dt$$

De nuevo tenemos

$$\left| \int_a^b f(t)\, dt \right| \leq \int_a^b |f(t)|\, dt$$

Esta desigualdad la podemos pensar como

$$|E[f(U)| \leq E(|f(U)|) \text{ donde } U \sim \mathcal{U}(a, b)$$

También podemos definir

$$\frac{d}{dt} f(t) = \frac{d}{dt} x(t) + i \frac{d}{dt} y(t)$$

si a y b son derivables.

10.2. Funciones Características

Para la siguiente definición, recordemos que para $x \in \mathbb{R}$, la función exponencial e^{ix} de exponente imaginario puro puede definirse por medio de la fórmula de Euler

$$e^{ix} = \cos x + i \operatorname{sen} x$$

que puede justificarse a partir de los correspondientes desarrollos de Taylor.

Definición 10.2.1 *Si X es una variable aleatoria tal que $E(|X|)$ es finita, su* función característica *se define por*

$$\varphi_X(t) = E\left[e^{itX}\right] \quad t \in \mathbb{R}$$

Teniendo en cuenta la definición de la esperanza, esto puede escribirse como

$$\varphi_X(t) = \int_\infty^\infty e^{itx}\, dF_X(x)$$

siendo F_X la función de distribución de X, y si X es una variable con densidad continua $f_X(x)$ entonces

$$\varphi_X(t) = \int_{\infty}^{\infty} e^{itx}\, f_X(x)\, dx$$

Notemos entonces que en la teoría de probabilidades se llama función característica a lo que en muchos otros contextos de la matemática se conoce como transformada de Fourier*. De hecho, existe toda una rama de la matemática dedicada al estudio de este tipo de transformadas, el* análisis armónico*. Para nosotros, serà una herramienta ùtil para estudiar la convergencia en distribución de las variables aleatorias (ver el teorema de continuidad en la sección siguiente).*

Observemos también que la función característica sólo depende de la distribución de la variable aleatoria X, por lo que tiene sentido hablar de funciones característica de una determinada distribución de probabilidades F. Por eso, a veces escribiremos φ_F en lugar de φ_X para enfatizar este hecho.

Observación 10.2.2 *Si X es una variable aleatoria discreta que toma valores en $\mathbb{N}_0$, tenemos que*

$$\varphi_X(t) = \sum_{k=0}^{\infty} e^{itk} P\{X = k\} = \sum_{k=0} (e^{it})^k P\{X = k\} = g_X(e^{it})$$

donde g_X es la función generatriz que introdujimos en la sección 3.7. Por ejemplo, usando esto deducimos que:

- *Si $X \sim Bi(n,p) \Rightarrow \varphi_X(t) = (p+qe^{it})^n = \left(1 + p(e^{it} - 1)\right)^n$ donde $q = 1-p$, por (3.6).*
- *Si $X \sim \mathcal{P}(\lambda) \Rightarrow \varphi_X(t) = e^{\lambda(\exp(it)-1)}$ por (3.7).*
- *Si $X \sim Ge(p) \Rightarrow \varphi_X(t) = \frac{pe^{it}}{1-qe^{it}}$ donde $q = 1-p$, por (3.11)*

10.2.1. Funciones características de variables aleatorias continuas

Si X es una variable aleatoria absolutamente continua con densidad de probabilidad $f(x)$, entonces

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} f(x) e^{itx}\, dx$$

La función

$$\widehat{f}(t) = \mathcal{F}(f)(t) = \int_{-\infty}^{\infty} f(x) e^{itx}\, dx$$

se llama **transformada de Fourier** de la función f. Está definida para cualquier $f : \mathbb{R} \to \mathbb{C}$ tal que

$$\int_{-\infty}^{\infty} |f(x)|\, dx < \infty \quad \textbf{Notación: } f \in L^1(\mathbb{R})$$

Para ser precisos, acá tendríamos que usar la **integral de Lebesgue** que se ve en los cursos de análisis real. Pero en esta materia, lo usaremos simplemente como una notación (pueden pensar la integral como una integral impropia).

Ejemplo 10.2.3 *Para la distribución uniforme, la función característica puede determinarse a partir de la definición. Si* $X \sim \mathcal{U}(a,b)$*, entonces*

$$\varphi_X(t) = \int_a^b e^{itx} \, \frac{dx}{b-a} = \frac{e^{itb} - e^{ita}}{it(b-a)}$$

En particular cuando $X \sim \mathcal{U}(-1,1)$

$$\varphi_X(t) = \frac{e^{it} - e^{.it}}{2it} = \frac{\sin t}{t}$$

Notemos que $\varphi_X \notin L^1(\mathbb{R})$ *pues*

$$\int_{-\infty}^{\infty} \left| \frac{\sin t}{t} \right| \, dt = +\infty$$

Algunas observaciones:

- Existe toda una rama de la matemática dedicada al estudio de las series de Fourier y la transforma de Fourier, el **análisis armónico**.
- Para nosotros, será una herramienta útil para estudiar la convergencia en distribución de las variables aleatorias, y nos permitirá probar uno de los resultados centrales de la teoría de probabilidades: el teorema del límite central.
- Pero las series y transformadas de Fourier tiene innumerables aplicaciones en muchas ramas de la matemática y la física: ecuaciones diferenciales, análisis de señales, ondas, procesamiento de imágenes, mecánica cuántica, teoría de números, etc.
- De hecho, Joseph Fourier introdujo sus series para estudiar la propagación del calor en una barra de metal, que describió por medio de una ecuación diferencial (eso lo van a ver en el curso de ecuaciones diferenciales).
- Por eso es una herramienta que vale la pena aprender, más allá de la aplicación inmediata en la que estamos interesados (a la teoría de probabilidades).

Para los lectores interesados en saber más sobre series y transformadas de Fourier recomiendo [Duo03].

10.2.2. Propiedades de las funciones características

Proposición 10.2.4 *La función característica de una variable aleatoria X con $E(|X|) < \infty$ tiene las siguientes propiedades:*

i) La función característica $\varphi_X(t)$ es uniformemente continua.

ii)
$$|\varphi_X(t)| \leq 1$$

iii)
$$\varphi_X(0) = 1$$

iv) Si hacemos un cambio lineal de variable, $Y = aX + b$
$$\varphi_Y(t) = e^{itb}\varphi_X(ta)$$

Prueba: Probemos i):

$$\begin{aligned} |\varphi_X(t+h) - \varphi_X(t)| &= |E[e^{i(t+h)X}] - E[e^{itX}]| & &= |E[e^{i(t+h)X} - e^{itX}]| \\ &= |E[e^{itX} \cdot e^{ihX} - e^{itX})| & &= |E[e^{itX} \cdot (e^{ihX} - 1)| \\ &\leq E[|e^{itX}| \cdot |e^{ihX} - 1|] & &= E[|e^{ihX} - 1|] \\ &\leq E[|hX|) \\ &= |h|E(|X|) < \varepsilon \end{aligned}$$

si
$$|h| < \delta = \frac{\varepsilon}{E|X|}$$
(si $E[|X|] = 0$, $X = 0$ con probabilidad 1 y $\varphi_X \equiv 1$].

ii) Es inmediata pues
$$|\varphi_X(t)| = |E(e^{itX})| \leq E(|e^{itX})| = E(1) = 1$$

iii) También es inmediata pues
$$\varphi_X(0) = E(e^{i0}) = E(1) = 1$$

Para probar iv) notamos que
$$E(Y) = E(e^{itY}) = E(e^{it(aX+b)}] = E[e^{itaX}e^{itb}] = e^{itb}\varphi_X(ta)$$

□

Proposición 10.2.5 *Si X e Y son variables aleatorias independientes con esperanza finita entonces*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

Prueba: Como X e Y son independientes, e^{itX} y e^{itY} también lo son entonces

$$\varphi_{X+Y}(t) = E^{it(X+Y)} = E[e^{itX}]E[e^{itY}] = \varphi_X(t)\varphi_Y(t)$$

□

Proposición 10.2.6 *Sea $k \in \mathbb{N}$.Si $E(|X|^k) < \infty$, entonces $\varphi_X(t)$ es de clase C^k y*

$$\varphi_X^{(k)}(t) = E((iX)^k e^{itX})$$

En particular

$$\varphi_X^{(k)}(t) = i^k m_k(X)$$

donde

$$\mu_k(X) = E(X^k)$$

es el k-ésimo momento de la variable X (respecto del origen).

Prueba: Se obtiene derivando bajo el signo de esperanza. Para justificar esto, se requiere un teorema de derivación de integrales con respecto a un parámetro, que se ve en análisis real. □

Ejemplo 10.2.7 *Si $X \sim \Gamma(\alpha, \lambda)$, su función característica viene dada por*

$$\begin{aligned}\varphi(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{itx}\ x^{\alpha-1}\ e^{-\lambda x}\ dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1}\ e^{-(\lambda-it)x}\ dx\end{aligned}$$

Usando la fórmula (4.13) (que sigue valiendo para valores complejos de λ con $Re(\lambda) > 0$) se deduce que [1]

$$\varphi_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{\lambda - it} = \left(\frac{\lambda}{\lambda - it}\right)^\alpha \tag{10.2}$$

Observación 10.2.8 *Cuando X es una variable aleatoria con una densidad integrable, se tiene que*

$$\varphi_X(t) \to 0 \text{ cuando } |t| \to \infty$$

en virtud del lema de Riemann-Lebesgue (un resultado importante del análisis armónico). Sin embargo, esta propiedad no es cierta para variables aleatorias cualesquiera. Por ejemplo, si X es una variable aleatoria, tal que $X = 0$ con probabilidad 1, entonces $\varphi_X(t) \equiv 1$.

[1] Se requieren algunos conocimientos de análisis complejo para darle sentido a esta fórmula, z^α se puede definir en el plano complejo menos el eje real negativo, usando la fórmula $z^\alpha = \exp(\alpha \log(z))$ y tomando la rama principal del logaritmo.

10.3. La Función Característica de la Distribución Normal

El siguiente teorema es clave para la prueba que haremos del teorema central del límite, uno de los resultados fundamentales de la teoría de probabilidades:

Teorema 10.3.1 *Si* $X \sim N(\mu, \sigma^2)$, *entonces* $\varphi_X(t) = e^{it\mu}e^{-(\sigma t)^2/2}$

Existen varias pruebas de este teorema. Presentaré una prueba que aprendí en el curso de V. Yohai que utiliza argumentos probabilísticos. Notemos que el teorema dice esencialmente que la densidad normal estándar es un punto fijo de la transformada de Fourier. Hay también demostraciones que utilizan argumentos de análisis complejo o de ecuaciones diferenciales. La idea de dicha prueba es usar las propiedades de invariancia de la distribución normal para obtener una ecuación funcional para la función característica buscada.

Prueba: Usando el resultado del ejemplo 4.4.1, vemos que basta probarlo para la variable normalizada

$$X^* = \frac{X - \mu}{\sigma}$$

que tiene distribución $N(0, 1)$.

Consideramos entonces dos variables aleatorias $X, Y \sim N(0, 1)$ independientes, y sea $Z = aX + bY$, con $a, b > 0$. Tendremos entonces

$$\varphi_Z(t) = \varphi_{aX}(t)\varphi_{bY}(t) = \varphi_X(ta)\varphi_Y(tb)$$

y como la función característica sólo depende de la distribución esto es igual a

$$\varphi_Z(t) = \varphi_X(at)\varphi_X(tb)$$

Por otra parte, sabemos por la proposición 4.5.6 y el ejemplo 4.4.1, que

$$Z \sim N(0, a^2 + b^2)$$

Entonces de nuevo por el ejemplo 4.4.1,

$$Z^* = \frac{Z}{\sqrt{a^2 + b^2}} \sim N(0, 1)$$

y se deduce utilizando el item iv) de la proposición 10.2.4 que

$$\varphi_Z(t) = \varphi_X\left(\sqrt{a^2 + b^2}\, t\right)$$

Comparando las dos expresiones para $\varphi_Z(t)$ obtenemos la ecuación funcional buscada:

$$\varphi_X\left(\sqrt{a^2 + b^2}\, t\right) = \varphi_X(at)\varphi_X(tb)$$

En particular eligiendo $t = 1$, tenemos que

$$\varphi_X\left(\sqrt{a^2 + b^2}\right) = \varphi_X(a)\varphi_X(b)$$

Llamemos $\psi(s) = \varphi_X(\sqrt{s})$. Entonces

$$\psi(a^2 + b^2) = \psi(a^2)\psi(b^2)$$

y poniendo $a = \alpha^2, b = \beta^2$ deducimos que

$$\psi(\alpha + \beta) = \psi(\alpha)\psi(\beta) \quad \text{para todo } \alpha, \beta \geq 0$$

(Si α o β son cero, esto vale pues $\varphi_X(0) = 1$). Entonces por el lema 4.8.1, deducimos que

$$\psi(t) = e^{tb} \quad \text{para algún } b \in \mathbb{R}$$

ya que $\psi(0) = 1$, y por lo tanto

$$\varphi_X(t) = e^{bt^2}$$

Para encontrar el valor de b, derivamos dos veces

$$\varphi'_X(t) = 2bt\ e^{bt^2}$$

$$\varphi''_X(t) = (2b + 2bt)\ e^{bt^2}$$

En particular,

$$\varphi''_X(0) = 2b = -\mu_2(X)$$

por la proposición 10.2.6. Pero

$$\mu_2(X) = Var(X) = 1$$

luego $b = -1/2$, y obtenemos que

$$\varphi_X(t) = e^{-t^2/2}$$

□

10.4. La identidad de Plancherel

Lema 10.4.1 (Identidad de Plancherel) *Sea X una variable aleatoria con $E(|X|) < +\infty$, función de distribución F_X y función característica φ_X. Entonces si $g : \mathbb{R} \to \mathbb{R}$ es una función en $L^1(\mathbb{R})$,*

$$\int_{-\infty}^{\infty} \widehat{g}(x)\, dF_X(x) = \int_{-\infty}^{\infty} \varphi_X(y)\, g(y)\, dy$$

donde $\widehat{g} = \mathcal{F}(g)$ es la transformada de Fourier de g.

Prueba:

$$\begin{aligned}\int_{-\infty}^{\infty} \widehat{g}(x)\, dF_X(x) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(y)e^{ixy}\, dy\right]\, dF_X(x)\\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(y)e^{ixy}\, dF_X(x)\right]\, dy\\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{ixy}\, dF_X(x)\right] g(y)\, dy = \int_{-\infty}^{\infty} \varphi_X(y)\, g(y)\, dy\end{aligned}$$

El cambio en el orden de integración se puede justificar pues

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g(y)e^{ixy}|\, dF_X(x)\, dy &= \left(\int_{-\infty}^{\infty} |g(y)|\, dy\right) \cdot \left(\int_{-\infty}^{\infty} dF_X(x)\right)\\ &= \int_{-\infty}^{\infty} |g(y)|\, dy < \infty\end{aligned}$$

□

10.5. La Fórmula de Inversión: unicidad de la función característica

Un hecho fundamental es que es posible reconstruir la distribución de probabilidades de una variable aleatoria, a partir de su función característica.

Teorema 10.5.1 (Fórmula de inversión de Feller) *Sea X una variable aleatoria con $E(|X|) < +\infty$, función de distribución F_X y función característica φ_X. Entonces*

$$F_X(x_0) = \frac{1}{2\pi} \lim_{\sigma\to 0} \int_{-\infty}^{x_0} \left[\int_{-\infty}^{\infty} \varphi_X(y)\, e^{-iz\cdot y} \cdot e^{-(\sigma y)^2/2}\, dy\right]\, dz$$

en cada punto de continuidad x_0 de F_X.

Prueba: Usamos la identidad de Plancherel con la elección

$$g(y) = \frac{1}{2\pi} e^{-iz\cdot y} \cdot e^{-(\sigma_n y)^2/2}$$

donde (σ_n) es una sucesión tal que $\sigma_n \to 0$ y $z \in \mathbb{R}$. Entonces

$$\widehat{g}(x) = \frac{1}{\sigma_n\sqrt{2\pi}}\, e^{-(x-z)^2/(2\sigma_n^2)}$$

usando las propieddes que vimos antes. Queda:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma_n\sqrt{2\pi}}\, e^{-(x-z)^2/(2\sigma^2)}\, dF_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_X(y)\, e^{-iz\cdot y} \cdot e^{-(\sigma_n y)^2/2}\, dy$$

Por el corolorario 4.5.3, la primera integral es la densidad de probabilidad de $X_n = X + Y_n$ donde $Y_n \sim N(0, \sigma_n^2)$ es independiente de X. Integrando

$$F_{X_n}(x_0) = \frac{1}{2\pi}\int_{-\infty}^{x_0} \left[\int_{-\infty}^{\infty} \varphi_X(y)\, e^{-ix_0\cdot y} \cdot e^{-(\sigma_n y)^2/2}\, dy\right] dz$$

Pero nos acordamos que $X_n = X + Y_n$ donde $Y_n \sim N(0, \sigma_n^2)$ es independiente de X.

Cuando $\sigma_n \to 0$, $Y_n \xrightarrow{P} 0$, luego $X_n \xrightarrow{P} X$,y por lo tanto $X_n \xrightarrow{D} X$. Entonces

$$F_{X_n}(t) \to F_X(t)$$

en cada punto de continuidad de F_X. Esto prueba el teorema. □

Una variante de este teorema es (Véase [Jam02], capítulo 6):

Teorema 10.5.2 (Otra versión de la Fórmula de inversión) *Si X es una variable aleatoria, con función de distribución $F = F_X$ y función característica $\varphi = \varphi_X$, y x e y son puntos de continuidad de F $x < y$ entonces*

$$F(y) - F(x) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-itx} - e^{ity}}{it}\, \varphi(t)\, dt$$

Corolario 10.5.3 (Unicidad de la función característica) *Si F_1 y F_2 son dos distribuciones de probabilidad, y $\varphi_{F_1}(t) = \varphi_{F_2}(t)$ para todo $t \in \mathbb{R}$ (es decir: sus funciones características coinciden) entonces $F_1 = F_2$.*

Prueba: La fórmula de inversión implica que $F_1(x) = F_2(x)$ si x es un punto de continuidad. Si x no lo fuera, basta observar que como los puntos de discontinuidad de F_1 y F_2 son a lo sumo numerables, entonces podemos elegir una sucesión (x_n) tal que $x_n \searrow x$, tal que x_n sea un punto de continuidad tanto de F_1 como de F_2, entonces $F_1(x_n) = F_2(x_n)$ y como F_1 y F_2 son continuas por la derecha, deducimos que $F_1(x) = F_2(x)$. □

10.5.1. Otra versión de la fórmula de inversión

En general, $\varphi_X \notin L^1(\mathbb{R})$, como muestra el ejemplo de la distribución uniforme [o también si X fuera una variable discreta no nula]. Pero si esto ocurriera, podríamos pasar al límite cuando $\sigma \to 0$ en la integral (usando el teorema de convergencia mayorada), y obtener una fórmula más sencilla

$$F_X(x_0) = \frac{1}{2\pi} \int_{-\infty}^{x_0} \int_{-\infty}^{\infty} \varphi_X(y)\, e^{-ix_0 \cdot y}\, dy$$

[Si φ_X no estuviera en L^1, esto no tendría sentido pues la integral podría diverger como se ve tomando $x_0 = 0$]

Se deduce que entonces, X es una variable continua con la densidad:

$$f_X(x_0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi_X(y)\, e^{-ix_0 \cdot y}\, dy$$

En términos de la transformada de Fourier, esto se formularía así:

Teorema 10.5.4 (Fórmula clásica de inversión de Fourier) *Sea $f \in L^1(\mathbb{R})$ continua tal que $\widehat{f} \in L^1(\mathbb{R})$, entonces podemos reconstruir f a partir de su transformada mediante la fórmula de inversión*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(y)\, e^{-ix \cdot y}\, dy$$

Comparemos esto con la definición de la transformada:

$$\widehat{f}(y) = \int_{-\infty}^{\infty} f(x)\, e^{ix \cdot y}\, dx$$

Esta forma de la fórmula de inversión de Fourier es más simétrica, pero no es cierta sin la restricción de que $\widehat{f} \in L^1(\mathbb{R})$.

Un ejemplo: la distribución de Laplace

Una variable aleatoria X tiene la **distribución de Laplace** o **distribución exponencial doble** con parámetros $\mu \in \mathbb{R}$ y $b > 0$ si tiene la densidad de probabilidad

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b} \right)$$

Calculemos su función caracerística. Usando las propiedades, basta saber hacerlo con $\mu = 0$ y $b = 1$. En ese caso,

$$\begin{aligned}\varphi_X(t) = \widehat{f}(t) &= \int_{-\infty}^{\infty} e^{ixt}\frac{1}{2}e^{-|x|}\,dx \\ &= \int_{-\infty}^{0} e^{ixt}\frac{1}{2}e^{-|x|}\,dx + \int_{0}^{\infty} e^{ixt}\frac{1}{2}e^{-|x|}\,dx \\ &= \int_{-\infty}^{0} e^{ixt}\frac{1}{2}e^{x}\,dx + \int_{0}^{\infty} e^{ixt}\frac{1}{2}e^{-x}\,dx \\ &= \frac{1}{2}\int_{-\infty}^{0} e^{x(it+1)}\,dx + \frac{1}{2}\int_{0}^{\infty} e^{x(it-1)}\,dx\end{aligned}$$

Siguiendo, las integrales que nos quedaron se pueden calcular con la definición de integral impropia y la regla de Barrow. Nos queda:

$$\begin{aligned}\varphi_X(t) = \widehat{f}(t) &= \frac{1}{2}\left[\frac{1}{it+1} - \frac{1}{it-1}\right] \\ &= \frac{1}{1+t^2}\end{aligned}$$

En general, si μ y b son cualesquiera, la función característica de una variable aleatoria con distribución de Laplace va a ser

$$\varphi_X(t) = \frac{e^{it\mu}}{1+b^2t}$$

Otro ejemplo: La distribución de Cauchy

La **distribución de Cauchy** $\mathcal{C}(\mu, \lambda)$ tiene densidad de probabilidad dada por:

$$f(x) = \frac{1}{\pi}\,\frac{\lambda}{\lambda^2 + (x-\mu)^2}$$

De vuelta, nos bastaría calcular su función característica con $\mu = 0$ y $\lambda = 1$. Sería

$$\varphi_X(t) = \int_{-\infty}^{\infty} \frac{1}{\pi}\,\frac{1}{x^2+1}\,e^{ixt}\,dx$$

No es fácil calcular esta integral directamente. Pero si observamos que la densidad $f(x) = \frac{1}{1+x^2}$ es el resultado del ejemplo anterior, podemos calcularla usando la fórmula de inversión de Fourier (como f es par, no cambia la integral si reemplazamos x por $-x$). Obtenemos

$$\varphi_X(t) = e^{-|t|}$$

En general, si μ y λ son cualesquiera,

$$\varphi_X(t) = e^{i\mu t - \lambda|t|}$$

10.6. Transformada de Fourier de una derivada

Proposición 10.6.1 *Si $f : \mathbb{R} \to \mathbb{C}$ es una función en L^1 que es de clase C^1 y $f(x) \to 0$ cuando $|x| \to +\infty$,*

$$\mathcal{F}(f')(t) = (-it)\mathcal{F}f(t)$$

Prueba:

$$\begin{aligned}
\mathcal{F}(f')(t) &= \int_{-\infty}^{\infty} f'(x)e^{ixt}dx \\
&= \lim_{R\to+\infty} \int_{-R}^{R} f'(x)e^{ixt}dx \\
&= \lim_{R\to+\infty} \left\{ f(x)e^{ixt}\Big|_{-R}^{R} - \int_{-R}^{R} f(x)ite^{ixt}dx \right\} \\
&= -\int_{-\infty}^{\infty} f(x)ite^{ixt}dx \\
&= (-it)\mathcal{F}f(t)
\end{aligned}$$

□

10.7. Derivada de la transformada de Fourier

Proposición 10.7.1 *Si $f : \mathbb{R} \to \mathbb{C}$ es una función en L^1 tal que $x \cdot f(x) \in L^1$ entonces $\widehat{f}$ es derivable y*

$$\frac{d}{dt}\mathcal{F}f(t) = \mathcal{F}(ixf)(t)$$

Prueba:

$$\begin{aligned}
\frac{d}{dt}\mathcal{F}f(t) &= \int_{-\infty}^{\infty} f(x)\, \frac{d}{dt}[e^{ixt}]\, dx \\
&= \int_{-\infty}^{\infty} f(x)ixe^{ixt}\, dx \\
&= \mathcal{F}(ixf)(t)
\end{aligned}$$

Para justificar la derivación bajo el signo de integral, se usa un teorema de análisis real (corolario del teorema de convergencia mayorada). □

10.8. El espacio de Schwartz

Definimos el **espacio de Schwartz** $\mathcal{S}(\mathbb{R})$ como el conjunto de funciones $f : \mathbb{R} \to \mathbb{C}$ de clase C^∞ tales que para todo par de índices j y k en $\mathbb{N}_0$ existe una constante $M_{j,k}$ tal que

$$|x^j f^{(k)}(x)| \leq M_{j,k} \text{ para todo } x \in \mathbb{R}$$

La idea es que si una función está en $\mathcal{S}(\mathbb{R})$, ella y todas sus derivadas decaen en infinito más rápido que x^{-k} para todo k. Es un espacio muy chico, pero las funciones C^∞ de soporte compacto están en él, así como las funciones gaussianas

$$f(x) = e^{-ax^2} \quad \text{con } a > 0$$

En particular, si $f \in \mathcal{S}(\mathbb{R})$, $f^{(k)}$ y $x^k f$ estarán en $\mathcal{S}(\mathbb{R})$ para todo $k \in \mathbb{N}$.

Teorema 10.8.1 *Sea $\mathcal{S} = \mathcal{S}(\mathbb{R})$ el espacio de Schwartz. La transformada de Fourier pensada como una transformación lineal $\mathcal{F} : \mathcal{S} \to \mathcal{S}$ es biyectiva. Su inversa viene dada por la fórmula de inversión clásica*

$$\mathcal{F}^{-1}(g)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(y)\, e^{-ix\cdot y} f(y)\, dy$$

Para todo índice k tenemos:

$$\mathcal{F}(f^{(k)})(t) = (-it)^k \mathcal{F}f(t)$$

$$\frac{d^k}{dt^k}\mathcal{F}f(t) = \mathcal{F}((ix)^k f)(t)$$

10.9. El Teorema de Continuidad de Paul Lévy

Teorema 10.9.1 *Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias finitas en casi todo punto, y X otra variable aleatoria finita en casi todo punto. Entonces*

$$X_n \xrightarrow{D} X \Leftrightarrow \varphi_{X_n}(t) \to \varphi(t) \; \forall \, t \in \mathbb{R}$$

Nota: En realidad vamos a ver que si

$$\varphi_{X_n}(t) \to \varphi(t) \text{ para casi todo } t$$

entonces

$$X_n \xrightarrow{D} X$$

Prueba: Supongamos primero que $X_n \xrightarrow{D} X$. Para ver que $\varphi_{X_n}(t) \to \varphi(t)$ basta aplicar el corolario 9.2.3 aplicado a la función $\varphi(t) = e^{itx}$ (Este corolario se extiende a funciones con valores complejos, separando la parte real y la imaginaria). Ahora queremos probar el recíproco. Supongamos que $\varphi_{X_n}(t) \to \varphi(t)$ para todo t. Queremos probar que $X_n \xrightarrow{D} X$. Usando el recíproco fuerte del teorema de Helly-Bray, esto es equivalente a probar que

$$E[\psi(X_n)] \to E[\psi(X)]$$

para toda ψ de clase C^∞ con soporte compacto. Como obervamos antes, ψ está en el espacio de Schwartz, así que podemos escribir $\psi = \mathcal{F}(g)$ donde

$$g(x) = \mathcal{F}^{-1}(\psi)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi(y)\, e^{-ix\cdot y}\, dy$$

será otra función en el espacio de Schwartz.

Entonces escibimos

$$\begin{aligned} E[\psi(X_n)] &= \int_{-\infty}^{\infty} \psi(x)\, dF_{X_n}(x) & &= \int_{-\infty}^{\infty} \widehat{g}(x)\, dF_{X_n}(x) \\ &= \int_{-\infty}^{\infty} g(x)\, \varphi_{X_n}(x)\, dx \end{aligned}$$

por la indentidad de Plancherel. Como

$$|g(x)\, \varphi_{X_n}(x)| \leq |g(x)|$$

y g está en L^1, podemos pasar al límite cuando $n \to +\infty$ usando el teorema de convergencia mayorada. Y se obtiene:

$$E[\psi(X_n)] \to \int_{-\infty}^{\infty} g(x)\, \varphi_X(x)\, dx = E[\psi(X)]$$

haciendo la misma cuenta que antes, con X_n en lugar de X. Como vale para toda ψ con soporte compacto, deducimos que

$$X_n \xrightarrow{D} X$$

□

Nota: Este teorema será la clave de la demostración del Teorema del Límite Central en el capítulo siguiente. La demostración que dimos está adaptada de [HN01, Teorema 11.50] (si bien en este libro está escrita asumiendo que las distribuciones F_n son absolutamente continuas). En el apéndice H se da una prueba alternativa (con un enunciado más general) tomada de [Jam02].

10.9.1. Un ejemplo

Supongaamos que las X_n son variables de Rademacher con probabilidad de éxito 1/2, o sea

$$P\{X_n = -1\} = P\{X_n = 1\} = 1/2$$

y son independientes. Vamos a probar que

$$Y_n = \sum_{k=1}^{n} \frac{X_k}{2^k} \xrightarrow{D} \mathcal{U}(-1, 1) \text{ cuando } n \to +\infty$$

usando el teorema de Lévy.

Para ello vamos a calcular la función característica de Y_n Notamos que

$$\varphi_{X_n}(t) = E[e^{itX_n}] = \frac{1}{2} \cdot e^{it} + \frac{1}{2} e^{-it} = \cos(t)$$

Como las X_n son independientes:

$$\varphi_{Y_n}(t) = \prod_{k=1}^{n} \varphi_{X_k}\left(\frac{t}{2^k}\right) = \prod_{k=1}^{n} \cos\left(\frac{t}{2^k}\right)$$

¿Cómo calcular este producto ? La identidad trigonométrica

$$\operatorname{sen} t = 2 \operatorname{sen}\left(\frac{t}{2}\right) \cdot \cos\left(\frac{t}{2}\right)$$

permite probar por inducción que

$$\operatorname{sen} t = 2^n \operatorname{sen}\left(\frac{t}{2^n}\right) \left[\prod_{k=1}^{n} \cos\left(\frac{t}{2^k}\right)\right]$$

[Estas fórmulas las saqué del artículo de Wikipedia sobre la fórmula de Viète para π].

Entonces despejando vemos que si $t \notin \pi\mathbb{Z}$,

$$\varphi_{Y_n}(t) = \frac{\sin(t)}{2^n \operatorname{sen}\left(\frac{t}{2^n}\right)} = \frac{\sin(t)}{t} \cdot \frac{\frac{t}{2^n}}{\operatorname{sen}\left(\frac{t}{2^n}\right)}$$

Como

$$\lim_{x \to 0} \frac{\operatorname{sen}\ x}{x} = 1$$

vemos que

$$\varphi_{Y_n}(t) \to \frac{\operatorname{sen} t}{t} \quad \forall\, t \notin \pi\mathbb{Z}$$

Deducimos que

$$\varphi_{Y_n}(t) \to \varphi_Y(t) \text{ para todo } t \notin \pi\mathbb{Z}$$

donde $Y \sim \mathcal{U}(-1, 1)$. Por el teorema de continuidad de Paul Lévy,

$$Y_n \xrightarrow{D} Y$$

Capítulo 11

El Teorema del Límite Central

En este capítulo, presentaremos el Teorema del Límite Central, que es uno de los resultados fundamentales de la teoría de probabilidades. Informalmente, este teorema dice que la suma de un número grande de variables aleatorias independientes con varianza finita , donde la varianza de cada variable contribuye poco (en algún sentido) a la varianza total se distribuye en forma aproximadamente normal (formalizaremos esta idea más adelante). Este teorema justifica el papel central que juega la distribución normal en la estadística. Por ejemplo, los errores de medición en un experimento suelen tener una distribución normal, y esto es esperable por el teorema central del límite, si suponemos que el error de medición puede originarse en distintas fuentes independientes de error, cada una de las cuales contribuye en pequeña medida al error total.

Comenzaremos presentando una versión para la distribución binomial, conocida como el **teorema de De Moivre-Laplace**. Es históricamente la primera versión que se conoció del teorema del límite central. Y la demostraremos "a mano" utilizando la aproximación del factorial por medio de la fórmula de Stirling. Después demostraremos una versión del teorema del límite central para variables independientes y uniformemente distribuidas (con varianza finita), por medio de la técnica de las funciones características que desarrollamos en el capítulo anterior. Finalmente, haremos algunos comentarios sobre sus generalizaciones y versiones más refinadas.

11.1. El Teorema Local de De Moivre-Laplace

Sea X una variable aleatoria con segundo momento finito. Entonces la variable reescaldada (o "normalizada")

$$X^* = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

satisface que $E(X^*) = 0$ y $\text{Var}(X^*) = 1$.

Sea S_n el número de éxitos en n ensayos de Bernoulli con probabilidad $p \in (0,1)$. Sabemos que S_n tiene distribución binomial:

$$P\{S_n = k\} = b(k,n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \quad (0 \leq k \leq n), \quad q = 1-p$$

y que $E[S_n] = np$, $\text{Var}(S_n) = npq$. Consideramos entonces la variable normalizada:

$$S_n^* = \frac{S_n - np}{\sqrt{npq}} \tag{11.1}$$

Nuestro objetivo es estudiar el límite de la distribución de S_n^* cuando $n \to +\infty$:

Comenzamos aproximando la distribución binomial, utilizando la fórmula de Stirling (ver apéndice):

$$n! \sim \sqrt{2\pi}\ n^{n+1/2} e^{-n} e^{O(1/n)}$$

Obtenemos[1]:

Teorema 11.1.1 (Teorema local de De Moivre-Laplace)

$$b(k,n,p) = \frac{1}{\sqrt{2\pi npq}}\ e^{-x_k^2/2}\ (1+\beta_{n,k})$$

donde

$$x_k = \frac{k-np}{\sqrt{npq}}$$

y para $M \geq 0$,

$$\max_{|x_k| \leq M} |\beta_{n,k}| \to 0 \text{ cuando } n \to \infty \tag{11.2}$$

Prueba:

$$b(k,n,p) = \frac{\sqrt{2\pi}\ n^{n+1/2}\ e^{-n}\ e^{O(1/n)}}{\sqrt{2\pi}\ k^{k+1/2}\ e^{-k}\ e^{O(1/k)} \sqrt{2\pi}\ (n-k)^{n-k+1/2}\ e^{-(n-k)} e^{O(1/(n-k))}}\ p^k q^{n-k}$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} e^{O(1/n)+O(1/k)+O(1/(n-k))}$$

Notemos que:

$$k = np + x_k\sqrt{npq} = np\left(1 + x_k\sqrt{\frac{q}{np}}\right)$$

[1] La prueba que presentamos del teorema de De Moivre-Laplace está basada en unas notas del curso de probabilidad y estadística del profesor N. Fava.

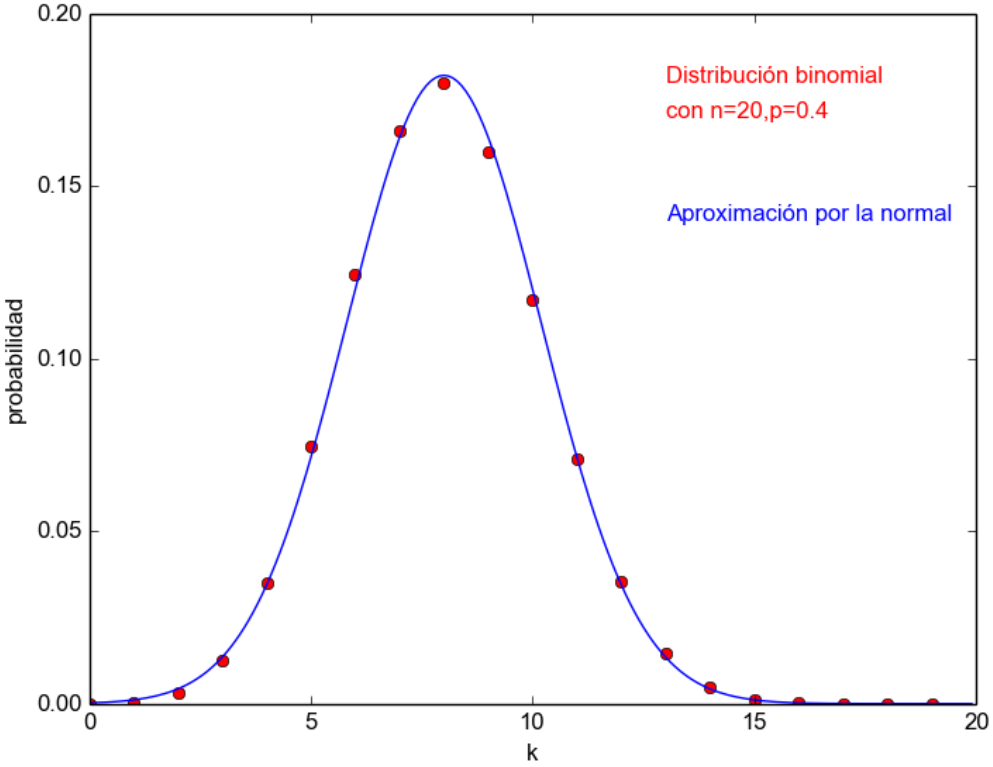


Figura 11.1: Ilustración de la bondad de la aproximación a la distribución binomial por la distribución normal dada por el teorema de local de De Moivre-Laplace, con $n = 20$ y $p = 0{,}4$.

y que:

$$n - k = nq - x_k\sqrt{npq} = nq\left(1 - x_k\sqrt{\frac{p}{nq}}\right)$$

Estimaremos en forma separada el valor de cada uno de los factores a medida que $n \to +\infty$:

$$\sqrt{\frac{n}{k(n-k)}} = \sqrt{\frac{n}{np\left(1 + x_k\sqrt{\frac{q}{np}}\right) nq\left(1 - x_k\sqrt{\frac{p}{nq}}\right)}} = \frac{1}{\sqrt{npq}}(1 + \alpha_{n,k})$$

donde

$$\max_{|x_k| \leq M} |\alpha_{n,k}| \to 0 \text{ cuando } n \to +\infty$$

Para estimar el segundo factor, tomamos logaritmo y hacemos uso del desarrollo de Taylor: $\log(1+t) = t - \frac{t^2}{2} + O(t^3)$ cuando $t \to 0$.

En consecuencia:

$$\begin{aligned}
\log\left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} &= (-k)\log\left(\frac{k}{np}\right) - (n-k)\log\left(\frac{n-k}{nq}\right) \\
&= (-k)\log\left(1 + x_k\sqrt{\frac{q}{np}}\right) - (n-k)\log\left(1 - x_k\sqrt{\frac{p}{nq}}\right) \\
&= (-np - x_k\sqrt{npq})\left\{x_k\sqrt{\frac{q}{np}} - \frac{1}{2}x_k^2\frac{q}{np} + O\left(\frac{1}{n^{3/2}}\right)\right\} \\
&+ (-nq + x_k\sqrt{npq})\left\{-x_k\sqrt{\frac{p}{nq}} - \frac{1}{2}x_k^2\frac{p}{nq} + O\left(\frac{1}{n^{3/2}}\right)\right\} \\
= -x_k\sqrt{npq} + \frac{1}{2}qx_k^2 - qx_k^2 + O\left(\frac{1}{n^{1/2}}\right) &+ x_k\sqrt{npq} + \frac{1}{2}px_k^2 - px_k^2 + O\left(\frac{1}{n^{1/2}}\right) \\
&= -\frac{1}{2}x_k^2 + O\left(\frac{1}{n^{1/2}}\right)
\end{aligned}$$

Deducimos que:

$$\left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} = e^{-x_k^2/2} \cdot e^{O(1/n^{1/2})}$$

Finalmente consideramos el término de error $e^{O(1/n)-O(1/k)-O(1/(n-k))} = e^E$ donde

$$E = O\left(\frac{1}{n}\right) + O\left(\frac{1}{np\left(1 + x_k\sqrt{\frac{q}{np}}\right)}\right) + O\left(\frac{1}{nq\left(1 - x_k\sqrt{\frac{p}{nq}}\right)}\right) = O\left(\frac{1}{n}\right)$$

En consecuencia, utilizando las estimaciones que hemos obtenido para cada factor, y teniendo en cuenta que $O(1/n^{1/2}) + O(1/n) = O(1/n^{1/2})$, obtenemos que:

$$b(k,n,p) = \frac{1}{\sqrt{2\pi npq}}\, e^{-x_k^2/2} \cdot (1 + \alpha_n(x_k))\, e^{O(1/n^2)}$$

Finalmente, observamos que el factor de error dado por

$$(1 + \alpha_n(x_k))e^{O(1/n^{1/2})}$$

tiende a 1 cuando $n \to +\infty$, uniformemente para los k tales que $|x_k| \leq M$, por lo que podremos representarlo en la forma $1 + \beta_{n,k}$ donde

$$\max_{|x_k| \leq M} |\beta_{n,k}| \to 0$$

□

Observación 11.1.2 *La fórmula 11.2 significa que la aproximación dada por el teorema de local De Moivre-Laplace es buena en el centro de la distribución binomial, pero no en las colas de la misma. Por ejemplo, si n es grande y p es muy pequeño, como se ilustra en la figura 11.2. En esta situación es mejor la aproximación por la distribución de Poisson que discutimos en la sección 3.6. Por simetría, tampoco es buena si p está muy cerca de* 1.

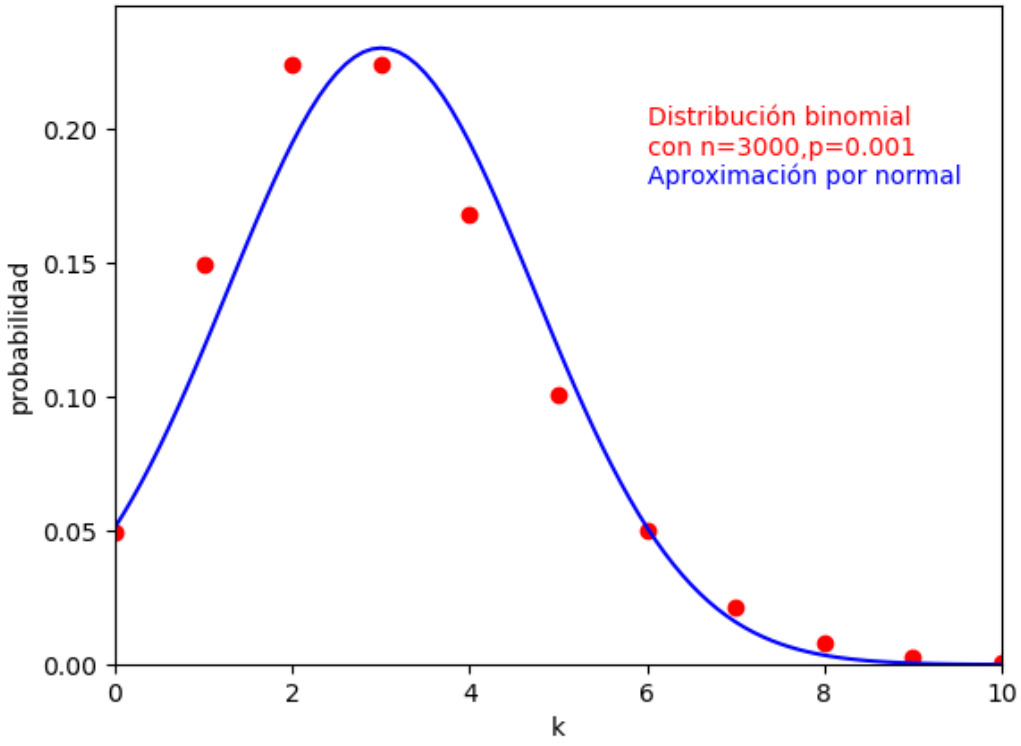


Figura 11.2: Ilustración de la bondad de la aproximación a la distribución binomial por la distribución normal dada por el teorema local de De Moivre-Laplace, con $n = 3000$ y $p = 0{,}01$. Vemos que no resulta tan buena si n es grande y p es pequeña.

11.2. El Teorema de De Moivre-Laplace

En este capítulo, notaremos por

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{11.3}$$

a la densidad normal estándar (que introdujimos en el ejemplo 4.1.6), y por

$$G(x) = \int_{-\infty}^{x} g(t)\, dt \tag{11.4}$$

a la correspondiente función de distribución normal (acumulada).

El siguiente teorema afirma que la distribución límite de la variable normalizada S_n^* está dada por la integral definida de $g(x)$:

Teorema 11.2.1 (De Moivre-Laplace)

$$P\{a < S_n^* \leq b\} \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx = G(b) - G(a)$$

uniformemente en a y en b cuando $n \to +\infty$.

Observación 11.2.2 *De acuerdo con [McD05a], el teorema 11.1.1 fue enunciado por De Moivre en 1754 en su trabajo* Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi*, pero sólo lo demostró para $p = 1/2$. La primera prueba completa fue dada por Laplace (1795) en su libro* Théorie analytiquedes probabilités. *Análogamente el teorema 11.2.1 fue demostrado por De Moivre para $p = 1/2$, y por Laplace para cualquier $p \in (0, 1)$.*

.

La idea básica de la demostración es la siguiente:

$$P_n(a, b) = P\{a < S_n^* \leq b\} = \sum_{a < x_k \leq b} b(k, n, p)$$

ya que si S_n^* toma el valor x_k, entonces S_n toma el valor k.

Los puntos x_k están cada vez más próximos a medida que $n \to +\infty$, ya que

$$x_{k+1} - x_k = \frac{1}{\sqrt{npq}}$$

y por el teorema anterior $b(k, n, p) \approx g(x_k)(x_{k+1} - x_k)$ entonces,

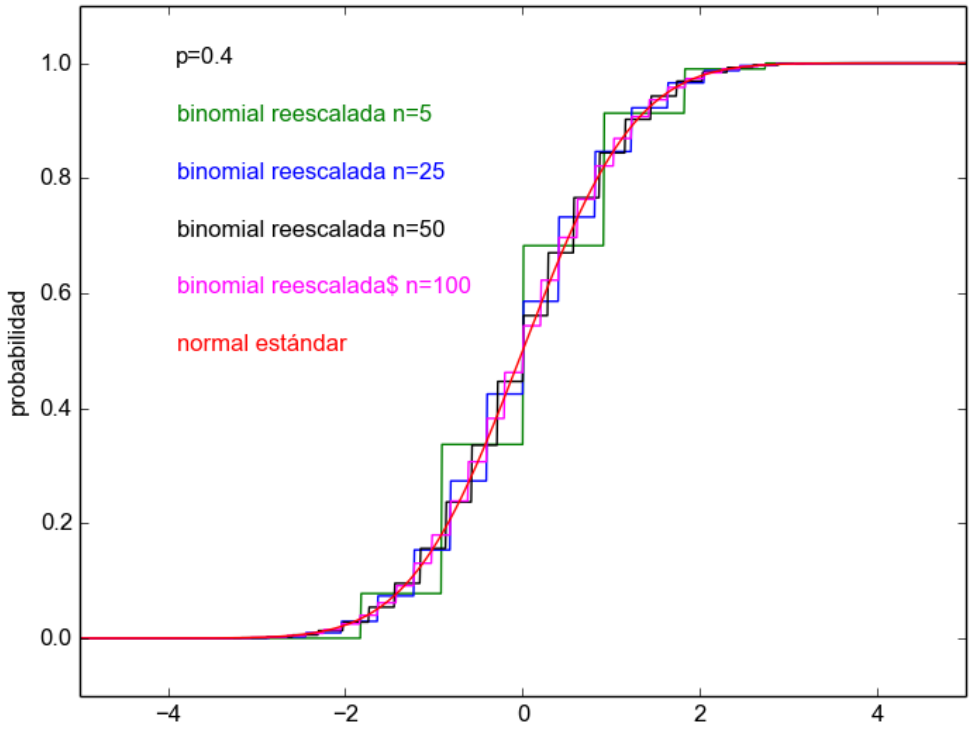


Figura 11.3: Ilustración del teorema de De Moivre-Laplace: para $p = 0{,}4$ y distintos valores de n, dibujamos la función de distribución de la distribución binomial, junto con la de la normal estándar.

$$P_n(a,b) = P\{a < S_n^* \leq b\} \approx \sum_{a<x_k\leq b} g(x_k)(x_{k+1} - x_k)$$

y esta es una suma de Riemann para la integral $\int_a^b g(x)\ dx$. Por lo tanto, conforme $n \to +\infty$, es razonable que podamos aproximar $P_n(a,b)$ por dicha integral.

La demostración consiste en una formalización de esta idea:

Prueba: Dado $\varepsilon > 0$, elegimos M de modo que

$$G(M) - G(-M) = 1 - \varepsilon$$

y además

$$\frac{1}{M^2} < \varepsilon$$

por consiguiente:

$$G(-M) = 1 - G(M) = \varepsilon/2$$

Consideramos primero el caso en que el intervalo (a, b) está contenido en el intervalo $(-M, M)$. La función g_n definida por $g_n(x) = g(x_k)$ para $x_k < x \leq x_{k+1}$ converge uniformente a $g(x)$ cuando $n \to +\infty$, en virtud de la continuidad uniforme de g.

Denotamos por k_0 el mínimo entero tal que $a < x_{k_0}$ y sea k_1 el máximo entero tal que $x_{k_1} \leq b$.

En virtud del teorema 11.1.1,

$$P_n(a,b) = \sum_{a<x_k\leq b} (1+\beta_{n,k}) g(x_k)(x_{k+1} - x_k)$$

$$= \sum_{a<x_k\leq b} g(x_k)(x_{k+1} - x_k) + \sum_{a<x_k\leq b} \beta_{n,k}\, g(x_k)(x_{k+1} - x_k)$$

$$= \int_{x_{k_0}}^{x_{k_1+1}} g_n(x)\, dx + \sum_{a<x_k\leq b} \beta_{n,k}\, g(x_k)(x_{k+1} - x_k)$$

En consecuencia,

$$P_n(a,b) = \int_a^b g_n(x)\, dx - \int_a^{x_{k_0}} g_n(x)\, dx + \int_b^{x_{k_1+1}} g_n(x)\, dx$$

$$\sum_{a<x_k\leq b} \beta_{n,k}\, g(x_k)(x_{k+1} - x_k)$$

o sumando y restando g_n:

$$P_n(a,b) = \int_a^b g(x)\, dx + \int_a^b [g_n(x) - g(x)]\, dx + \int_a^{x_{k_0}} g_n(x)\, dx$$

$$+ \int_b^{x_{k_1+1}} g_n(x)\, dx + \sum_{a<x_k\leq b} \beta_{n,k}\, g(x_k)(x_{k+1} - x_k)$$

El segundo término de esta expresión podemos acotarlo del siguiente modo:

$$\left| \int_a^b [g_n(x) - g(x)]\, dx \right| \leq (b-a) \sup_{x\in[a,b]} |g_n(x) - g(x)| \leq 2M \sup_{x\in[a,b]} |g_n(x) - g(x)|$$

Además como g y por consiguiente g_n están acotadas por $(2\pi)^{-1/2}$, deducimos que:

$$\left| \int_a^{x_{k_0}} g_n(x)\, dx \right| \leq \frac{1}{\sqrt{2\pi npq}}$$

, Similarmente:

$$\left|\int_b^{x_{k_1+1}} g_n(x)\ dx\right| \leq \frac{1}{\sqrt{2\pi npq}}$$

Finalmente, último término podemos acotarlo del siguiente modo,

$$\left|\sum_{a<x_k\leq b} \beta_{n,k}\ g(x_k)(x_{k+1}-x_k)\right| \leq \max_{|x_k|\leq M} |\beta_{n,k}| \sum_{k=k_0}^{k_1} g(x_k)(x_{k+1}-x_k)$$

$$\leq \frac{1}{\sqrt{2\pi}} 2M \max_{|x_k|\leq M} |\beta_{n,k}| \to 0 \text{ cuando } n \to +\infty$$

Como todas las estimaciones efectuadas, son independientes de a y b, concluimos que cuando $n \to +\infty$,

$$P_n(a,b) \to \int_a^b g(x)\ dx$$

uniformemente en a y b. Es decir: existe un entero $n_0 = n_0(\varepsilon)$ independiente de a y de b tal que

$$\left|P_n(a,b) - \int_a^b g(x)\ dx\right| < \varepsilon$$

para cualquier $a, b \in (-M, M)$. En particular, deducimos que:

$$\left|P_n(-M,M) - \int_{-M}^{M} g(x)\ dx\right| \leq \varepsilon$$

para $n \geq n_0$.

Si (a,b) no está contenido en $(-M,M)$, tenemos que:

$$P_n(a,b) = P_n(a,-M) + P_n(-M,M) + P_n(M,b)$$

y

$$\int_a^b g(x)\ dx = \int_a^{-M} g(x)\ dx + \int_{-M}^{M} g(x)\ dx + \int_M^b g(x)\ dx$$

Utilizando entonces la desigualdad triangular tenemos que:

$$\left|P_n(a,b) - \int_a^b g(x)\ dx\right| \leq \left|P_n(-M,M) - \int_{-M}^{M} g(x)\ dx\right| +$$

$$+P_n(a,-M) + P_n(M,b) + \int_a^{-M} g(x)\ dx + \int_M^b g(x)\ dx$$

Pero

$$\int_a^{-M} g(x)\,dx + \int_M^b g(x)\,dx \leq \int_{-\infty}^{-M} g(x)\,dx + \int_M^{\infty} g(x)\,dx = G(-M) + [1 - G(M)] < \varepsilon$$

y

$$P_n(a, -M) + P_n(M, b) \leq P\{|S_n^*| \geq M\} \leq \frac{1}{M^2} < \varepsilon$$

por la desigualdad de Chebyshev, pues $E(S_n^*) = 0$ y $\mathrm{Var}(S_n^*) = 1$ (teniendo en cuenta nuestra elección de M al comienzo de la demostración). En consecuencia,

$$\left| P_n(a, b) - \int_a^b g(x)\,dx \right| \leq 3\varepsilon$$

si $n \geq n_0(\varepsilon)$ Esto concluye la demostración del teorema. □

11.3. Una Aplicación a la Estadística

Veremos ahora una aplicación del teorema de De Moivre-Laplace y de la distribución normal, a la estadística.

Consideremos por ejemplo, una encuesta electoral para una elección donde participan dos candidatos A y B, y supongamos que cada persona puede votar por uno de ellos (y para simplificar que no hay votos en blanco). Podemos modelizar esto utilizando la distribución binomial, para ello imaginemos un experimento aleatorio donde se elige una persona al azar y se le pregunta por quien vota. Y llamemos p a la probabilidad de que vote por A ("éxito") y $q = 1 - p$ a la probabilidad de que vote por B. Alternativamente, podemos pensar que tenemos una elección en la que participan varios candidatos y que nos interesa medir la intención de voto de un determinado candidato A. En este caso, consideramos el experimento aleatorio que consiste en elegir una persona al azar, preguntarle por quien vota, y hay dos resultados posibles que nos interesan: si vota por A (con probabilidad p) o si no vota por A con probabilidad q=1-p.

Nuestro objetivo es estimar la probabilidad desconocida p. Como resulta extraordinariamente costoso y complicado preguntarle a cada votante del padrón electoral por quién piensa votar, lo que suele hacerse es elegir una **muestra**, digamos formada por n personas. Entonces, conforme a la ley de los grandes números, si llamamos S_n a la cantidad de personas de la muestra que votan por el candidato A, podemos aproximar la probabilidad desconocida p por la frecuencia:

$$f_n = \frac{S_n}{n}$$

observada en la muestra (Estamos suponiendo que las elecciones de las distintas personas pueden considerarse independientes unas de otras, de modo que la elección de n

personas encuestadas, puede considerarse como realizar n ensayos de Bernoulli, y la distribución de S_n sea dada por la distribución binomial.)

Otro ejemplo análogo se da en el control de calidad en un proceso industrial. Por ejemplo, imaginemos que tenemos un lote de 10.000 lamparitas y queremos saber cuantas están falladas. Llamemos p a la probabilidad de que una lamparita elegida al azar funcione, y $q = 1 - p$ a la probabilidad de que esté fallada. Nuevamente, sería extraordinariamente costoso probar una por una las lamparitas, por lo que se hace es elegir una muestra, y aproximar p por la frecuencia f_n observada en la muestra.

Una pregunta fundamental es entonces: ¿Cómo elegir el tamaño de la muestra?. Para ello, elegimos un margen de error ε, y un nivel de confianza $1-\alpha$ donde ε y α son números pequeños, y nos proponemos elegir el tamaño de la muestra de modo que podamos asegurar que la probabilidad de que f_n diste de p como mucho en ε es por lo menos $1-\alpha$, o sea:

$$P\{|f_n - p| \leq \varepsilon\} \geq 1 - \alpha \tag{11.5}$$

Por ejemplo: supongamos que queremos que muestra encuesta (o control de calidad) se equivoque como mucho en un 2 % en el 95 % de las veces que realizamos la encuesta. Entonces, elegimos $\varepsilon = 0,02$ y $\alpha = 0,05$.

Elegimos entonces x_α de modo que:

$$G(-x_\alpha) = \frac{\alpha}{2}$$

donde G es la función de distribución normal estándar (dada por 11.4). Por la simetría de la curva normal,

$$G(x_\alpha) = 1 - \frac{\alpha}{2}$$

Llamando S_n^* a la variable normalizada dada por (11.1), por el teorema de De Moivre Laplace:

$$P\left\{-x_\alpha \leq S_n^* \leq x_\alpha\right\} \approx \frac{1}{\sqrt{2\pi}} \int_{-x_\alpha}^{x_\alpha} e^{-x^2/2}\, dx = G(x_\alpha) - G(-x_\alpha) = 1 - \alpha$$

si n es suficientemente grande. En consecuencia, recordando la definición de S_n^* y despejando:

$$P\left\{-x_\alpha \sqrt{npq} \leq S_n - np \leq x_\alpha \sqrt{npq}\right\} \approx 1 - \alpha$$

$$P\left\{np - x_\alpha \sqrt{npq} \leq S_n \leq np + \alpha \sqrt{npq}\right\} \approx 1 - \alpha$$

$$P\left\{p - x_\alpha \sqrt{\frac{pq}{n}} \leq \frac{S_n}{n} \leq p + x_\alpha \sqrt{\frac{pq}{n}}\right\} \approx 1 - \alpha$$

O sea:

$$P\left\{\left|\frac{S_n}{n}-p\right| \leq x_\alpha \sqrt{\frac{pq}{n}}\right\} \approx 1-\alpha$$

Esta relación dice que con probabilidad $1-\alpha$ podemos asegurar que p está en el intervalo:

$$I_\alpha = \left[\frac{S_n}{n} - x_\alpha\sqrt{\frac{pq}{n}}, \frac{S_n}{n} + x_\alpha \sqrt{\frac{pq}{n}}\right]$$

I_α se llama un **intervalo de confianza** (asintótico) para p de nivel de confianza $1-\alpha$. En realidad en esta forma, esta relación no resulta todavía muy útil ya que no conocemos p y entonces tampoco conocemos el ancho del intervalo I_α. Pero podemos observar que:

$$pq = p(1-p) \leq \frac{1}{4} \; \forall \; p \in [0,1]$$

En consecuencia, podemos asegurar que

$$I_\alpha \subset \left[\frac{S_n}{n} - x_\alpha\frac{1}{2\sqrt{n}}, \frac{S_n}{n} + x_\alpha \frac{1}{2\sqrt{n}}\right]$$

y que (si n es grande):

$$P\left\{\left|\frac{S_n}{n}-p\right| \leq x_\alpha \frac{1}{2\sqrt{n}}\right\} \geq 1-\alpha$$

En consecuencia, si queremos que valga la relación (11.5) debemos elegir n para que:

$$x_\alpha \frac{1}{2\sqrt{n}} \leq \varepsilon$$

o sea:

$$n \geq n_0 = \left(\frac{x_\alpha}{2\varepsilon}\right)^2$$

Esta relación nos dice cuál es el tamaño (mínimo) de la muestra que necesitamos para poder garantizar un determinado margen de error con un determinado nivel de confianza. Por ejemplo, si $\alpha = 0,05$ y $\varepsilon = 0{,}02$, obtenemos que: $x_\alpha = 1,96$ y $n \geq 2401$.

Observación: Notamos que cuando $\alpha \to 0$, $x_\alpha \to +\infty$ por lo que $n_0 \to +\infty$.

11.4. El Teorema del Límite Central

El siguiente teorema generaliza al de De Moivre-Laplace:

Teorema 11.4.1 (Teorema del Límite Central, versión sencilla) *Sea* $(X_k)_{k\in\mathbb{N}} : \Omega \to \mathbb{R}$ *una sucesión de variables aleatorias independientes e idénticamente distribuidas con*

$0 < \sigma^2 = Var(X_k) < +\infty$. *Sea* $\mu = E[X_k]$ *(como suponemos que las* X_k *tienen todas la misma distribución, tendrán todas la misma esperanza y varianza). Notemos:*

$$S_n = X_1 + X_2 + \ldots + X_n$$

$$S_n^* = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

Entonces

$$S_n^* \xrightarrow{D} N(0,1)$$

Observación 11.4.2 *Para comprender el significado de este teorema, observemos que si consideramos el esquema de ensayos de Bernoulli, y las* (X_k) *son las variables aleatorias de la sección 3.4 entonces* S_n *representa el número total de éxitos en* n *ensayos, y el teorema del límite central se reduce al teorema de De Moivre-Laplace.*

Observación 11.4.3 *El nombre del teorema se debe a que proporciona una buena aproximación en el centro de la distribución, pero no tan buena en las colas de la misma, como vimos en la observación 11.1.2 para el caso de la distribución binomial. En inglés se denomina* central limit theorem*, pero por esta observación resulta más correcto traducirlo por* teorema del límite central *que por* teorema central del límite*, como muchas veces se hace.*

Para la prueba necesitamos un lema elemental sobre números complejos (que el lector fácilmente puede demostrar usando la rama principal del logaritmo).

Lema 11.4.4 *Si* (c_n) *es una sucesión de números complejos tal que* $c_n \to c$*, entonces*

$$\left(1 + \frac{c_n}{n}\right)^n \to e^c$$

Pasaremos entonces a la demostración del teorema del límite central:
Prueba: Sin pérdida de generalidad, podemos suponer que $\mu = 0$, cambiando sino las X_k por las variables centradas

$$\widetilde{X}_k = X_k - \mu$$

Calculemos la función característica de S_n^*. Como las (X_k) son idependientes, y tienen todas la misma distribución será

$$\varphi_{S_n^*}(t) = \varphi\left(\frac{t}{\sigma\sqrt{n}}\right)^n$$

donde $\varphi(t) = \varphi_{X_k}(t)$ para todo k. Hagamos el desarrollo de Taylor de $\varphi(t)$ a segundo orden. Usando la proposición 10.2.6 (que relaciona los momentos de X_k con las derivadas de la función característica en $t = 0$), vemos que es

$$\begin{aligned}\varphi(X_k)(t) &= 1 + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2 + t^2 e_2(t) \\ &= 1 - \frac{\sigma^2}{2}t^2 + t^2 e_2(t) \\ &= 1 + \left[-\frac{\sigma^2}{2} + e_2(t)\right] t^2\end{aligned}$$

donde

$$\lim_{t\to+\infty} e_2(t) = 0 \tag{11.6}$$

por la propiedad que tiene el resto de Taylor. Entonces:

$$\begin{aligned}\varphi_{S_n^*}(t) &= \left\{1 + \left[-\frac{\sigma^2}{2} + e_2\left(\frac{t}{\sigma\sqrt{n}}\right)\right]\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right\}^n \\ &= \left\{1 + \left[-\frac{1}{2} + \frac{1}{\sigma^2}\, e_2\left(\frac{t}{\sigma\sqrt{n}}\right)\right]\frac{t^2}{n}\right\}^n\end{aligned}$$

Fijado un t, si llamamos

$$c_n = \left[-\frac{1}{2} + \frac{1}{\sigma^2}\, e_2\left(\frac{t}{\sigma\sqrt{n}}\right)\right] t^2$$

como

$$c_n \to c = -\frac{t^2}{2}$$

cuando $n \to \infty$, por (11.6), vemos aplicando el lema que

$$\varphi_{S_n^*}(t) \to e^c = e^{-t^2/2}$$

pero esta función es justamente la función característica de la distribución normal estándar $N(0, 1)$. Por el corolario el teorema de continuidad de Paul Lévy, se deduce que S_n^* converge en distribución a la normal estándar, como afirma el teorema. □

Una prueba alternativa del teorema del límite central sin utilizar funciones características se presenta en [Chi22]. Otra demostración interesante es la de [Tro59] (pero usa la noción de operadores en un espacio de Banach).

11.4.1. Aplicación a las distribuciones χ^2_n

Para dar un ejemplo del teorema del límite central, consideremos nuevamente las variables

$$Z_n = X_1^2 + X_2^2 + \ldots + X_n^2$$

donde las (X_k) son variables con distribución normal estándar independientes, que introdujimos en la sección 4.9.1. Entonces, por definición Z_n tiene distribución χ^2_n y sabemos que $E[Z_n] = n$ y $\text{Var}(Z_n) = 2n$. Por el teorema del límite central, para n grande, la distribución normal proporciona una buena aproximación de la distribución χ^2_n en el sentido que las variables normalizadas

$$Z_n^* = \frac{Z_n - n}{\sqrt{2n}}$$

convergen en distribución a una normal estándar. El siguiente gráfico compara las funciones de distribución de Z_n^* con la de la distribución normal, para n grande:

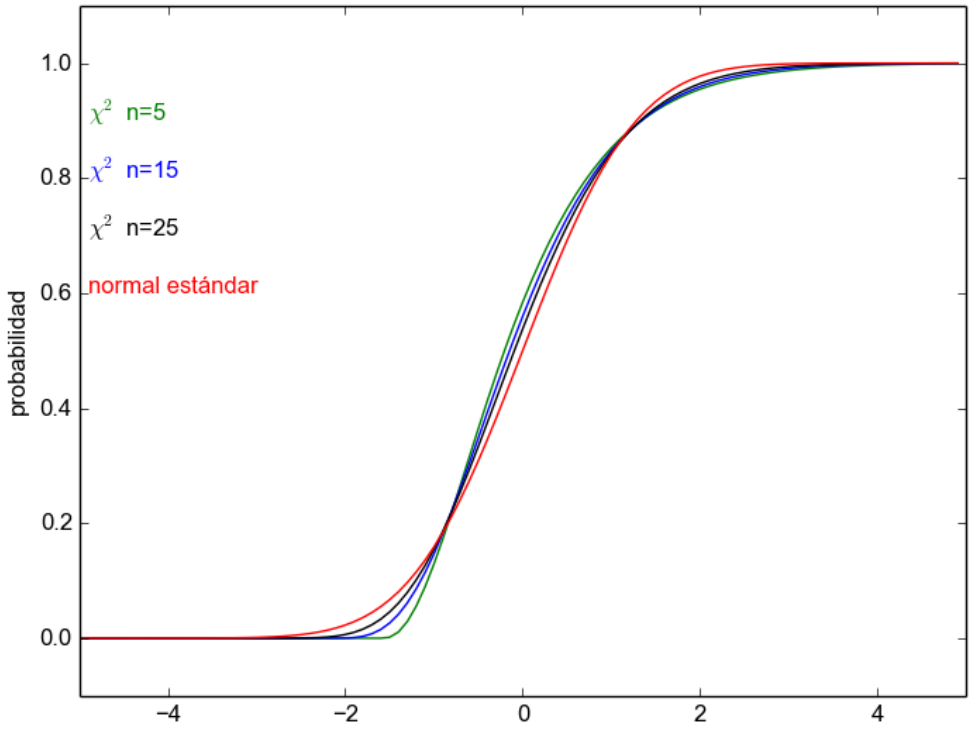


Figura 11.4: Convergencia en distribución de la distribución χ^2_n normalizada (distribución de Z_n^*) a la normal estándar.

11.5. Generalizaciones y comentarios adicionales

El teorema del límite central no está limitado al caso de variables idénticamente distribuidas. Como dijimos en la introducción, se aplica en general a sumas de variables aleatorias independientes con varianza finita, donde la varianza de cada variable contribuye (en algún sentido) a la varianza total. Una condición muy general para su validez está dada por el siguiente teorema de Lindeberg:

Teorema 11.5.1 (Teorema del Límite central de Lindeberg) *Sea $(X_k)_{k\in\mathbb{N}}$ una sucesión de variables aleatorias independientes tales que $\mu_k = E[X_k]$ y $\sigma_k^2 = Var(X_k)$, donde σ_k es finita y al menos algún $\sigma_{k_0} > 0$. Sean*

$$S_n = X_1 + X_2 + \ldots + X_k$$

$$s_n = \sqrt{Var(S_n)} = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$$

y supongamos que se cumple la siguiente condición de Lindeberg*:*

$$\forall\, \varepsilon > 0 \ \lim_{\varepsilon\to 0} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{|x-\mu_k|>\varepsilon s_n} (x-\mu_k)^2 \, dF_{X_k}(x) = 0$$

entonces si definimos

$$S_n^* = \frac{S_n - E[S_n]}{s_n} = \frac{S_n - (\mu_1 + \mu_2 + \ldots + \mu_n)}{s_n}$$

tenemos que

$$S_n^* \xrightarrow{D} N(0,1)$$

El teorema de Lindeberg implica el siguiente teorema de Lyapunov que da una condición más fuerte, pero quizás más fácil de entender:

Teorema 11.5.2 (Teorema Límite central de Lyapunov) *Sea $(X_k)_{k\in\mathbb{N}}$ una sucesión de variables aleatorias independientes tales que $\mu_k = E[X_k]$ y $\sigma_k^2 = Var(X_k)$, donde σ_k es finita y al menos algún $\sigma_{k_0} > 0$. Sean*

$$S_n = X_1 + X_2 + \ldots + X_k$$

$$s_n = \sqrt{Var(S_n)} = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$$

y supongamos que existe algún $\delta > 0$ tal que se cumple la siguiente condición de Lyapunov*:*

$$\forall\, \varepsilon > 0 \ \lim_{\varepsilon\to 0} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - \mu_k|^{2+\delta}] = 0$$

entonces si definimos

$$S_n^* = \frac{S_n - E[S_n]}{s_n} = \frac{S_n - (\mu_1 + \mu_2 + \ldots + \mu_n)}{s_n}$$

tenemos que

$$S_n^* \xrightarrow{D} N(0,1)$$

La demostración de estos resultados puede verse en [Jam02] (capítulo 7). También emplea el método de las funciones características, aunque resulta mucho más técnica.

Una pregunta que podemos hacernos es ¿cuál es la velocidad de convergencia a la distribución normal en el teorema del límite central? Una respuesta es dada por el teorema de Bery-Essen[2] cuya versión más sencilla (correspondiente a la situación del teorema 11.4.1) es la siguiente:

Teorema 11.5.3 (Teorema de Berry-Essen, versión sencilla) *Si (X_k) es una sucesión de variables independientes identicamente distribuidas, con $E(X_k) = \mu$, $E(X_k^2) = Var(X_k) = \sigma^2 > 0$ y si suponemos además que el tercer momento respecto de la media μ de las X_k*

$$\rho = E[|X_k - \mu|^3] < \infty$$

es finito, y si definimos como antes

$$S_n = X_1 + X_2 + \ldots + X_n$$

$$S_n^* = \frac{S_n - E[S_n]}{\sqrt{VarS_n}} = \frac{S_n - n\mu}{n\sqrt{\sigma}}$$

entonces

$$|F_{S^*}(x) - G(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

donde G denota la función de distribución de la normal estándar y C es una constante fija.

También debemos mencionar que el teorema del límite central se generaliza sin dificultades esenciales a vectores aleatorios, debiendo considerar en este caso para la distribución límite a la distribución normal multivariada (ver [Jam02], teorema 7.2). Y que existen versiones "locales" del teorema central del límite, que generalizan al teorema 11.1.1, para una discusión al respecto ver [McD05a].

[2]Este teorema fue descubierto independientemente por los matemáticos Andrew C. Berry [Ber41] y Carl-Gustav Esseen [CG42]. La prueba en el primero de ellos también emplea el método de las funciones características.

11.6. Una Aplicación a la Teoría de Números

Resulta sorprendente encontrar aplicaciones del teorema del límite central, en ramas de la matemática aparentemente alejadas de las probabilidades. En esta sección comentaremos brevemente una de ellas: una aplicación a la teoría de números. Esta rama de la matemática se ocupa fundamentalmente de las propiedades de los números enteros.

Comencemos con una pregunta muy básica: ¿qué quiere decir elegir un número natural al azar?. Para ello, fijado un $N \in \mathbb{N}$ consideramos el conjunto $\Omega_N = \{n \in \mathbb{N} : 1 \leq n \leq N\}$ como un espacio muestral discreto en el que asignamos probabilidades de acuerdo con la definición clásica de Laplace:

$$P_N(A) = \frac{\#(A)}{N} \quad A \subset \Omega_N$$

Si queremos asignar a eventos $A \subset \Omega = \mathbb{N}$, resulta natural entonces tomar el límite cuando $N \to \infty$, y definir

$$P(A) = \lim_{N\to\infty} P_N(A \cap \Omega_N) \quad A \subset \mathbb{N}$$

siempre que este límite exista

Por ejemplo: ¿cuál es la probabilidad de que un número natural elegido al azar sea par? De acuerdo a esta definición si $D_2 = \{n \in \mathbb{N} : n \text{ es par }\}$, entonces

$$P(D_2) = \lim_{N\to\infty} \frac{1}{N}\left[\frac{N}{2}\right] = \lim_{N\to\infty} \frac{1}{N}\left(\frac{N}{2} + O(1)\right) = \frac{1}{2}$$

(donde los corchetes indican la parte entera de $\frac{N}{2}$), que está de acuerdo con nuestra intuición. Más generalmente, si $d \in \mathbb{N}$, y consideramos el evento

$$D_d = \{n \in \mathbb{N} : n \text{ es divisible por } d\}$$

un argumento similar muestra que

$$P(D_d) = \frac{1}{d} \tag{11.7}$$

como esperamos[3].

Sin embargo, hay que ser cuidadosos, porque esta noción de probabilidad no es σ-aditiva (es decir: se sale del marco de Kolmogorov en el que venimos trabajando[4].). Por ejemplo, $P(\mathbb{N}) = 1$ pero

$$\mathbb{N} = \bigcup_{n\in\mathbb{N}} \{n\}$$

[3]Para una discusión más detallada de este concepto, ver [San55]

[4]Sin embargo, es posible formalizarla en el contexto más general de las álgebras de probabilidad condicional propuesto por Renyi [Ren78]

y $P(\{n\}) = 0$.

Para $n \in \mathbb{N}$ consideremos ahora la función $\omega(n)$ que cuenta el número de divisores primos distintos de n. Por ejemplo,

$$360 = 2^3 \times 3^2 \times 5^1 \Rightarrow \omega(360) = 3$$

.

Entonces se tiene el siguiente teorema:

Teorema 11.6.1 (Teorema del límite central de Erdős–Kac,[GS07]) *La distribución de $\omega(n)$ es asintóticamente normal, en el siguiente sentido:*

$$\lim_{N\to\infty} P_N\left(\left\{n \leq N : a \leq \frac{\omega(n) - \log\log n}{\sqrt{\log\log n}} \leq b\right\}\right) = G(b) - G(a)$$

Podemos interpretar la intuición detrás de este teorema de la siguiente manera: consideremos el conjunto de los primos numerado en forma creciente

$$\mathbb{P} = \{p_1 = 2, p_2 = 3, p_3 = 5, p_4 = 7, p_5 = 11, p_6 = 13, \ldots\},$$

y para cada $k \in \mathbb{N}$ definamos la función (variable aleatoria)

$$X_k(n) = \begin{cases} 1 & \text{si} \quad p_k \text{ divide a n} \\ 0 & \text{si} \quad 0 \text{ si no} \end{cases}$$

Las X_k se comportan como variables aleatorias independientes pues de acuerdo con 11.7:

$$P\{X_j = 1, X_k = 1\} = \frac{1}{p_j p_k} = \frac{1}{p_j} \cdot \frac{1}{p_k} = P\{X_j = 1\} \cdot P\{X_k = 1\}$$

En consecuencia como

$$\omega(n) = \sum_{k=1}^{\infty} X_k(n)$$

(Esta suma es en realidad finita para cada n, pues basta sumar los primos con $p_k \leq n$), vemos que ω se comporta como una suma de variables aleatorias independientes, y esto explica porqué el teorema del límite central se aplique a ella. Sin embargo, hacer riguroso este argumento requiere argumentos de teoría de las cribas. Una prueba relativamente sencilla aparece en [GS07].

Capítulo 12

Esperanza Condicional

12.1. Esperanza condicional respecto de un evento

Sea B un evento de probabilidad positiva. Recordamos que la probabilidad condicional de que ocurra el evento A sabiendo que ocurre el evento B, notada $P(A/B)$ se define por:

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

Sea $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria discreta. Recordamos que la esperanza de X se define como la serie

$$E[X] = \sum_i x_i P\{X = x_i\}$$

donde Im $= \{x_i\}$ es por hipótesis a lo sumo numerable; siempre que dicha serie sea absolutamente convergente.

En consecuencia, resulta natural definir la esperanza de X dado que ocurre el evento A de probabilidad positiva, por:

$$E[X/A] = \sum_i x_i P\{X = x_i/A\}$$

Teniendo en cuenta la definición de probabilidad condicional esto es equivalente a:

$$E[X/A] = \sum_i x_i \frac{P(\{X = x_i\} \cap A\}}{P(A)} = \frac{1}{P(A)} \sum_i x_i I_A(x_i) P\{X = x_i\}$$

Es decir que:

$$E[X/A] = \frac{1}{P(A)} E[I_A X] \tag{12.1}$$

Notemos que esta fórmula puede adoptarse como definición de la esperanza condicional respecto de un evento para cualquier variable aleatoria (sea discreta o no) mientras tenga esperanza finita, y el evento A tenga probabilidad positiva.

12.1.1. Un ejemplo con una variable discreta

Supongamos que $X \sim \mathcal{P}(\lambda)$ donde $\lambda > 0$. Recordamos que su distribución puntual viene dada por

$$p_k = P\{X = k\} = e^{-\lambda} \cdot \frac{\lambda^k}{k!} \quad k \in \mathbb{N}_0$$

y que $E[X] = \lambda$. Pero supongamos que ahora sabemos que $X \geq 1$. Entonces nuestra estimación de las probabilidades cambiará Notamos que

$$P\{X = 0\} = p_0 = e^{-\lambda} \Rightarrow P(A) = 1 - e^{-\lambda}$$

Tendremos la **distribución condicional**

$$P\{X = k/A\} = \begin{cases} 0 & \text{si} \quad k = 0 \\ \frac{e^{-\lambda}}{1-e^{-\lambda}} \cdot \frac{\lambda^k}{k!} & \text{si} \quad k \geq 1 \end{cases}$$

Estamos interesados en calcular $E[X/A]$ siendo $A = \{X \geq 1\}$.

$$\begin{aligned} E[X/A] &= \frac{1}{P(A)} \sum_k x_k \cdot I_A(x_k) P\{X = x_k\} \\ &= \frac{e^{-\lambda}}{1 - e^{-\lambda}} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} \end{aligned}$$

Pero haciendo un cambio de índice $j = k - 1$:

$$\sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \sum_{j=0}^{\infty} \frac{\lambda^{j+1}}{j!} = \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda e^{\lambda}$$

(esta cuenta es la misma que para calcular $E[X]$!). Nos queda:

$$E[X/A] = \frac{\lambda}{1 - e^{-\lambda}}$$

12.1.2. Un ejemplo con una variable continua

Supongamos que $X \sim N(0, 1)$. Entonces $E[X] = 0$.

Pero supongamos que además sabemos que $X > 0$. Entonces nuestra estimación de las probabilidades cambia, y ahora estamos interesados en calcular $E[X/A]$ siendo $A = \{X > 0\}$. Notamos que

$$P(A) = \int_0^\infty \phi(x)\, dx = \frac{1}{2} \text{ donde } \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$E[X/A] = \frac{1}{P(A)} E[I_A X] = \frac{1}{P(A)} E[X^+] = \frac{1}{1/2} \int_{-\infty}^\infty x^+ \cdot \phi(x)\, dx$$
$$= \frac{2}{\sqrt{2\pi}} \int_0^\infty x \cdot e^{-x^2/2}\, dx \text{ donde } x^+ = \begin{cases} x & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

Haciendo el cambio de variable $y = x^2/2$ vemos que:

$$E[X/A] = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-y}\, dy = \sqrt{\frac{2}{\pi}}$$

Generalización

Si X es una variable continua con densidad $f(x)$, $U \subset \mathbb{R}$ abierto y $A = \{X \in U\}$.

$$E[X/A] = \frac{1}{P(A)} E[I_A X] = \frac{1}{P(A)} E[g_U(X)]$$

$$= \frac{1}{P(A)} \int_{-\infty}^\infty g_U(x) f(x)\, dx = \frac{1}{P(A)} \int_U f(x)\, dx$$

donde $g_U(x) = x$ si $x \in U$ y 0 si $x \notin U$.

12.2. Esperanza comdicional de una variable con respecto a otra: caso discreto

Ahora consideremos dos variables discretas $X, Y : \Omega \to \mathbb{R}$. Nos proponemos definir el concepto de esperanza condicional $E[X/Y]$ de X dada Y. Supondremos que X tiene esperanza finita.

Sean $\{y_j\}$ los distintos valores que toma la variable Y, y notemos que los eventos $A_j = \{\omega \in \Omega : Y(\omega) = y_j\}$ forman una partición del espacio muestral Ω.

Si $P\{Y = y_j\} > 0$, podemos definir

$$E[X/Y = y_j] = E[X/A_j]$$

utilizando la definición introducida en la sección anterior.

Más explícitamente:

$$E[X/Y = y_j] = \sum_i x_i P\{X = x_i/Y = y_j\} \tag{12.2}$$

Las probabilidades $P\{X = x_i/Y = y_j\}$ que aparecen en esta definición se llaman la distribución condicional de probabilidades de X dada Y.

Notemos que depende del valor y_j de la variable Y. En consecuencia, $E[X/Y]$ puede considerarse como una nueva variable aleatoria. Más explícitamente, definimos $E[X/Y] : \Omega \to \overline{\mathbb{R}}$ por:

$$E[X/Y](\omega) = E[X/Y = Y(\omega)]$$

Observación 12.2.1 *Si X e Y son variables discretas independientes, entonces*

$$P\{X = x_i/Y = y_j\} = P\{X = x_i\}$$

Luego

$$E[X/Y = y_j] = \sum_i x_i \cdot P\{X = x_i/Y = y_j\} = \sum_i x_i \cdot P\{X = x_i\} = E[X]$$

En consecuencia, con lo que $E[X/Y] = E[X]$ (una variable aleatoria constante), en este caso.

Observación 12.2.2 *En el otro extremo, ¿qué pasa cuando $Y = f(X)$ siendo $f : \mathbb{R} \to \mathbb{R}$?*

$$P\{Y = y_j/X = x_i\} = \begin{cases} 1 & si \quad y_j = f(x_i) \\ 0 & si \quad y_j \neq f(x_i) \end{cases}$$

Entonces:

$$E[Y/X = x_i] = \sum_j y_j \cdot P\{Y = y_j/X = x_i\} = f(x_i)$$

Es decir:

$$E[f(X)/X] = f(X)$$

En particular:

$$E[X/X] = X$$

Otras propiedades útiles son:

- Linealidad: si $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$E[\lambda_1 X_1 + \lambda_2 X_2/Y] = \lambda_1 \cdot E[X_1/Y] + \lambda_2 \cdot E[X_2/Y]$$

- Más generalmente, podemos sacar afuera de la esperanza condicional funciones de la variable con respecto a la que estamos condicionando:

$$E[f(Y)X/Y] = f(Y)E[X/Y]$$

porque:

$$\begin{aligned} E[f(Y)X/Y = y_j] &= \sum_i f(y_j \cdot) \cdot x_i \cdot P\{X = x_i/Y = y_j\} \\ &= f(y_j) \sum_i x_i \cdot P\{X = x_i/Y = y_j\} = f(y_j) \cdot E[X/Y = y_j] \end{aligned}$$

12.2.1. Un ejemplo

Consideramos el siguiente ejemplo. Tiramos dos dados en forma sucesiva. Nuestro espacio muestral es:

$$\Omega = \{\omega = (\omega_1, \omega_2) : \omega_i \in D\}$$

donde $D = \{1, 2, 3, 4, 5, 6\}$. Consideramos la suma S de los puntos obbtenidos. $S : \Omega \to \mathbb{R}$ Tenemos que $S = X_1 + X_2$ donde $X_1(\omega) = \omega_1, X_2(\Omega) = \omega_2$.

Tenemos que

$$E[S] = E[X_1] + E[X_2] = 3,5 + 3,5 = 7$$

Pero si sabemos cuánto salió en la primera tirada (o sea, cuándo vale X_1), nuestra estimación de las probabilidades para S cambia.

$$E[S/X_1] = E[X_1/X_1] + E[X_2/X_1] = X_1 + E[X_2] = X_1 + 3,5$$

12.2.2. Fórmula de la probabilidad total

$E[X/Y]$ es una nueva variable aleatoria. ¿Qué pasa si calculamos su esperanza? Recordamos que $A_j = P\{Y = y_j\}$ es una partición de Ω.

$$\begin{aligned} E[E[X/Y]] &= \sum_j E[X/Y = y_j] \cdot P(A_j) \\ &= \sum_j \frac{1}{P(A_j)} E[XI_{A_j}] \cdot P(A_j) \\ &= \sum_j E[XI_{A_j}] = \\ &= E\left[X\left(\sum_j I_{A_j}\right)\right] = E[X] \end{aligned}$$

Proposición 12.2.3 *Fórmula de la probabilidad total*

$$E[E[X/Y]] = E[X]$$

12.3. Esperanza condicional de una variable continua respecto de una discreta

La definición anterior de $E[X/Y]$,

$$E[X/Y] = \sum_j E[X/Y = y_j] \cdot I_{A_j} \text{ donde } A_j = \{Y = y_j\}$$

también se puede aplicar si X es una variable aleatoria continua, e Y una variable discreta (siempre que $P(A_j) > 0$).

Veamos un ejemplo (ejercicio de un parcial):

Enunciado:

Se tira un dado equilibrado de tres caras (o sea: se elege un número del 1 al 3 con idénticas probabilidades). Sea I el número obtenido en el dado. A continuación se define $Z = \sum_{j=1}^{I} X_j$ donde las variables aleatorias X_j tienen distribución exponencial de parámetro 1, y son todas independientes entre sí y del lanzamiento del dado.

i) Encuentre una expresión explícita para la densidad de probabilidad de Z.

ii) Utilizando dicha expresión, calcule $E[Z]$.

iii) Calcular $P(Z > 3)$.

Solución del item i)

Si conociéramos el valor i de I, tendríamos la variable

$$Z_i = \sum_{j=1}^{i} X_j$$

Sabemos que $Z_i \sim \Gamma(i, 1)$ por ser suma de suma de i variables aleatorias indepedendientes con distribución $\text{Exp}(1) = \Gamma(1, 1)$. [por un resultado que vimos en la clase 11]

Esta es una **distribución condicional**. ¡ Pero I es aleatoria! La verdadera distribución de Z se encuentra **mezclando** estas distribuciones condicionales, pesándolas de acuerdo a

la distribución de probabilidades de I,

$$\begin{aligned} f_Z(z) &= \sum_{i=1}^{3} f_{Z_i}(z) \cdot P\{I = i\} \\ &= \frac{1}{3}\left[\sum_{i=1}^{3} \frac{z^{i-1}}{(i-1)!}\right] \cdot I_{(0+\infty}(z) \end{aligned}$$

Recordamos que esta es otra aplicación de la fórmula de la probabilidad total.

Una vez determinada la distribución de Z su esperanza se encuentra mediante la fórmula de siempre.

$$E[Z] = \int_{-\infty}^{\infty} z \cdot f_Z(z)\, dz$$

Pero ahora podríamos pensar esta cuenta de otra manera

$$E[Z/I = i] = E[Z_i] = i$$

dado que ya calculamos la esperanza de una variable con distribución $\Gamma(i, 1)$ (en la clase 8). Entonces

$$E[Z] = E[E[Z/I]] = \sum_{i=1}^{3} E[Z/I = i] \cdot P\{I = i\} = \frac{1}{3}\sum_{i=1}^{3} i = \frac{1+2+3}{3} = 2$$

12.4. Esperanza condicional de variables continuas

La definición anterior tiene un serio problema si queremos generalizar el concepto de esperanza condicional $E[X/Y]$ cuando la variable aleatoria Y es continua: en general

$$P\{Y = y_0\}$$

puede ser cero, por lo que las probabilidades condicionales:

$$P\{X \in I/Y = y_0\}$$

donde I es un intervalo, no va a estar definida.

Vamos a investigar primero el caso en que X e Y admiten una densidad conjunta $f_{XY}(x, y)$ continua. Recordamos que en esta situación Y se distribuye según la densidad marginal

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx$$

Consideramos un pequeño intervalo $J = [yo, y_0 + h]$, entonces:

$$P\{X \in I / Y \in J\} = \frac{P\{X \in I, Y \in J\}}{P\{Y \in J\}} = \frac{\int_I \int_J f_{XY}(x,y)\, dxdy}{\int_J f_Y(y)\, dy}$$

Entonces elegimos $I = (-\infty, x]$ y dividimos arriba y abajo por h

$$P\{X \leq x / Y \in J\} = \frac{\frac{1}{h}\int_{-\infty}^{x} \int_{y_0}^{y_0+h} f_{XY}(x,y)\, dxdy}{\frac{1}{h}\int_{y_0}^{y_0+h} f_Y(y)\, dy}$$

Cuando $h \to 0$ esta expresión converge a

$$F_{X/Y=y_0}(x) = \frac{\int_{-\infty}^{x} f_{XY}(x, y_0)\, dx}{f_Y(y_0)}$$

por el teorema fundamental del cálculo. Esta expresión se llama **función de distribución condicional** de X dada Y. Esta cuenta tiene sentido sólo si $f_Y(y) > 0$.

De donde obtenemos la densidad **densidad condicional** de X dada Y dada por

$$f_{X/Y=y_0}(x) = \frac{f_{XY}(x, y_0)}{f_Y(y_0)}$$

que podemos pensar como una versión infinitesimal de la definición de probabilidad condicional.

Entonces podemos definir la esperanza condicional en este caso, integrando la densidad condicional:

$$E[X/Y = y_0] = \int_{\infty}^{\infty} x\, dF_{X/Y=y_0}(x) = \int_{-\infty}^{\infty} x\, f_{X/Y=y_0}(x)\, dx$$

Todas las propiedades anteriores van a seguir valiendo con esta definición.

12.4.1. Un ejemplo: Esperanzas condicionales en la distribución normal bivariada

Recordamos que la **distribución normal bivariada** es la distribución de un vector aleatorio

$$\begin{pmatrix} X \\ Y \end{pmatrix} = A \cdot \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} + \mu$$

donde $Z_1, Z_2 \sim N(0,1)$ son independientes,

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \in \mathbb{R}^2$$

y $A \in \mathbb{R}^{2\times 2}$ es una matriz no singular. Encontramos que su densidad conjunta es

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2-2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]\right\}}$$

donde

$$\Sigma = A\cdot A^t = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

es la matriz de covariancias, y ρ es el **coeficiente de correlación** entre X e Y.

Proposición 12.4.1 *Si el vector (X,Y) se distribuye según la densidad normal bivariada $N(\mu,\Sigma)$, entonces*

$$E[Y|X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)$$

Esto dice, que en este caso la esperanza condicional está dada por la recta de regresión lineal.

Corolario 12.4.2 *Por simetría, en eesta situación*

$$E[X|Y] = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y-\mu_Y)$$

Prueba: Buscamos la **descomposición de Cholesky** de la matriz de covariancia. Es decir buscamos $A = \text{Chol}(Z)$ triangular tal que

$$A\cdot A^t = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}\cdot\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} a^2 & ab \\ ab & b^2+c^2 \end{pmatrix} = \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Nos quedan tres ecuaciones con tres incógnitas

$$a^2 = \sigma_x^2, \quad ab = \rho\sigma_X\sigma_Y, \quad b^2+c^2 = \sigma_Y^2$$

Entonces

$$\begin{aligned} a &= \sigma_X \\ b &= \rho\sigma_X\sigma_Y/a = \rho\sigma_Y \\ c &= \sqrt{\sigma_Y^2 - b^2} = \sigma_Y(1-\rho^2)^{1/2} \end{aligned}$$

Usando el resultado del ejercicio que mencionamos antes podemos escribir:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \mu + \text{Chol}(\Sigma)Z$$

donde Z es un vector con distribución normal bivariada estándar, es decir con componentes Z_1, Z_2 que son $N(0,1)$ independientes.

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} + \begin{pmatrix} \sigma_X & 0 \\ \rho\sigma_Y & \sigma_Y(1-\rho^2)^{1/2} \end{pmatrix} \cdot \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

de donde

$$\begin{aligned} X &= \mu_X + \sigma_X Z_1 \\ Y &= \mu_Y + \sigma_Y[\rho Z_1 + (1-\rho^2)^{1/2} Z_2] \end{aligned}$$

Entonces:

$$\begin{aligned} E[Y/X] &= E\left[\mu_Y + \sigma_Y[\rho Z_1 + (1-\rho^2)^{1/2} Z_2/X\right] \\ &= E\left[\mu_Y + \sigma_Y\left(\rho\frac{X-\mu_X}{\sigma_X}\right) + (1-\rho^2)^{1/2} Z_2/X\right] \\ &= E\left[\mu_Y/X\right] + E\left[\sigma_Y\left(\rho\frac{X-\mu_X}{\sigma_X}\right)/X\right] + E\left[(1-\rho^2)^{1/2} Z_2/X\right] = \\ &= \mu_Y + \sigma_Y\left(\rho\frac{X-\mu_X}{\sigma_X}\right) + (1-\rho^2)^{1/2} E\left[Z_2/X\right] \end{aligned}$$

Ahora X es una función de Z_1, y Z_1, Z_2 eran independientes. Se deduce que X es independiente de Z_2. Y la relación de independencia entre las variables es simétrica. Luego:

$$E[Z_2/X] = E[Z_2] = 0$$

Se deduce que:

$$E[Y/X] = \mu_Y + \sigma_Y\left(\rho\frac{X-\mu_X}{\sigma_X}\right)$$

□

Lema 12.4.3 *La variable aleatoria $h(Y) = E[X/Y]$ tiene las siguientes propiedades:*

- *Tiene esperanza finita.*
- *Para cualquier función $f : \mathbb{R} \to \mathbb{R}$ acotada, se verifica que:*

$$E[f(Y)h(Y)] = E[f(Y)X]$$

Más aún: la esperanza condicional $E[X/Y]$ está caracterizada por estas dos propiedades. en el siguiente sentido: si $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ son dos funciones que verifican estas dos propiedades, entonces

$$P\{h_1(Y) = h_2(Y)\} = 1$$

Prueba: Para probar que $h(Y)$ tiene esperanza finita, debemos mostrar que la serie

$$\sum_j h(y_j) P\{Y = y_j\}$$

donde (y_j) recorre los posibles valores que la variable Y toma con probabilidad positiva, es absolutamente convergente.

$$\sum_j |h(y_j)| P\{Y = y_j\} = \sum_j \left| \sum_i x_i P\{X = x_i / Y = y_j\} \right| P\{Y = y_j\}$$

$$\leq \sum_i \sum_j |x_i| P\{X = x_i, Y = y_j\} = E(|X|) < +\infty$$

Para probar la segunda afirmación calculamos:

$$E[f(Y)h(Y)] = \sum_j f(y_j) h(y_j) P\{Y = y_j\ \}$$

$$= \sum_i f(y_j) P\{Y = y_j\} \sum_i x_i P\{X = x_i / Y = y_j\}$$

$$= \sum_i \sum_j f(y_j) x_i P\{X = X_i, Y = y_j\} = E[f(Y)X]$$

donde el reordenamiento de la serie se justifica utilizando que dicha serie converge absolutamente (dado que f es acotada).

Ahora probaremos la unicidad: supongamos que $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ son funciones que verifican las propiedades anteriores. Entonces para cualquier función $f : \mathbb{R} \to \mathbb{R}$ acotada, tenemos que:

$$E[f(Y)h_1(Y)] = E[f(Y)h_2(Y)] = E[f(Y)X]$$

En consecuencia, si llamamos $h = h_1 - h_2$ por la linealidad de la esperanza:

$$E[f(Y)h(Y)] = 0$$

Eligiendo $f(t) = I_{\{y_j\}}(t)$ deducimos que:

$$h(y_j) P\{Y = y_j\} = 0$$

Por lo tanto si $h(y_j) \neq 0$, $P\{Y = y_j\} = 0$. En consecuencia:

$$P\{h(Y) \neq 0\} = \sum_{y_j : h(y_j) \neq 0} P\{Y = y_j\} = 0$$

Es decir que: $P\{h_1(Y) = h_2(Y)\} = 1$. □

Corolario 12.4.4

$$E[E[X/Y]] = E[X]$$

(Se deduce tomando $f \equiv 1$ en la fórmula anterior).

12.4.2. Un detalle muy técnico

Recordamos que la σ-álgebra de Borel se define como la σ-álgebra de subconjuntos de $\mathbb{R}$ generada por los intervalos (abiertos).

Definición 12.4.5 *Una función $h : \mathbb{R} \to \mathbb{R}$ se dice boreliana si es medible respecto a la σ-álgebra de Borel, o sea que $h^{-1}(I)$ es un conjunto boreliano para todo intervalo abierto $I \subset \mathbb{R}$.*

Nota: Esta definición garantiza que si $X : \Omega \to \mathbb{R}$ es una variable aleatoria, $h(X) = h \circ X : \Omega \to \mathbb{R}$ también lo es. pues

$$(h \circ X)^{-1}(I) = X^{-1}(h^{-1}(I)$$

Entonces si I es un intervalo, $h^{-1}(I)$ es un conjunto boreliano y entonces $X^{-1}(h^{-1}(I))$ es un evento (le podemos asignar una probabilidad).

Notemos que si $h : \mathbb{R} \to \mathbb{R}$ es continua, es boreliana.

Teniendo en cuenta las observaciones anteriores, es posible adoptar la siguiente **definición axiomática** de la esperanza condicional:

Definición 12.4.6 *Sean $X, Y : \Omega \to \overline{\mathbb{R}}$ variables aleatorias. Decimos que una variable aleatoria $Z = h(Y)$ es una versión de la esperanza condicional $E[X/Y]$ si donde $h : \mathbb{R} \to \mathbb{R}$ es una función boreliana, si se verifican las siguiente propiedades:*

1. *$h(Y)$ tiene esperanza finita.*
2. *Para cualquier función boreliana acotada $f : \mathbb{R} \to \mathbb{R}$ se verifica que:*

$$E[f(Y)h(Y)] = E[f(Y)X]$$

12.4.3. El caso continuo

Haciendo las mismas cuentas de antes, pero con integrales en lugar de sumas, y densidades en lugar de distribuciones puntuales, se prueba:

Teorema 12.4.7 *Si el vector (X, Y) se distribuye según la densidad de probabilidad conjunta f_{XY} y $E(|X|) < \infty$. Supongamos además que*

$$f_Y(y) > 0 \ \forall y \in \mathbb{R}$$

entonces

$$h(y) = \int_{-\infty}^{\infty} x \ f_{X/Y=y}(x) \ dx$$

donde

$$f_{X/Y=y_0}(x) = \frac{f_{XY}(x, y_0)}{f_Y(y_0)}$$

es la densidad condicional, proposciona una versión de la esperanza condicional $E[X/Y]$.

12.4.4. Teorema de existencia

El siguiente teorema afirma que siempre existe una versión de la esperanza condicional, aunque no proporciona ninguna fórmula para calcularla. No demostraremos este teorema ya que su demostración depende de un teorema de análisis real (el teorema de Radon-Nikodym)

Teorema 12.4.8 *Si $X, Y : \Omega \to \overline{\mathbb{R}}$ son variables aleatorias, siempre existe una versión de la esperanza condicional $E[X/Y]$. Además si $h_1(Y)$, $h_2(Y)$ son dos versiones de la esperanza condicional $E[X/Y]$, entonces*

$$P\{h_1(Y) = h_2(Y)\} = 1$$

Proposición 12.4.9 (Unicidad) *Si $\widehat{Y}_1$ y $\widehat{Y}_2$ verifican la definición axiomática de la esperanza condicional, entonces $\widehat{Y}_1 = \widehat{Y}_2$ con probabilidad 1.*

Prueba: Sea $W = \widehat{Y}_1 - \widehat{Y}_2$. $\widehat{Y}_1 = h_1(X), \widehat{Y}_2 = h_2(X)$. Entonces

$$W = h(X) \text{ con } h = h_1 - h_2$$

$$E[WZ] = E[\widehat{Y}_1 Z] - E[\widehat{Y}_2 Z] = E[YZ] - E[YZ] = 0$$

para toda $Z = f(X)$ con f boreliana acotada. Elegimos $f(x) = I_{\{h(x)>\delta\}}$. Tenemos

$$\delta \cdot P(A_\delta) \leq E[W \cdot I_{A_\delta}] = 0 \text{ donde } A_\delta = \{W > \delta\}$$

Luego $P(A_\delta) = 0$ para todo $\delta > 0$, se deduce que $W \leq 0$ con probabilidad 1. Cambiando W por $-W$, vemos que también $W \geq 0$ con probabilidad 1, luego $W = 0$ o sea $\widehat{Y}_1 = \widehat{Y}_2$ con probabilidad 1. □

12.5. Propiedades de la esperanza condicional

Las propiedades de la esperanza condicional se pueden deducir de la definición axiomática. Por ejemplo:

Proposición 12.5.1 (Linealidad) *Sean $Y_1, Y_2 \in L^1(\Omega)$. Si $c_1, c_2 \in \mathbb{R}$,*

$$E[c_1 \cdot Y_1 + c_2 \cdot Y_2/X] = c_1 \cdot E[Y_1/X] + c_2 E[Y_2/X]$$

Prueba: Sean $\widehat{Y_1} = E[Y_1/X]$, $\widehat{Y_2} = E[Y_2/X]$, $Y = c_1 \cdot Y_1 + c_2 \cdot Y_2$ Hay que verificar que $\widehat{Y} = c_1\widehat{Y_1} + c_2\widehat{Y_2}$ cumple con la defnición axiomática de esperanza condicional.

- $\widehat{Y}$ es función de X porque $\widehat{Y_1}$ e $\widehat{Y_2}$ lo son.
- $\widehat{Y}$ tiene esperanza finita, pues $E[\widehat{Y}] = c_1 E[\widehat{Y_1}] + c_2 E[\widehat{Y_2}]$
- Si $Z = f(X)$ con f acotada, entonces

  $$E[\widehat{Y}Z] = c_1 E[\widehat{Y_1}Z] + c_2[\widehat{Y_2}Z] = c_1 E[Y_1 Z] + c_2[Y_2 Z] = E[YZ]$$

  Por la unicidad de la esperanza condicional, vale la propiedad.

□

Otras propiedades que también pueden demostrarse a partir de la definición axiomática son:

Proposición 12.5.2

- *Si $Y \in L^1$, $E[E[Y/X]] = E[Y]$*
- *Sea $Y \in L^1$, $g : \mathbb{R} \to \mathbb{R}$ boreliana acotada, X otra variable aleatoria: $E[Y \cdot g(X)/X] = g(X)E[Y/X]$.*
- *Monotonía: si $Y_1, Y_2 \in L^1$, $Y_1 \leq Y_2$ con probabilidad 1,*

  $$E[Y_1/X] \leq E[Y_2/X]$$

- *Desigualdad de Cauchy-Schwartz: Si $X, Y \in L^2$,*

  $$E(Y_1 Y 2/X) \leq E[Y_1/X]^{1/2} \cdot E[Y_2/X]^{1/2}$$

- *Desigualdad de Jensen: si $Y \in L^1$, $\varphi : \mathbb{R} \to \mathbb{R}$ es convexa y $\varphi(Y) \in L^1$,*

  $$\varphi(E[Y/X]) \leq E[\varphi(Y)/X)$$

  En particular:

  $$|E[Y/X])|^p \leq E[|Y|^p/X) \text{ si } p \geq 1$$

- *Condicionamiento reiterado:*

  $$E[E[X/Y]/Z] = E[X/Z]$$

12.6. La esperanza condicional como proyección ortogonal

Planteo del problema

Recordamos que la idea al definir $E[X/Y]$ es estimar X por medio de una función de Y. Formalizaremos esta intuición usando las ideas que introdujimos en el capítulo 7.

El enfoque que vamos a desarrollar sólo funciona para variables aleatorias con segundo momento finito (mientras que como vimos anteriormente $E[X/Y]$ se puede definir en general con sólo asumir que $E[|X|] < \infty$).

En dicho capítulo, planteamos el problema de aproximar una variable $Y \in L^2(\Omega)$ por un elemento $\widehat{Y}$ de un subespacio S, de modo de minimizar el **error cuarático medio**

$$\text{ECM}(Y, \widehat{Y}) = E(|Y - \widehat{Y}|^2) = \|Y - \widehat{Y}\|^2$$

Dadas otra variable aleatoria X, vamos a considerar ahora el subespacio:

$$S = \{Y \in L^2(\Omega) : Y = h(X) \text{ donde } h : \mathbb{R} \to \mathbb{R}\}$$

Por razones técnicas tenemos que pedir que h sea una función boreliana como mencionamos antes: es decir que $h^{-1}(I)$ sea un conjunto boreliano para cada intervalo abierto I en $\mathbb{R}$. Por ejemplo, cualquier función continua va a cumplir esto.

Entonces, si $X, Y \in L^2(\Omega)$ podemos definir la **esperanza condicional** $E[Y/X]$ como la solución $\widehat{Y}$ de este problema de optimización.

Aplicando el lema 7.3.1, vemos que la esperanza condicional $\widehat{Y} = E[Y/X]$ se define por las siguientes dos propiedades:

- $\widehat{Y} \in S$.
- $E[(Y - \widehat{Y}) \cdot Z] = 0$ para toda $Z \in S$, o sea:

$$E[Y \cdot Z] = E[\widehat{Y} \cdot Z] \text{ para todo } Z \in S$$

Son esencialmente las mismas condiciones de la **definición axiomática de la esperanza condicional** que vimos antes. La única diferencia, es que allí trabajamos con variables con esperanza finita, entonces tuvimos que pedir que $Z = f(Y)$ con $f : \mathbb{R} \to \mathbb{R}$ acotada (para poder garantizar que las esperanzas que aparecen aquí sean finitas).

12.6.1. El caso en que la variable Y es discreta

Como vimos en la clase pasada, el caso más sencillo de la esperanza condicional $E[Y/X]$ es cuando la variable X es discreta. Para simplificar vamos a suponer que

$$Im(X) = \{x_1, x_2, \ldots, x_n\}$$

es finita y que $p_j = P\{X = x_j\} > 0$. Notamos que los eventos

$$A_j = \{X = x_j\} = \{\omega \in \Omega : X(\omega) = x_j\}$$

forman una **partición** de Ω. Vamos a suponer que $P(A_j) > 0$. En este caso S es de dimensión finita, y una base de S está formada por sus funciones indicadoras

$$B = \{I_{A_1}, I_{A_2}, \ldots, I_{A_n}\}$$

La condición de que $\widehat{Y} = E[Y/X] \in S$ dice que

$$E[Y/X] = \sum_{k=1}^{n} c_j \cdot I_{A_j}$$

para ciertos escalares c_j que queremos determinar.

Ahora miramos la condición

$$E[Y \cdot Z] = E[\widehat{Y} \cdot Z] \text{ para todo } Z \in S$$

Como B es una base de S, alcanza mirar esta condición para $Z = I_{A_j}$. Por otra parte, como los A_j son disjuntos, resulta que B es una **base ortogonal** (pero no ortonormal) de S, pues

$$\langle I_{A_j}, I_{A_k} \rangle = E(I_{A_j} \cdot I_{A_k}) = \begin{cases} P(A_j) & \text{si} \quad j = k \\ 0 & \text{si} \quad j \neq k \end{cases}$$

Nos queda:

$$E[Y \cdot I_{A_j}] = c_j \cdot P(A_j)$$

entonces:

$$c_j = \frac{1}{P(A_j)} E[Y \cdot I_{A_j}] = E[Y/A_j]$$

que coincide con la defínción que vimos en la clase pasada.

En resumen, cuando X es dicreta con imagen finita:

$$E[Y/X] = \sum_{i=1}^{n} E[Y/A_j] \cdot I_{A_j}$$

Esta fórmula puede generalizarse al caso en que X tiene imagen numerable (en este caso S no es de dimensión finita, y en lugar de una suma finita tenemos una serie, pero esencialmente funciona igual).

$$E[Y/X] = \sum_{i=1}^{\infty} E[Y/A_j] \cdot I_{A_j}$$

En este caso, debemos comprobar que esta fórmula define en efecto una función en $L^2(\Omega)$. Como las I_{A_j} son ortogonales

$$\begin{aligned} E[E[Y/X]^2] = \|E[Y/X]\|^2 &= \sum_{j=1}^{\infty} \|E[Y/A_j] \cdot I_{A_j}\|^2 \\ &= \sum_{j=1}^{\infty} |E[Y/A_j]|^2 \|I_{A_j}\|^2 \\ &= \sum_{j=1}^{\infty} |E[Y/A_j]|^2 P(A_j) \end{aligned}$$

Ahora por la **desigualdad de Cauchy-Schwarz**:

$$|E[Y/A_j]| \leq \frac{1}{P(A_j)} E(|Y I_{A_j}|) \leq \frac{1}{P(A_j)} \|Y I_{A_j}\| \cdot \|I_{A_j}\| = \|Y\| = E(I_{A_j} Y^2)^{1/2} \frac{1}{P(A_j)^{1/2}}$$

Entonces

$$E[E[Y/X]^2] \leq \sum_{i=1}^{\infty} |E[I_{A_j} Y]|^2 = E(|Y|^2) < +\infty$$

Esperanzas condicionales en el caso continuo

El otro caso que vimos anteriormente es cuando X e Y son variables continuas con una densidad conjunta $f_{XY}(x, y)$.

Recordamos que en este caso definimos la densidad condicional

$$f_{Y/X=x}(y) = \frac{f_{XY}(x, y)}{f_X(x)}$$

donde

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy$$

es la **densidad marginal** de X, suponiendo que $f_X(y) > 0$ para todo y. En la clase anterior, definimos $E[Y/X] = h(X)$ donde

$$h(x) = \int_{-\infty}^{\infty} y \cdot f_{Y/X=x}(y)\, dy$$

Si Y tiene esperanza finita, podemos calcular

$$\begin{aligned} E[|E[Y/X]] = E[|h(X)|] &= \int_{-\infty}^{\infty} |h(x)|\ f_X(x)\ dx \\ &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} y \cdot f_{Y/X=x}(x)\ dy \right|\ f_X(x)\ dx \\ &\leq \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} |y| \cdot f_{Y/X=x}(x)\ dy \right]\ f_X(x)\ dx \\ &= \int_{-\infty}^{\infty} |y| \left[\int_{-\infty}^{\infty} f_{Y/X=x}(x)\ f_X(x) dx \right]\ dy \\ &= \int_{-\infty}^{\infty} |y| \left[\int_{-\infty}^{\infty} f_{XY}(x,y) dx \right]\ dy \\ &= \int_{-\infty}^{\infty} |y| \cdot f_Y(y)\ dy = E(|Y|) < \infty \end{aligned}$$

Se deduce en particular que $h(x)$ es finita para casi todo x, por lo que $E[X/Y]$ está bien definida.

Vamos a comprobar ahora que si $Z = g(X)$ con $g : \mathbb{R} \to \mathbb{R}$ acotada, entonces $E[Y \cdot Z] = E[\widehat{Y} \cdot Z]$

Para calcular $E[Y \cdot Z]$ la pensamos como la esperanza de una función del vector aleatorio (X, Y).

$$E[Y \cdot Z] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \cdot g(x) \cdot f_{XY}(x,y)\ dx\ dy$$

mientras que como $\widehat{Y} \cdot Z$ es una función de X sola,

$$E[\widehat{Y} \cdot Z] = \int_{-\infty}^{\infty} h(x) \cdot g(x) \cdot f_X(x)\ dx$$

pero por la definición de h,

$$h(x)\ f_X(x) = \left[\int_{-\infty}^{\infty} y \cdot f_{Y/X=x}(y)\ dy \right] \cdot f_X(x) = \int_{-\infty}^{\infty} y \cdot f_{XY}(x,y)\ dy$$

Reemplazando vemos que

$$\begin{aligned} E[\widehat{Y} \cdot Z] &= \int_{-\infty}^{\infty} g(x) \left[\int_{-\infty}^{\infty} y \cdot f_{XY}(x,y)\ dy \right]\ dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\ g(x) f_{XY}(x,y)\ \ dx\ dy = E[Y \cdot Z] \end{aligned}$$

Vemos que $h(Y)$ verfica las dos condiciones de la definición axiomática de esperanza condicional que vimos en la clase anterior.

Si supiéramos que $Y \in L^2$ va a resultar que $\widehat{Y}$ está en L^2 y que la misma cuenta se puede hacer suponiendo que Z está en L^2 (aunque g no fuera acotada).

Para probar que en efecto $\widehat{Y} = E[Y/X] \in L^2$ cuando $Y \in L^2$, la idea es aproximar h por funciones acotadas:

$$h_n(x) = \begin{cases} h(x) & \text{si} \quad |h(x)| \leq n \\ n & \text{si} \quad h(x) > n \\ -n & \text{si} \quad h(x) < -n \end{cases} \qquad n \in \mathbb{N}$$

Entonces tomando $Z_n = h_n(X)$, tenemos que:

$$E[Y \cdot Z_n] = E[\widehat{Y} \cdot Z_n]$$

por lo que ya probamos. Luego por la **desigualdad de Cauchy-Schwarz**:

$$|E[\widehat{Y} \cdot Z_n]| \leq E(Y^2)^{1/2} \cdot E(Z_n^2)^{1/2}$$

pero como $|h_n(x)| \leq |h(x)|$, $|Z_n| \leq |\widehat{Y}|$, luego como $Y \in L^2$:

$$|E[\widehat{Y} \cdot Z_n]| \leq E(Y^2)^{1/2} \cdot E(\widehat{Y}^2)^{1/2} = E(\widehat{Y}^2)$$

Ahora bien, explícitamente

$$E[\widehat{Y} \cdot Z_n] = \int_{-\infty}^{\infty} h_n(x)\ h(x)\ f_X(x)\ dx$$

cuando $n \to +\infty$ esta integral va a converger a

$$E[\widehat{Y}^2] = \int_{-\infty}^{\infty} h(x)^2\ f_X(x)\ dx$$

porque $h_n(x)$ converge en forma monótona creciente hacia h. [Por el **teorema de convergencia monótona**, otro resultado clave de análisis real). Resulta que:

$$E[\widehat{Y}^2] \leq E[Y^2]$$

o sea:

$$E(E(Y/X)^2) \leq E(Y^2)$$

que es la misma desigualdad que obtuvimos antes en el caso discreto.

Capítulo 13

Estadística: Estimación de parámetros

13.1. Estimadores de máxima verosimilitud

Uno de los problemas centrales de la estadística es la **estimación de parámetros** de una distribución.

Supongamos que tenemos una población y queremos medir una cierta variable aleatoria, cuya distribución F no conocemos, pero sabemos o suponemos que $F \in \mathcal{F}$, una cierta familia de distribuciones.

Para estimar un parámetro $\theta = \theta(F)$, tomamos una muestra eleatoria de tamaño n de nuestra población. Esto nos dará variables

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

todas con distribución F e independientes. Entonces queremos estimar θ mediante un **estimador**

$$\widehat{\theta}(X_1, X_2, \ldots, X_n)$$

Por ejemplo, si μ es la esperanza de la distribución F, entonces:

$$\widehat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(conocido como **media muestral** es un estimador razonable de μ ya que por la **ley fuerte de los grandes números**

$$\overline{X}_n \xrightarrow{c.s.} \mu \quad \text{cuando } n \to +\infty$$

Se dice que $\overline{X}_n$ es un estimador fuertemente consistente para μ.

Similarmente, ¿Cómo podríamos estimar $\sigma^2 = \text{Var}(X) = E[(X-\mu)^2]$. Un estimador que podríamos considerar razonable es

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Este sería el valor de la varianza de la distribución empírica generada a partir de la muestra.

Vamos a ver que nuevamente, este estimador es fuertemente consistente, o sea:

$$\widehat{\sigma}_n \xrightarrow{c.s.} \sigma$$

Recordando que $\text{Var}(X) = E(X^2) - E(X)^2$ también tenemos:

$$\widehat{\sigma}_n^2 = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \left(\overline{X}_n\right)^2$$

Por la ley fuerte de los grandes números

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{c.s.} E(X^2)$$

Como elevar al cuadrado es una función continua

$$\widehat{\sigma}_n^2 \to E[X^2] - E(X)^2 = \sigma^2$$

y por lo tanto

$$\widehat{\sigma}_n \xrightarrow{c.s.} \sigma$$

Ahora bien: dado un parámetro λ pueden pensarse diferentes estimadores para λ que pueden parecer igualmente razonables.

Por ejemplo, supongamos que tenemos una población cuya distribución F sabemos que es normal $N(\mu, \sigma^2)$ con ciertos parámetros μ y σ como vimos antes. Entonces, para estimar μ podríamos usar la media muestral como vimos antes, porque μ es la esperanza de F.

Pero para la distribución normal μ también es la mediana. Por lo tanto otra forma de estimar μ podría ser usar la mediana muestral Me. Para definirla ordenamos las variables (o sea consideramos los estadísticos de orden):

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

y definimos

$$\text{Me} = X_{((n+1)/2)} \text{ si } n \text{ es impar}$$

$$\text{Me} = \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) \text{ si } n \text{ es par}$$

Esto lleva a preguntarnos qué propiedades es deseable que tenga un estimador, para tener un criterio para elegir un estimador sobre otro.

13.1.1. Sesgo de un estimador

Dado un estimador $\widehat{\lambda}_n$ de un parámetro $\lambda = \lambda(F)$, se define el **sesgo** del estimador como

$$\text{sesgo}(\widehat{\lambda}_n) = E[\widehat{\lambda_n}] - \lambda$$

Un estimador se dice **insesgado** si

$$\text{sesgo}(\widehat{\lambda}_n) = 0$$

y **asintóticamente insesgado** si

$$\text{sesgo}(\widehat{\lambda}_n) \to 0$$

13.1.2. Sesgo de la media muestral

Consideramos el estimador

$$\widehat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

para la esperanza $\mu = E[X]$.

Por la linealidad de la esperanza,

$$E[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Luego $\overline{X}_n$ es un estimador insesgado de μ.

13.1.3. Sesgo para el estimador de la varianza

Ahora repitamos la cuenta con el estimador de la varianza que definimos antes. Recordamos que:

$$\widehat{\sigma}_n^2 = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \overline{X}_n^2 \tag{13.1}$$

Así que empezemos calculando:

$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] = c$$

donde

$$c = E(X_i^2) = \text{Var}(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

Por otra parte, necesitamos calcular $E[\left(\overline{X}_n\right)^2]$. Para ello, la observación clave es que como las variables X_i son **independientes**

$$\mathrm{Var}(\left(\overline{X}_n\right)^2) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

Entonces:

$$E(\left(\overline{X}_n\right)^2) = \mathrm{Var}(\left(\overline{X}_n\right)^2) + E[\overline{X}_n]^2 = \frac{\sigma^2}{n} + \mu^2$$

Volviendo a (13.1) obtenemos que:

$$E[\widehat{\sigma}_n^2] = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2\left(1 - \frac{1}{n}\right) = \sigma^2\left(\frac{n-1}{n}\right)$$

Luego este estimador no resulta insesgado, pero sí asintóticamente insesgado.

13.1.4. Estimador insesgado de la varianza

Si queremos tener un estimador insegado de la varianza, debemos reemplazarlo por:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

que aparece en el ejercicio 24 de la práctica 8, ya que como

$$S_n^2 = \frac{n}{n-1}\cdot\widehat{\sigma}_n^2$$

ahora tendremos que:

$$E[S_n^2] = \sigma^2$$

13.2. Estimadores de Máxima Verosimilitud

Veremos ahora un método general para obtener estimadores con buenas propiedades: los estimadores de máxima verosimilitud.

Notación vectorial

En muchos ejemplos la distribución estará caracterizada por un número finito k de parámetros, que podemos pensar como componentes de un vector

$$\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k$$

que se mueve en una cierta región $A \subset \mathbb{R}^k$ de parámetros admisibles.

Por ejemplo, podemos pensar en la familia de distribuciones normales:

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

En este caso $\theta = (\mu, \sigma) \in A$, donde

$$A = \{(mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$$

En general, podemos escribir

$$\mathcal{F} = \{F_\theta : \theta \in A\}$$

13.3. Verosimilitud en el caso discreto

Comenzemos considerando el caso discreto. En este caso la distribución F_θ vendrá dadas por las probabilidades puntuales, que dependerán del vector de parámetros θ:

$$p_\theta(x) = P\{X_i = x\} \quad \text{(las mismas para todo)}$$

(que serán cero salvo para numerables valores de de x)

Por ejemplo, supongamos que tenemos una urna con un cierto número de bolitas blancas B y otro tanto de rojas R, y que extraemos n bolitas con reposición pero no conocemos cuántas bolitas de cada color hay. Definimos las variables aleatorias de Bernoulli

$$X_i = \begin{cases} 1 & \text{si} \quad \text{si sale roja} \\ 0 & \text{si} \quad \text{si sale blanca} \end{cases}$$

Entonces $X_i \sim \text{Be}(\theta)$ donde $\theta = \frac{R}{B+R} \in [0, 1] = A$. Aquí los posibles valores de las X_i son 0 y 1, y sus pobablidades

$$p_\theta(1) = \theta, \quad p_\theta(0) = 1 - \theta$$

Ahora nos preguntamos: si el parámetro θ tuviera un cierto valor, ¿cuál sería la probabilidad de observar ciertos valores $x_1, X_2, \ldots, x_n$?. Esto vendrá dado por la **función de verosimilitud**

$$\begin{aligned} \mathcal{L}(\theta) = \mathcal{L}(x_1, x_2, \ldots, x_n; \theta) &:= P_\theta\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \\ &= \prod_{i=1}^{n} P_\theta\{X = x_i\} \quad \text{por } \textbf{independencia} \\ &= \prod_{i=1}^{n} p_\theta(x_i) \end{aligned}$$

Aquí usamos la notación P_θ para indicar que las probabilidades indicadas dependen del parámetro θ.

Para cada muestra particular $(x_1, \ldots, x_n)$, la estimación de máxima verosimilitud de θ es el valor $\widehat{\theta}_{VM}$ que maximiza la verosimilitud. Es decir:

$$\mathcal{L}(x_1, x_2, x_n; \widehat{\theta}_{MV}) = \max_{\theta \in A} \mathcal{L}(x_1, \ldots, x_n; \theta)$$

El estimador de máxima verosimilitud, $\widehat{\theta}_{VM}(X_1, X_2, \ldots, X_n)$, es aquél que evaluado en cada muestra particular nos da la estimación de máxima verosimilitud

$$\widehat{\theta}_{MV}(x_1, x_2, \ldots, x_n)$$

Como $\mathcal{L}$ es un producto, conviene maximizar $\ell(s) = \log\ \mathcal{L}(\theta)$.

13.3.1. Estimación del parámetro de la distribución de Bernoulli

En el ejemplo que vimos antes de la distribución $\text{Be}(\theta)$:

$$\mathcal{L}(\theta) = \theta^s (1-\theta)^{n-s}$$

donde

$$s = x_1 + x_2 + \ldots + x_n$$

Luego:

$$\ell(\theta) = \log\ \mathcal{L}(\theta) = s \log \theta + (n-s) \log(1-\theta)$$

$$\ell'(\theta) = s \cdot \frac{1}{\theta} - (n-s) \cdot \frac{1}{1-\theta}$$

El máximo se va a alcanzar cuando $\ell'(s) = 0$, o sea:

$$\frac{s}{\theta} = \frac{n-s}{1-\theta} \Leftarrow \frac{1-\theta}{\theta} = \frac{n-s}{s} \Leftrightarrow \frac{1}{\theta} - s = \frac{n}{s} - 1 \Leftrightarrow \theta = \frac{s}{n}$$

Así que en este caso el mejor

$$\widehat{\theta}_{MV} = \overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

13.4. Verosimilitud en el caso continuo

Cuando trabajamos con variables continuas, la distribución F_θ estará caracterizada por una densidad de probabilidad f_θ. entonces definimos la función de **función de verosimilitud** como la densidad conjunta del vector aleatorio $(X_1, X_2, \ldots, X_n)$ correspondiente a

un determinado valor del parámetro θ, que de nuevo por la independencia de la muestra será:

$$\mathcal{L}(\theta) = \mathcal{L}(x_1, x_2, \ldots, x_n; \theta) := \prod_{i=1}^{n} f_\theta(x_i)$$

13.4.1. Estimación de los parámetros de la distribución normal

Volvamos al ejemplo de la familia de las distribuciones normales. Son distribuciones continuas con la densidad:

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)} \quad \theta = (\mu, \sigma)$$

Entonces:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{\sigma^n\sqrt{2\pi}} \prod_{i=1}^{n} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

luego

$$\ell(\theta) = \log\, \mathcal{L}(\theta) = -n\log\sigma - \frac{1}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Como ahora tenemos dos parámetros, para encontrar el máximo vemos donde se anulan simultáneamente ambas derivadas parciales:

$$\frac{\partial \ell}{\partial \mu}(\theta) = -\sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0 \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial \ell}{\partial \sigma}(\theta) = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \Rightarrow \sigma = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{1/2}$$

O sea que los **estimadores de máxima verosimilitud** para los parámetros de la distribución normal son:

$$\widehat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widehat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X}_n)^2\right)^{1/2}$$

13.5. Intervalos de confianza

13.5.1. Planteo del problema

Hasta ahora vimos como estimar los parámetros de una distribución. Por ejemplo si sabemos (o conjeturamos) que tenemos una muetra de la distribución normal $N(\mu, \sigma^2)$ podemos estimar los parámetros.

¿Pero cómo podemos estimar el error cometido en la estimación? Primero consideraremos el caso más sencillo aunque poco realista en que σ es conocido y queremos estimar μ. Sabemos que podemos estimar μ usando la medida muestral $\widehat{\mu}_n = \overline{X}_n$.

Nos gustaría encontrar un **intervalo de confianza** para μ, es decir un intervalo alrededor de $\widehat{\mu}_n$ tal que

$$P\{\mu \in I_\alpha\} = 1 - \alpha$$

donde $0 < \alpha < 1$ es un **nivel de confianza** elegido (típicamente $\alpha = 0,05$).

Ya nos encontramos con este concepto en un ejemplo que vimos en la clase 7 sobre las aproximaciones de la normal (encuesta electoral).

13.5.2. Solución cuando la varianza es conocida

Cuando la distribución es normal y σ es conocida podemos razonar así: $\overline{X}_n$ tendrá distribución $N\left(\mu, \frac{\sigma^2}{n}\right)$ Entonces:

$$Z_n = \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} \sim N(0,1)$$

Ahora elegimos $z_{\alpha/2}$ de modo que $P(Z_n > z_{\alpha/2}) = \alpha/2$, y por la simetría de la curva normal tenemos que

$$P\{-z_{\alpha/2} \leq -Z_n \leq z_{\alpha/2}\} = 1 - \alpha$$

Depejando obtenemos el intervalo de confianza

$$I_\alpha = \left[\overline{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \overline{X}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right]$$

para el que podemos garantizar que

$$P\{\mu \in I_\alpha\} = 1 - \alpha$$

13.5.3. Intervalos de confianza asintóticos

En la realidad, no es realista suponer que la distribución es conocida, o que la varianza lo es. De todos modos, podemos definir un intervalo de confianza asintótico para $\mu = E[X]$,

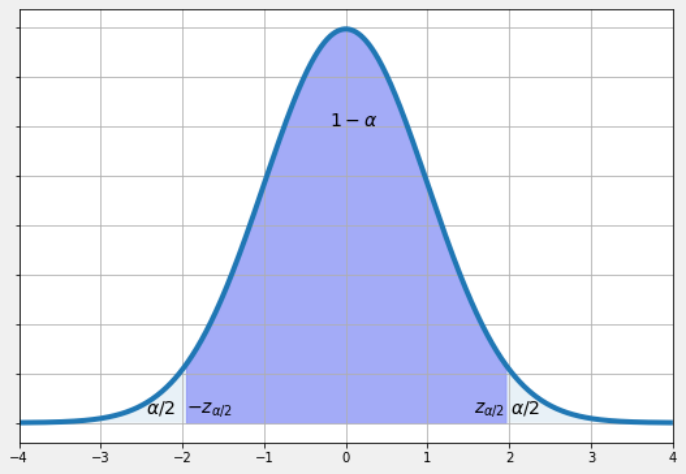


Figura 13.1: ¿Cómo elegimos $z_{\alpha/2}$ para determinar un intervalo de confianza?.

reemplazando a σ por un estimador fuertemente consistente $\widehat{\sigma}_n$ de los que vimos antes (da igual cuál consideremos)

$$I_\alpha = \left[\overline{X}_n - \frac{z_{\alpha/2}\widehat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{z_{\alpha/2}\widehat{\sigma}_n}{\sqrt{n}}\right]$$

Con sólo suponer que la variancia de la distribución tendremos que:

$$Z_n = \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{D} N(0,1)$$

por el teorema del límite central, siempre que $\sigma^2 = \text{Var}(X_i) < \infty$.

Como

$$\frac{\sigma}{\widehat{\sigma}_n} \xrightarrow{c.s.} 1$$

tendremos que la convergencia en distribución no se ve alterada:

$$\widehat{Z}_n = \frac{\sigma}{\widehat{\sigma}_n} Z_n = \sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\widehat{\sigma}_n} \xrightarrow{D} N(0,1)$$

por el teorema de Slutsky.

Por lo que nuestro intervalo

$$I_\alpha = \left[\overline{X}_n - \frac{z_{\alpha/2}\widehat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{z_{\alpha/2}\widehat{\sigma}_n}{\sqrt{n}}\right]$$

verfica que

$$\lim_{n\to+\infty} P\{\mu \in I_\alpha\} = 1 - \alpha$$

Capítulo 14

Paseos al azar y Ecuaciones Diferenciales

14.1. Introducción

A lo largo del curso, hemos tratado de mostrar las relaciones que existen entre la teoría de probabilidades y las distintas ramas de la matemática, particularmente con el análisis con la que está estrechamente ligada. Continuando esta linea, en éste capítulo, exploraremos qué relación existe entre entre la teoría de probabilidades y la de ecuaciones diferenciales parciales. Esta conexión es de gran importancia para los desarrollos actuales en ambas áreas.

Una **ecuación diferencial** es una relación entre las derivadas de una función (que puede involucrar derivadas de distintos órdenes o con respecto a diferentes variables, e incluso a la misma función que es su derivada de orden cero). Las ecuaciones diferenciales se utilizan habitualmente en muchas aplicaciones de la matemática, particularmente en física, para modelar distintos fenómenos.

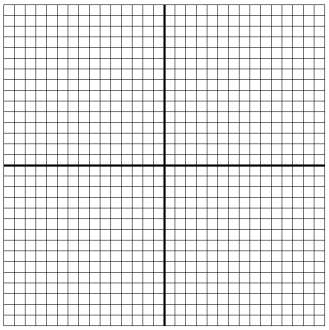
14.2. Un modelo sin tiempo: Paseos al azar y funciones armónicas

La Grilla

Consideramos una **grilla** o **retículo** en el plano

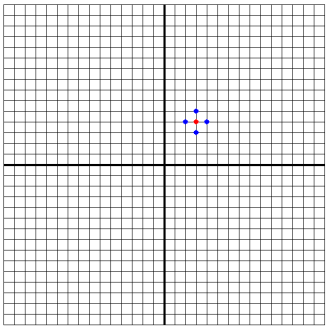
$$G = \{(ih, jh) : i, j \in \mathbb{Z}\}$$

donde $h > 0$ es un parámetro. A los puntos de la grilla los llamaermos **nodos**.



Nodos Vecinos

Dado un nodo (ih, jh) de la grilla, sus **vecinos** son los puntos $((i-1)h, jh), (i+1)h, jh), (ih, (j-1)h), (ih, (j+1)h)$.



Los vecinos del nodo rojo son los nodos azules.

Paseos al azar

Consideramos un bichito que efectua **paseo al azar sobre la grilla**. Trabajaremos con un tiempo discreto $t \in \mathbb{N}_0$. Empezamos en una posición inicial X_0. Llamamos X_t a la posición al tiempo t. Será un **vector aleatorio** con valores en G.

En cada tiempo, suponiendo que estamos en un nodo X_{t-1} elegimos con probabilidad $1/4$ uno de sus vecinos y nos movemos a él.

$$P\{X_{t+1} = q / X_t = p\} = 1/4$$

para todo nodo q vecino a p.

Notamos que este proceso define una **cadena de Markov** donde los posibles estados son los puntos de la grilla.

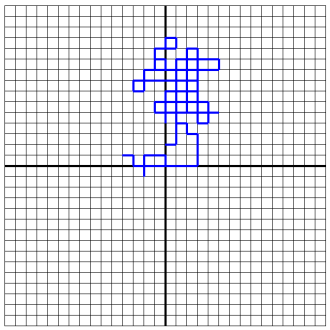
Tener una variable aleatoria X_t para cada tiempo t usualmente se denomina un **proceso estocástico**.

Trayectorias

Aunque nuestro proceso tiene tiempo discreto, podríamos convertirlo en un proceso con **tiempo continuo** $t \in \mathbb{R}_{\geq 0}$, especificando que cuando $n < t < n+1$ con $i \in \mathbb{N}_0$, nuestro bichito se mueve del nodo X_n al X_{n+1} en linea recta a velocidad uniforme.

$$X_n = X_n + (t-n)(X_{n+1} - X_n), \quad n < t < n+1$$

Ahora las trayectorias de nuestro proceso serán **curvas continuas** (poligonales).



Tiempo de salida de un dominio

Ahora consideramos un abierto acotado $U \subset \mathbb{R}^2$ con frontera suave (por ejemplo: un círculo).

Supongamos que nuestro proceso (X_t) comienza en un punto $X_0 \in \overline{U}$.

Notamos

$$\tau = \text{mín}\{t \in \mathbb{R}_{\geq 0} : X_t \notin U\}$$

al tiempo que nuestro proceso tarda en salir del dominio U (o $\tau = +\infty$ si nunca salimos). Se llama el **tiempo de parada** para nuestro proceso.

Como las trayectorias son continuas no podemos salir de U sin cruzar la **frontera** ∂U, es decir

$$X_\tau \in \partial U$$

¿Por dónde salimos?

Ahora consideramos una parte de la frontera $\Gamma \subset \partial U$. Por ejemplo

$$U = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$$

$$\partial U = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$$

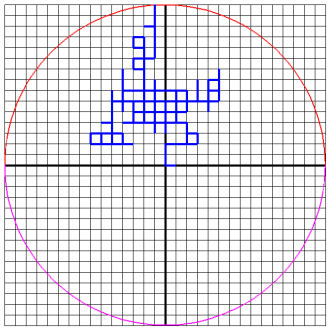
$$\Gamma = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1, y > 0\}$$

Nos preguntamos ¿cuál es la probabilidad de salir por Γ suponiendo que a tiempo $t_0 \in \mathbb{N}$ arrancamos en $x_0 \in G$?

Definimos una función $u_h : G \to [0, 1]$

$$u_h(x_0) = P\{X_\tau \in \Gamma / X_{t_0} = x_0\}$$

Notemos que en realidad esta probabilidad **no depende** del tiempo inicial t_0 por la **falta de memoria** del proceso.



¿Qué propiedad cumple u_h?

Si a tiempo $t_0 \in \mathbb{N}$ estamos en un nodo x_0, a tiempo $t_0 - 1$ tenemos que haber estado en alguno de sus vecinos (y hay un $1/4$ de probabilidad para cada uno).

$$P\{X_\tau \in \Gamma / X_{t_0} = x_0\} = \frac{1}{4} \sum_{v \sim x_0} P\{X_\tau \in \Gamma / X_{t_0 - 1} = v\}$$

donde notamos por $\sim$ la relación de ser vecinos. O sea:

$$u_h(x_0) = \frac{1}{4}\sum_{v\sim x_0} u_h(v)$$

Es decir que u_h verifica la **propiedad discreta del valor medio**. El valor de u_h en un nodo x_0 es el promedio de los valores de u_h en los nodos vecinos.

Las funciones que la cumplen se llaman **funciones armónicas discretas**.

¿Dónde aparecen las ecuaciones diferenciales?

Ahora si $u : \mathbb{R}^2 \to \mathbb{R}$ es una función de clase C^2, tenemos usando **el desarrollo de Taylor** que

$$\frac{1}{4}\sum_{v\sim x_0} u_h(v) = u(x_0) + \frac{1}{2}\Delta u(x_0)h^2 + o(h^2)$$

donde Δu es el **Laplaciano** de u.

$$\Delta u(x_0) = u_{xx}(x_0) + u_{yy}(x_0)$$

La ecuación para u_h puede escribirse

$$\frac{1}{2h^2}\left[u_h(x_0) - \frac{1}{4}\sum_{v\sim x_0} u_h(v)\right] = 0$$

Entonces, cuando $h \to 0$ es esperable que u_h converja a la solución del **problema de Dirichlet**

$$\begin{cases} \Delta u &= 0 & \text{en } U \\ u &= 1 & \text{en } \Gamma \\ u &= 0 & \text{en } \partial U - \Gamma \end{cases}$$

(Para dominios buenos, hay una solución única)

Solución para el círculo

Por ejemplo si U es el círculo como antes, la solución del problema de Dirichlet

$$\begin{cases} \Delta u &= 0 & \text{en } U \\ u &= f & \text{en } \partial U \end{cases}$$

viene dada por

$$u(r\cos\theta, r\,\text{sen}\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_r(\theta - t) f(cost, \text{sen}\, t)\,\mathrm{d}t, \quad 0 \le r < 1$$

donde el **núcleo de Poisson** es

$$P_r(\theta) = \sum_{n=-\infty}^{\infty} r^{|n|} e^{in\theta} = \frac{1-r^2}{1-2r\cos\theta + r^2} = \operatorname{Re}\left(\frac{1+re^{i\theta}}{1-re^{i\theta}}\right), \quad 0 \leq r < 1$$

Así que en nuestro ejemplo

$$u(r\cos\theta, r\ \operatorname{sen}\theta) = \frac{1}{2\pi}\int_0^{\pi} P_r(\theta - t)\,\mathrm{d}t, \quad 0 \leq r < 1$$

Estas fórmulas (que no hemos deducido) se ven en los cursos de ecuaciones diferenciales, y a veces en los de análisis complejo.

14.3. Un modelo con tiempo: La ecuación del calor o ecuación de difusión

Paseos al azar unidimensionales

Hasta aquíiconsideramos paseos al azar bidimensionales. Pero la misma idea puede considerarse en cualquier número de dimensiones.

Consideremos para simplificar el caso unidimensional. Consideramos la **grilla** dada por los múltiplos enteros de un parámetro $h > 0$, $G_h = h\mathbb{Z}$.

Vamos a considerar una partícula que se mueve por esta grilla, en ciertos tiempos $t_n = nk$ donde $k > 0$ es otro parámetro. Llamamos $T_k = k\mathbb{N}_0$ al conjunto de tiempos que vamos a considerar-

La partícula comienza en tiempo $t_0 = 0$ en una cierta posición $X_0 = x_0$ y después se mueve al azar según la regla.

$$X_{t_n} = \begin{cases} X_{t_{n-1}} + h & \text{con probabilidad } 1/2 \\ X_{t_{n-1}} - h & \text{con probabilidad } 1/2 \end{cases}$$

Podemos pensar que en cada tiempo discreto t_n tiramos una moneda y decidimos si ir para la izquierda o para la derecha una distancia h según el resultado de la moneda. Asumimos que las distintas tiradas de la moneda son independientes.

¿Cómo podríamos encontrar la distribución de X_{t_n} ? Podemos escribir

$$X_{t_n} = X_{t_{n-1}} + hU_n$$

donde la U_n son variables aleatorias independientes, con distribución de Rademacher

$$U_n = \begin{cases} 1 & \text{con probabilidad } 1/2 \\ -1 & \text{con probabilidad } 1/2 \end{cases}$$

Luego:

$$X_{t_n} = X_0 + h(U_1 + U_2 + \ldots + U_n)$$

Podemos escribir $U_n = 2V_n - 1$ donde $V_n \sim \text{Be}(1/2)$.

$$V_n = \begin{cases} 0 & \text{con probabilidad } 1/2 \\ 1 & \text{con probabilidad } 1/2 \end{cases}$$

Luego

$$X_{t_n} = x_0 + [2(V_1 + V_2 + \ldots + V_n)n]h = x_0 + [2S_n - n]h$$

donde S_n representa el número de éxitos en n ensayos de Bernoulli con probabilidad de éxito $1/2$, $S_n \sim \text{Bi}(1/2)$.

Podemos enconteces escibir una fórmula para la distribución de X_{t_0}.

$$p_h(x_m, t_n) = P\{X_{t_n} = x_m\} = b(d, n, 1/2) = \binom{n}{d}\left(\frac{1}{2}\right)^d\left(\frac{1}{2}\right)^{n-d} = \frac{1}{2^n}\binom{n}{d}$$

si $x_m = x_0 + [2d - n]h$ con $d \in \{0, 1, 2, \ldots, n\}$.

También podríamos obtener una **ecuación en diferencias** para u_h notando que si nuestra partícula está en la posición x_n en un tiempo t_n, en el tiempo t_{n-1} debe haber estado en las posiciones x_{m-1} o x_{m+1} con probabilidad $1/2$, dependiendo del valor de la variable aletoria de Rademacher U_n

Entonces si $x \in G$:

$$\begin{aligned} p_h(x_m, t_n) &= P\{X_{t_n} = x_m\} \\ &= P\{X_{t_n} = x_m / U_n = 1\} \cdot P\{U_n = 1\} \\ &+ P\{X_{t_n} = x_m / U_n = -1\} \cdot P\{U_n = -1\} \\ &= P\{X_{t_{n-1}} = x_{m-1}\} \cdot \frac{1}{2} + P\{X_{t_{n-1}} = x_{m+1}\} \cdot \frac{1}{2} \\ &= \frac{1}{2}\left(p_h(x_{m-1}, t_{n-1}) + p_h(x_{m+1}, t_{n-1})\right) \end{aligned}$$

La ecuación del calor o ecuación de difusión

Nos interesa entender el comportamiento asintótico de $p_h(x, x_0, t)$ cuando $h \to 0$. Esto va a depender de que relación exista entre el paso en el espacio h y el paso en el tiempo k.

Recordamos el desarrollo de Taylor:

Si $p : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ es una función C^2,

$$p(x, t + k) - p(x, t) = \frac{\partial p}{\partial t}(x, t)k + o(k)$$

$$p(x+h,t)+p(x-h,t)-2p(x,t)=\frac{\partial^2 p}{\partial^2 x}(x,t)h^2+o(h^2)$$

Vamos a asumir que $k=c\cdot h^2$ donde c es una constante, entonces:

$$\frac{p_h(x_m,t_n)-p_h(x,t_{n-1})}{k}=\frac{1}{2}\frac{p_h(x_m,t_{n-1})+p_h(x_m,t_{n-1})-2p_h(x_m,t_{n-1})}{ch^2}$$

Si suponemos que las densidades de probabilidad convergen

$$p_h(x,t)\to p(x,t)$$

obtenemos en el límite la ecuación diferencial en derivadas parciales

$$\frac{\partial p}{\partial t}(x,t)=\frac{1}{2c}\,\frac{\partial^2 p}{\partial^2 x}(x,t)$$

que se conoce como **ecuación del calor** o **ecuación de difusión**. Aunque esta ecuación fue formulada originalmente por J. Fourier para describir la propagación del calor en una barra de metal, se puede usar para describir muchos otros procesos de difusión (por ejemplo de la tinta en el agua [LLLT04]).

La solución fundamental de la ecuación del calor

Para cada $t>0$, la función $p(x,t)$ va a ser una densidad de probablidad, límite de las probabilidades $p_h(x_m,t_n)$ que dan la distribución discreta. Vab a depedender también del punto x_0 donde arranca nuestra partícula, así que las notaremos $p_h(x_m,x_0,t_n)$ y $p(x,x_0,t)$ para enfatizar esto.

Entonces si $f:\mathbb{R}\to\mathbb{R}$ es una función acotada:

$$u_h(x_0,t_n)=E[f(X_{t_n})/X_0=x_0]=\sum_{x_m}f(x_m)\cdot p_h(x_m,x_0,t_n)$$
$$\to u(x_0,t):=\int_{-\infty}^{\infty}f(x)\cdot p(x,x_0,t)\,dx$$

y u también va a satisfacer la ecuación del calor

$$\frac{\partial u}{\partial t}(x,t)=\frac{1}{2c}\,\frac{\partial^2 u}{\partial^2 x}(x,t)$$

con la **condición inicial**

$$u(x_0,0)=f(x_0)$$

Nota: Esto no es una justificación rigurosa, pero nos da una idea intuitiva de lo que esperamos que ocurra. Una justificación rigurosa puede darse usando la teoría de las soluciones viscosas (ver [Ros]).

Para encontrar explícitamente quien es p usamos el teorema local de De Moivre-Laplace (teorema 11.1.1). Nos acordamos de que

$$p_h(x_m, t_n) = P\{X_{t_n} = x_m\} = b(d, n, 1/2)$$

si $x_m = x_0 + [2d - n]h$ con $d \in \{0, 1, 2, \ldots, n\}$. Entonces

$$z_d = \frac{d - np}{\sqrt{npq}} = \frac{d - n/2}{\sqrt{n/4}} = \frac{2d - n}{\sqrt{n}} = \frac{\frac{x_m - x_0}{h}}{\sqrt{n}} = \frac{x_m - x_0}{\sqrt{n}h}$$

Luego como $t_n = kn = ch^2n$:

$$z_d^2 = \frac{(x_m - x_0)^2}{h^2 n} = \frac{c(x_m - x_0)^2}{t_n}$$

$$\frac{1}{\sqrt{2\pi npq}} = \frac{1}{\sqrt{2\pi n/4}} = \frac{1}{\sqrt{\pi t_n/(2k)}} = \frac{ch}{\sqrt{\pi t_n/2}}$$

Luego cuando $h \to 0$, obtenemos que:

$$\frac{p_h(x_m, t_n)}{h} \to p(x, x_0, t) := \frac{c}{\sqrt{\pi t}} \exp\left(\frac{-c(x - x_0)^2}{t/2}\right)$$

Entonces recapitulando, la solución general de la ecuación del calor con la condición inicial

$$u(x_0, 0) = f(x_0)$$

vendrá dada por:

$$u(x_0, t) := \int_{-\infty}^{\infty} f(x) \cdot p(x, x_0, t)\, dx$$

Esta fórmula se ve en los cursos de ecuaciones diferenciales.
Para profundizar en los temas de este capítulo, pueden consular [LL10] o [Law10].

Apéndice A

Repaso de Combinatoria

Los temas de este apéndice corresponden a álgebra I. Para más detalles recomiendo consultar el apunte de la profesora Krick (se incluyen referencias a dicho apunte en este apéndice).

A.1. Formalizando algunas cosas que sabemos desde la escuela primaria

¿Cómo podemos reconocer que dos conjuntos tienen la misma cantidad de elementos?

Definición A.1.1 *Decimos que dos conjuntos A y B son* ***coordinables*** *si existe una función biyectiva $f : A \to B$. Notación: $A \sim B$.*

Ejemplo: $A = \{1, 2, 3\}$ y $B = \{a, b, c\}$ son coordinables por medio de la función $f(1) = a$, $f(2) = b$ y $f(3) = c$.

Notamos que $\sim$ es una relación de equivalencia entre los conjuntos.

- Es **reflexiva**: pues $\mathrm{Id}_A : A \to A$ es biyectiva. Luego $A \sim A$.
- Es **simétrica** porque si $A \sim B \Rightarrow$ existe $f : A \to B$ es biyectiva. Pero entonces $f^{-1} : B \to A$ también lo es. Luego $B \sim A$.
- Es **transitiva**: Porque si $A \sim B$ y $B \sim C$ entonces existen $f : A \to B$ y $g : B \to C$ biyectivas. Pero entonces $g \circ f : A \to C$ es biyectiva [ya que $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$] Luego $A \sim C$.

Notamos por $\#(A)$ el **cardinal** o **cantidad de elementos** de un conjunto A. Formalmente esto podría definirse como sigue:

Definición A.1.2 *Para cada $n \in \mathbb{N}_0$, consideramos la **sección incial** de los números naturales*

$$I_0 = \emptyset, \quad I_n = \{1, 2, 3, \ldots, n\} = \{m \in \mathbb{N} : m \leq n\}$$

.

*Decimos que un conjunto A es **finito** si es coordinable con alguna sección inicial de los números naturales I_n. En este caso decimos que A tiene n elementos y escribimos*

$$\#(A) = n$$

Notamos que $\emptyset$ es finito y $\#(\emptyset) = 0$.

*En caso contrario, decimos que A es **infinito**.*

- Esta definición es correcta porque $I_n \sim I_m \Leftrightarrow n = m$.
- Si $A \sim B$ y A es finito, entonces B es finito y $\#(A) = \#(B)$

Formalizaremos ahora, algunas cosas que uno aprende en la escuela primaria:

Teorema A.1.3 (Número de elementos en una unión de conjuntos) *Si A y B son finitos, $A \cup B$ es finito y*

$$\#(A \cup B) = \#(A) + \#(B) - \#(A \cap B)$$

En particular si A y B son disjuntos ($A \cap B = \emptyset$),

$$\#(A \cup B) = \#(A) + \#(B)$$

Ejemplo: $A = \{1, 3, 4, 5\}$, $B = \{5, 6, 7\}$, $A \cup B = \{1, 3, 4, 5, 6, 7\}$, $A \cap B = \{5\}$ entonces $\#(A) = 4, \#(B) = 3$, $\#(A \cup B) = 7$,$\#(A \cap B) = 1$.

Teorema A.1.4 (Número de elementos en una diferencia de conjuntos) *Si B es finito y $A \subseteq B$, entonces A es finito y $\#(A) \leq \#(B)$.*

$$\#(B - A) = \#(B) - \#(A)$$

Teorema A.1.5 (Número de elementos de un producto cartesiano) *Si A y B son finitos, $A \times B$ es finito y*

$$\#(A \times B) = \#(A) \cdot \#(B)$$

En particular si

$$A^n = A \times A \times ... \times A \ (n \text{ veces}) = \{(a_1, a_2, \ldots, a_n) : a_i \in A\}$$

entonces

$$\#(A^n) = \#(A)^n$$

Ejemplo: $A = \{1, 2, 3\},\ B = \{a, b\}, \#(A) = 2, \#(B) = 3$

$$A \times B = \{(1, a), (1, b), (2, a), (2, b), (3, a), (3, c)\}, \#(A \times B) = 6$$

A.2. Usando estas ideas para contar algunos objetos matemáticos

A.2.1. ¿Cuántas funciones hay de A en B?

Teorema A.2.1 (Cantidad total de funcione entre dos conjuntos) *Si A y B son finitos,*

$$\#\{\textit{funciones } f : A \to B\} = \#(B)^{\#(A)}$$

Demostración: Sean $n = \#(A), m = \#(B)$ y escribamos

$$A = \{a_1, a_2, \ldots, a_n\}$$

A cada función $f : A \to B$ le podemos asociar la n-upla de elementos de B

$$(f(a_1), f(a_2), \ldots, f(a_n))$$

y recíprocamente cada una de estas n-úplas determina una función de A en B. Es una **correspondencia biyectiva**. Es decir que el conjunto que estamos tratando de contar **es coordinable** con B^n. En consecuencia, tiene m^n elementos.

A.2.2. ¿Cuántas partes tiene un conjunto?

Teorema A.2.2 (Cantidad total de funcione entre dos conjuntos) *Dado un conjunto A, consideramos su conjunto de partes*

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

Entonces si A es finito, $\mathcal{P}(A)$ es finito y

$$\#(\mathcal{P}(A)) = 2^{\#(A)}$$

Demostración: Sea $n = \#(A)$ y escribamos

$$A = \{a_1, a_2, \ldots, a_n\}$$

Consideremos $\mathcal{T} = \{V, F\}$ un conjunto de 2 elementos. A cada subconjunto $B \subseteq A$ le podemos asignar la la n-upla de elementos de $\mathcal{T}$

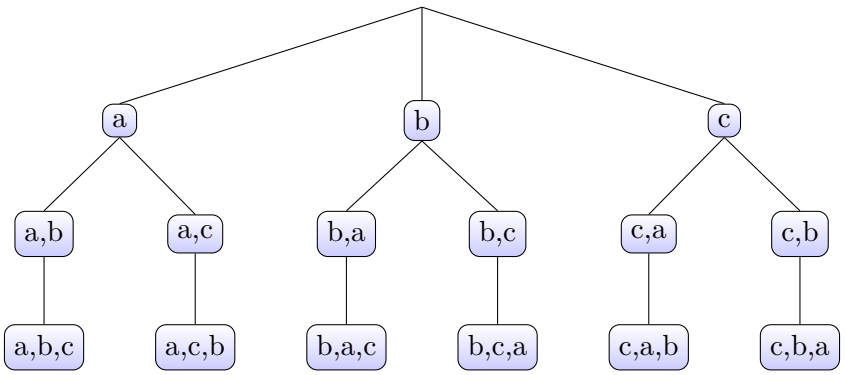
$$(t_1, t_2, \ldots, t_n) \text{ dada por } t_k = \begin{cases} V & \text{si} \quad a_k \in B \\ F & \text{si} \quad a_k \notin B \end{cases}$$

Entonces hay una **biyección** entre $\mathcal{P}(A)$ y $\mathcal{T}^n$ con lo que $\#\mathcal{P}(A) = \#(\mathcal{T})^n = 2^n$.

A.3. Permutaciones

A.3.1. Permutaciones de 3 elementos

¿De cuántas formas podemos ordenar 3 personas $a = Aldo$, $b = Blanca$, $c = Carlos$ en una fila?



Permutaciones de 3 elementos $\{a, b, c\} = 3 \times 2 \times 1 = 3! = 6$

A.3.2. Otra manera de pensar las permutaciones de 3 elementos

Notemos que cada manera de ordenar las personas como b, c, a puede pensarse como una **función biyectiva** del conjunto $I_3 = \{1, 2, 3\}$ en $\{a, b, c\}$ (que dice qué persona pusimos en cada lugar de la fila)

$$f(1) = b$$

$$f(2) = c$$

$$f(3) = a$$

Luego 3! también es el número de **funciones biyectivas** de un conjunto de 3 elementos en otro de 3 elementos.

Notamos que esta cantidad de funciones no depende de la naturaleza de los objetos que estemos considerando.

A.3.3. Permutaciones en general

Definición A.3.1 *Sea $n \in \mathbb{N}_0$. El número de permutaciones P_n de n objetos es el número de funciones biyectivas $f : A \to B$ cuando A y B son dos conjuntos cualesquiera con n elementos. Por ejemplo, podemos tomar $A = B$ o incluso $A = B = I_n$.*

Teorema A.3.2 *Si* $n \in \mathbb{N}$,

$$P_n = n! = 1 \cdot 2 \cdot 3 \cdots n = \prod_{k=1}^{n} k$$

Por ejemplo, ¿de cuántas maneras pueden ser ordenadas 5 personas en el orden de mérito de un concurso? (suponiendo que nadie queda afuera del concurso)

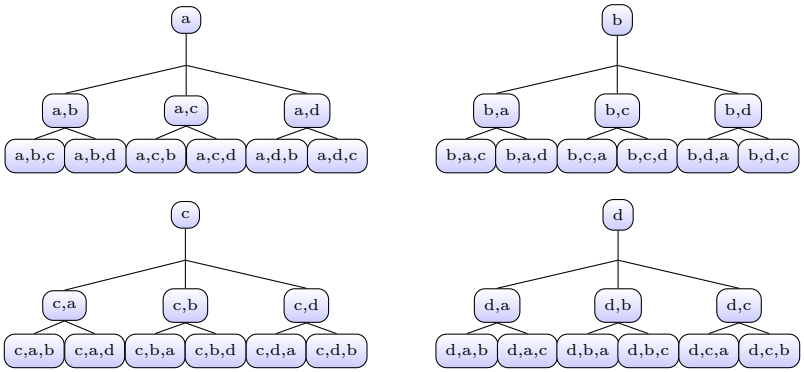
Rta: $5! = 120$.

Notar que si $n = 0$, $A = B = \emptyset$ y hay **una única** función $f : \emptyset \to \emptyset$ (la función vacía). Entonces $P_0 = 1$ por lo que la definición **0!=1** hace que el teorema sea cierto también en este caso.

A.4. Variaciones

A.4.1. Una variación del problema anterior

Supongamos ahora que tenemos 4 personas en un concurso $a = Aldo$, $b = Blanca$, $c = Carlos$ $d = Diana$ y tenemos que elegir una terna donde **importa el orden en que los ponemos**. ¿De cuántas formas diferentes podemos hacerlo?



Esto se conoce como el número de **variaciones (sin repetición)** de un conjunto de 4 elementos tomandos en tuplas de 3 elementos

$$V_3^4 = 4 \cdot 3 \cdot 2 = 24$$

A.4.2. Otra manera de pensar las variaciones

Notemos que cada posible terna de personas como b, c, a puede pensarse ahora como una **función inyectiva** del conjunto $I_3 = \{1, 2, 3\}$ en $\{a, b, c, d\}$ (que dice qué persona pusimos en cada lugar de la fila)

$$f(1) = b$$

$$f(2) = c$$
$$f(3) = a$$

A.4.3. Variaciones en general

En general, el número de **variaciones** V_k^n cuenta de cuántas maneras podemos elegir k objetos de un conjunto de n objetos **donde importa en qué orden los tomamos** (aquí $k \leq n$).

Formalmente, esto puede expresarse diciendo que V_k^n es el número de **funciones inyectivas** $f : A \to B$ donde A tiene k elementos y B tiene n elementos. [ver **proposición 3.2.2 del apunte**].

Generalizando el razonamiento anterior vemos que está dado por:

$$V_k^n = n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \prod_{j=1}^{k} (n-j+1)$$

También podemos escribirlo como

$$V_k^n = \frac{n \cdot (n-1) \cdot (n-2) \ldots (n-k+1)(n-k)(n-k-1) \cdots 1}{(n-k) \cdot (n-k-1) \cdots 1} = \frac{n!}{(n-k)!}$$

A.5. Combinaciones: ¿Y si no tenemos en cuenta el orden?

Volvamos al problema anterior, donde teníamos cuatro personas a, b, c, d y queríamos escoger una terna. Pero supongamos que ahora **no importa el orden** en que las elegimos. ¿Cuántas elecciones posibles tenemos?

Una manera de pensarlo es la siguiente: En el conjunto de ternas que obtuvimos antes, definimos **una relación de equivalencia** diciendo que dos ternas son equivlentes si una se obtiene de la otra permutando los elementos, Sabemos que esta relación va a partir el conjunto de ternas en **clases de equivalencia**. Por ejemplo la clase de equivalencia de la terna a, b, c está formada por las ternas

$$a, b, c \quad a, c, b \quad b, a, c \quad b, c, a \quad c, b, a \quad c, a, b$$

Cualquiera de ellas representa la misma elección de pesonas si no se tiene en cuenta el orden. Notemos que cada clase tiene $3! = 6$ elementos.

Como hay 24 ternas, y cada clase de equivalencia tiene 6 ternas tendremos en total

$$C_3^4 = \binom{4}{3} = \frac{V_3^4}{3!} = \frac{24}{6} = 4$$

En general, podemos considerar el número de combinaciones $C_k^n = \binom{n}{k}$ que cuenta cuántas elecciones podemos hacer de k elementos a partir de un conjunto de n elementos **sin tener en cuenta el orden**. O expresado en otras palabras: ¿cuántos subconjuntos de k elementos podemos obtener a partir de uno de n?.

$$\mathcal{P}_k(A) = \{B \subseteq A : \#(B) = k\} \Rightarrow \binom{n}{k} = \#\mathcal{P}_k(A) \text{ con } n = \#(A)$$

Generalizando el razonamiento anterior, se ve que:

$$\binom{n}{k} = \frac{V_k^n}{k!} = \frac{n!}{k!(n-k)!}$$

[Fórmula de la proposición 3.2.2 del apunte, aunque allí se prueba de otra forma.]

También se lo conoce como **número combinatorio**.

Teorema A.5.1 (Definición recursiva de los números combinatorios)

$$\binom{n}{0} = \binom{n}{n} = 1$$

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \qquad 1 \leq k \leq n-1$$

En el apunte se prueba esto a partir de la **interpretación combinatoria** de $\binom{n}{k}$ [proposición 3.3.3] y se deduce entonces la fórmula

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

por inducción [teorema 3.3.4]. También se puede seguir el camino inverso: deducir primero esta fórmula de la interpretación combinatoria (como hicimos) y a partir de ella la recurrencia haciendo cuentitas.

A.5.1. El Triangulo de Pascal

Ordenemos los números combinatorios de la siguiente forma

$$\binom{0}{0}$$

$$\binom{1}{0} \quad \binom{1}{1}$$

$$\binom{2}{0} \quad \binom{2}{1} \quad \binom{2}{2}$$

$$\binom{3}{0} \quad \binom{3}{1} \quad \binom{3}{2} \quad \binom{3}{3}$$

$$\binom{4}{0} \quad \binom{4}{1} \quad \binom{4}{2} \quad \binom{4}{3} \quad \binom{4}{4}$$

$$\binom{5}{0} \quad \binom{5}{1} \quad \binom{5}{2} \quad \binom{5}{3} \quad \binom{5}{4} \quad \binom{5}{5}$$

$$\binom{6}{0} \quad \binom{6}{1} \quad \binom{6}{2} \quad \binom{6}{3} \quad \binom{6}{4} \quad \binom{6}{5} \quad \binom{6}{6}$$

$$\binom{7}{0} \quad \binom{7}{1} \quad \binom{7}{2} \quad \binom{7}{3} \quad \binom{7}{4} \quad \binom{7}{5} \quad \binom{7}{6} \quad \binom{7}{7}$$

Obtenemos

$$1$$

$$1 \quad 1$$

$$1 \quad 2 \quad 1$$

$$1 \quad 3 \quad 3 \quad 1$$

$$1 \quad 4 \quad 6 \quad 4 \quad 1$$

$$1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1$$

$$1 \quad 6 \quad 15 \quad 20 \quad 15 \quad 6 \quad 1$$

$$1 \quad 7 \quad 21 \quad 35 \quad 35 \quad 21 \quad 7 \quad 1$$

A.5.2. Números combinatorios complementarios

Teorema A.5.2 (Números combinatorios complementarios)

$$\binom{n}{k} = \binom{n}{n-k} \quad 0 \leq k \leq n$$

Hay dos maneras de pensarla:

- **Interpretación combinatoria:** Hay tantos subconjuntos de $n-k$ elementos en uno de n elementos, como suboconjuntos de k elementos.

  Si A tiene n elementos, y $B \subseteq A$, entonces B tiene k elementos sí y sólo si $A - B$ tiene $n - k$. Esto establece una **biyeccción** entre $\mathcal{P}_k(A)$ y $\mathcal{P}_{n-k}(A)$

- También es inmediata a partir de la fórmula

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad 0 \leq k \leq n$$

A.5.3. Suma de todos los combinatorios para un n fijo

Teorema A.5.3 (Suma de todos los combinatorios para un n fijo)

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

- **Interpretación combinatoria:** La suma cuenta cuándos subconjuntos se pueden formar con un conjunto de n elementos ya que

$$\mathcal{P}(A) = \bigcup_{k=0}^{n} \mathcal{P}_k(A) \textbf{ unión disjunta}$$

A.5.4. Teorema del Binomio

Teorema A.5.4 *Sean $x, y \in \mathbb{C}$, $n \in \mathbb{N}_0$ entonces:*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Apéndice B

Cadenas de Markov

En este apéndice demostraremos dos resultados fundamentales sobre las cadenas de Markov.

En el espacio $\mathbb{R}^N$ usamos la norma

$$\|x\| = |x_1| + |x_2| + \ldots + |x_N|$$

Notamos

$$\mathcal{P}_N = \{x \in \mathbb{R}^N : x_k \geq 0, x_1 + x_2 + \ldots + x_N = 1\}$$

al conjunto de vectores de probabilidad N-dimensionales.

Teorema B.0.1 *Si $P \in \mathbb{R}^{N\times N}$ es una matriz estocástica, entonces la transformación lineal asociada a P aplica $\mathcal{P}_N$ en sí mismo, y tiene un punto fijo*

Prueba: Dado cualquier $U_0 \in \mathcal{P}_N$ considero los promedios

$$S_n = \frac{1}{n}\sum_{j=1}^{n} P^j U_0$$

$S_n \in \mathcal{P}_N$ como $\mathcal{P}_N$ es compacto, existe una subsucesión convergente S_{n_k}. Digamos que $S_{n_k} \to U_\infty$.

$$S_{n_k} = \frac{1}{n_k}\sum_{j=1}^{n_k} P^j U_0$$

$$\begin{aligned} PS_{n_k} &= \frac{1}{n_k}\sum_{j=1}^{n_k} P^{j+1}U_0 \\ &= \frac{1}{n_k}\sum_{j=2}^{n_k+1} P^{j}U_0 \\ &= \frac{1}{n_k}\sum_{j=1}^{n_k} P^{j}U_0 - PU_0 + P^{n_j+1}U_0 \\ &= \frac{1}{n_k}\sum_{j=1}^{n_k} P^{j}U_0 - PU_0 + P^{n_j+1}U_0 \end{aligned}$$

pero

$$\lim_{k\to\infty} \frac{1}{n_k}\| - PU_0 + P^{n_j+1}U_0\| = 0$$

y en el límite:

$$PU_\infty = U_\infty$$

o sea que U_∞ es una distribución estacionaria. □

También es posible probar este teorema como consecuencia directa del teorema de punto fijo de Brower, ya que se puede ver que $\mathcal{P}_N$ es homeomorfo a una bola cerrada de dimensión $N-1$.

Teorema B.0.2 *Si $p_{i,j} > 0$ para todo i, j, es una contracción en la métrica que definimos antes. Deducimos que tiene un único punto fijo $U_\infty \in \mathcal{P}_N$ y que para todo $U_0 \in V^n$*

$$\lim_{n\to+\infty} P^n U_0 = U_\infty$$

Prueba: Podemos elegir $\varepsilon > 0$ tal que $p_{i,j} > \varepsilon$ para todo i, j. Entonces como

$$\sum_{i=1}^{N} p_{ij} = 1$$

deducimos que $\varepsilon N \leq 1$.Achicando el ε podemos conseguir que $\varepsilon N < 1$

Escribimos $P = \alpha Q + \varepsilon J$ donde $\alpha = 1 - N\varepsilon$ y J es la matriz de $N \times N$ cuyas entradas son todos 1, o sea que si $Q = (q_i j)$

$$q_{i,j} = \frac{p_{i,j} - \varepsilon}{\alpha}$$

Afirmamos que Q es una matriz estocástica: es claro por nuestra elección de $\varepsilon > 0$ que $q_{i,j} > 0$. Calculemos la suma de las columnas de j

$$\sum_{i=1}^{N} q_{i,j} = \frac{1}{\alpha}\left[\sum_{i=1}^{N} p_{i,j}\right] - N\frac{\varepsilon}{\alpha} = \frac{1}{\alpha} - N\frac{\varepsilon}{\alpha} = \frac{\alpha}{\alpha} = 1$$

Sean $x, y \in \mathcal{P}_N$, entonces

$$\begin{aligned}(Px)_i - (Py)_i &= \sum_{j=1}^{N} p_{ij}(x_j - y_j) \\ &= \alpha\sum_{j=1}^{N} q_{ij}(x_j - y_j) + \varepsilon\left[\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i\right] \qquad = \alpha\sum_{j=1}^{N} q_{ij}(x_j - y_j)\end{aligned}$$

Luego

$$|x_i - y_j| \leq \alpha\sum_{j=1}^{N} q_{ij}|x_j - y_j|$$

$$\begin{aligned}\|Px - Py\|_1 = \sum_{i=1}^{n} |(Px)_i - (Py)_j| &\leq \alpha\sum_{i=1}^{n}\sum_{j=1}^{N} q_{ij}|x_j - y_j| \\ &\leq \alpha\sum_{j=1}^{N}\left[\sum_{i=1}^{n} q_{i,j}\right]|x_j - y_j| \\ &\leq \alpha\sum_{j=1}^{N}|x_j - y_j| = \alpha\|x - y\|_1\end{aligned}$$

pues las columnas de Q suman 1. □

Apéndice C

La Fórmula de Stirling

En muchas cuestiones del cálculo de probabilidades, resulta necesario disponer de una aproximación de $n!$ para n grande. Este es el contenido de la Fórmula de Stirling:

Teorema C.0.1 (Fórmula de Stirling)

$$n! \sim \sqrt{2\pi}\, n^{n+1/2} e^{-n}$$

Con más presición, se tienen las desigualdades:

$$\sqrt{2\pi}\, n^{n+1/2} e^{n} < n! < \sqrt{2\pi} e^{-n} \left(1 + \frac{1}{4n}\right)$$

C.1. La fórmula de Wallis para π

La siguiente notable fórmula expresa a π como un producto infinito. La utilizaremos para determinar la constante que aparece en la fórmula de Stirling:

Teorema C.1.1 (Producto infinito de Wallis para π)

$$\frac{\pi}{2} = \lim_{m \to +\infty} \left[\frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2m}{2m-1} \cdot \frac{2m}{2m+1} \right]$$

o en forma de producto infinito

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2m}{2m-1} \cdot \frac{2m}{2m+1} \cdots$$

Para demostrar esta fórmula, introduzcamos la cantidad

$$I_n = \int_0^{\pi/2} \operatorname{sen}^n x \, dx$$

Lema C.1.2 *Se verifica la relación de recurrencia:*

$$I_n = \frac{n-1}{n} I_{n-2} \ (n \geq 2)$$

Prueba: Integrando por partes:

$$I_n = \int_0^{\pi/2} \text{sen}^{n-1}\ x\ (-\cos x)' dx = \ -\text{sen}^{n-1} x\ \cos x\big|_0^{\frac{\pi}{2}} - \int_0^{\frac{\pi}{2}} \left(\text{sen}^{n-1}\ x\right)'\ (-\cos x)\ dx$$

Es decir:

$$I_n = \int_0^{\pi/2} (n-1)\text{sen}^{n-2}\ \cos^2\ dx = \int_0^{\pi/2} (n-1)\text{sen}^{n-2}(1-\cos^2 x)\ dx = (n-1)[I_{n-2} - I_n]$$

En consecuencia: $nI_n = (n-1)I_{n-2}$, o sea:

$$I_n = \frac{n-1}{n} I_{n-2}$$

□

Prueba de la fórmula de Wallis:
A fin, de calcular I_n observamos que

$$I_0 = \int_0^{\pi/2}\ dx = \frac{\pi}{2}$$

$$I_1 = \int_0^{\pi/2}\ dx = 1$$

En consecuencia, podemos calcular los valores de I_n para n par o impar, respectivamente:

$$I_{2m} = \frac{2m-1}{2m} \cdot \frac{2m-3}{2m-2} \cdots \frac{5}{6} \cdot \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{\pi}{2}$$

$$I_{2n+1} = \frac{2m}{2m+1} \cdot \frac{2m-2}{2m-1} \cdots \frac{8}{9} \cdot \frac{6}{7} \cdot \frac{4}{5} \cdot \frac{2}{3}$$

Podemos despejar $\pi/2$:

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{4}{3} \frac{6}{5} \cdot \frac{5}{7} \cdots \frac{2m}{2m-1} I_{2m}$$

y utilizando la expresión de I_{2m+1}

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2m}{2m-1} \cdot \frac{2m}{2m+1} \cdot \frac{I_{2m}}{I_{2m+1}}$$

Queremos estimar el cociente $\frac{I_{2m}}{I_{2m+1}}$: para ello observams que en el intervalo $0 < x < \frac{pi}{2}$ se tiene $0 < \text{sen } x < 1$, en consecuencia $0 < \text{sen}^{2m+1} x < \sin^{2m-1}$ e integrando resulta que:

$$0 \leq I_{2m+1} \leq I_{2m} \leq I_{2m-1}$$

luego

$$1 \leq \frac{I_{2m}}{I_{2m-1}} = \frac{2m+1}{2m} \cdot \frac{I_m}{I_{2m-1}} \leq \frac{2m+1}{2m} = 1 + \frac{1}{2m}$$

Por la propiedad del sandwich deducimos que $\frac{I_{2m}}{I_{2m+1}}$ tiende a 1 cuando $m \to +\infty$. En consecuencia:

$$\frac{\pi}{2} = \lim_{m\to+\infty} \left[\frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2m}{2m-1} \cdot \frac{2m}{2m+1} \cdot \frac{I_{2m}}{I_{2m+1}}\right]$$

Esto completa la demostración de la fórmula de Wallis.

C.1.1. Otra fórmula de la fñormula de Wallis

Podemos escribir el resultado anterior en la forma:

$$\frac{\pi}{2} = \lim_{m\to+\infty} \frac{2^2 \cdot 4^2 \cdot 6^2 \cdots (2m)^2}{3^2 \cdot 5^2 \cdot 7^2 \cdots (2m-1)^2(2m+1)}$$

Como $\lim_{m\to+\infty} \frac{2m+1}{2m} = 1$ obtenemos (producto de límites):

$$\frac{\pi}{2} = \lim_{m\to+\infty} \frac{2^2 \cdot 4^2 \cdot 6^2 \cdots (2m-2)^2}{3^2 \cdot 5^2 \cdot 7^2 \cdots (2m-1)^2} \cdot 2m$$

Tomando raíz cuadrada:

$$\sqrt{\frac{\pi}{2}} = \lim_{m\to+\infty} \frac{2 \cdot 4 \cdot 6 \cdots (2m-2)}{3 \cdot 5 \cdot 7 \cdots (2m-1)} \cdot \sqrt{2m}$$

Multiplicando el denominador y el denominador por $2 \cdot 4 \cdot 6 \ldots \cdot (2m-2)$ resulta:

$$\begin{aligned}
\sqrt{\frac{\pi}{2}} &= \lim_{m\to+\infty} \frac{2^2 \cdot 4^2 \cdot 6^2 \cdots (2m-2)^2}{2 \cdot 3 \cdot 5 \cdot 6 \cdot 7 \cdots (2m-1)} \cdot \sqrt{2m} \\
&= \lim_{m\to+\infty} \frac{2^2 \cdot 4^2 \cdot 6^2 \cdots (2m)^2}{(2m)!} \cdot \frac{\sqrt{2m}}{2m} \\
&= \lim_{m\to+\infty} \frac{2^{2m}(1^2 \cdot 2^2 \cdot 3^2 \cdots m^2}{(2m)!\sqrt{2m}} \\
&= \lim_{m\to+\infty} \frac{2^{2m}(m!)^2}{(2m)!\sqrt{2m}}
\end{aligned}$$

Multiplicando ambos miembros por $\sqrt{2}$, resulta:

Teorema C.1.3 (Otra forma de la fórmula de Wallis)

$$\sqrt{\pi} = \lim_{m \to +\infty} \frac{2^{2m}(m!)^2}{(2m)!\sqrt{m}}$$

C.2. Prueba de la fórmula de Stirling

La prueba de la fórmula de Stirling, se basa en la siguiente idea: tenemos que

$$\log(n!) = \sum_{k=1}^{n} \log(k) \tag{C.1}$$

Cuando n es grande, es razonable que esperar que el valor de $\log(n!)$ esté próximo del valor de la siguiente integral, que representa el área bajo la curva $y = \log x$ (en el intervalo $1 \leq x \leq n$) y que podemos calcular exactamente:

$$A_n = \int_1^n \log x \, dx = n \log n - n + 1$$

La suma en (C.1) representa una aproximación a esta integral por medio de rectángulos (sumas de Riemman). Una aproximación mejor se consigue utilizando la aproximación por medio de trapecios:

$$T_n = \sum_{k=1}^{n-1} \frac{\log(k) + \log(k+1)}{2} = \sum_{k=1}^{n-1} \log(k) + \frac{1}{2}\log n = \log(n!) - \frac{1}{2}\log n$$

Como la función $f(x) = \log x$ es cóncava, la secante a la curva $y = f(x)$ que une los puntos $(k, \log(k))$ y $(k+1, \log(k+1))$ queda por abajo de dicha curva. En consecuencia,

$$A_n \geq T_n$$

Nuestro objetivo es estimar el error $E_n = A_n - T_n$. Notamos que:

$$E_{k+1} - E_k = \int_k^{k+1} \log x \, dx - \frac{\log(k) + \log(k+1)}{2}$$

representa el área que queda entre la recta secante y la curva en el intervalo $[k, k+1]$. Como la función es cóncava, $E_{k+1} - E_k \geq 0$. Por otro lado el área entre la curva la secante podemos acotarla por el área entre la tangente a la curva en $x = k + 1/2$, es decir la recta:

$$y = T(x) = \log(k + 1/2) + \frac{1}{k + 1/2}(x - (k + 1/2))$$

y la secante (pues siendo f cóncava, tenemos que $f(x) \leq T(x)$). Deducimos que:

$$E_{k+1} - E_k \leq \int_k^{k+1} T(x)\, dx - \frac{\log(k) + \log(k+1)}{2}$$

es decir:

$$E_{k+1} - E_k \leq \log(k+1/2) - \frac{\log(k) + \log(k+1)}{2}$$
$$= \frac{1}{2}\left(1 + \frac{1}{2k}\right) - \frac{1}{2}\left(1 + \frac{1}{2(k+1/2)}\right) < \frac{1}{2}\left(1 + \frac{1}{2k}\right) - \left(1 + \frac{1}{2(k+1)}\right)$$

Sumando estas igualdades para $k = 1, 2, \ldots, n-1$, todos los términos del lado derecho se cancelan, excepto dos (serie telescópica), y como E_0, obtenemos que:

$$E_n < \frac{1}{2}\log\frac{3}{2} - \frac{1}{2}\log\left(1 + \frac{1}{2n}\right) < \frac{1}{2}\log\frac{3}{2}$$

Notamos que E_n es entonces, monótona creciente y acotada, por lo tanto E_n tiende a un límite E cuando $n \to +\infty$. Y la desigualdad para $E_{k+1} - E_k$ permite estimar la diferencia $E - E_n$:

$$E - E_n \leq \sum_{k=n}^{\infty}(E_{k+1} - E_k) < \frac{1}{2}\left(1 + \frac{1}{2n}\right)$$

Entonces como $A_n = T_n + E_n$, obtenemos que:

$$\log(n!) = (n + 1/2)\log(n) - n + 1 - E_n$$

o escribiendo $\alpha_n = e^{1-E_n}$, y tomando exponencial:

$$n! = \alpha_n n^{n+1/2} e^{-n}$$

La sucesión α_n es ahora monótona decreciente, y tiende al límite: $\alpha = e^{1-E}$. En consecuencia, por las estimaciones anteriores:

$$1 \leq \frac{\alpha_n}{\alpha} = e^{E-E_n} < e^{(1/2)\log(1+1/2n)} = \sqrt{1 + \frac{1}{2n}} \leq 1 + \frac{1}{2n}$$

En consecuencia, tenemos las desigualdades:

$$\alpha n^{n+1/2} e^{-n} \leq n! \leq \alpha\left(1 + \frac{1}{2n}\right) n^{n+1/2} e^{-n}$$

Nos queda determinar el valor de la constante α. Para ello utilizamos la fórmula de Wallis,

$$\sqrt{\pi} = \lim_{m\to+\infty} \frac{2^{2m}(m!)^2}{(2m)!\sqrt{m}} = \lim_{n\to+\infty} \frac{\alpha_n^2}{\alpha_{2n}\sqrt{2}} = \frac{\alpha^2}{\alpha\sqrt{2}}$$

por lo que deducimos que $\alpha = \sqrt{2\pi}$.

Apéndice D

Construcción de la Integral de Lebesgue, y equivalencia de las distintas definiciones de esperanza

Motivación

En este apéndice presentaremos una construcción de la integral de Lebesgue, que es una herramienta útil para definir esperanzas de variables aleatorias y operar con ellas (Se desarrolla en los cursos de análisis real, pero aquí presentaremos algunas nociones básicas, siempre teniendo en mente la interpretación probabilística).

Para ver porqué la integral de Stieltjes no es adecuada para muchos propósitos teóricos, consideremos la definición que hemos dado anteriormente de la esperanza de una variable aleatoria X en términos de una integral de Stieltjes:

$$E[X] = \int_{-\infty}^{+\infty} x \, dF(x)$$

siendo $F = F_X$ su función de distribución. Esta definición es muy útil desde el punto de vista del cálculo, ya que no necesitamos conocer cuál es el espacio muestral o cuál es la función P que asigna las probabilidades. Toda la información relevante sobre la variable X está contenida en su función de distribución F_X.

Sin embargo, por ejemplo resulta complicado por ejemplo, con esta definición probar que la esperanza es lineal, ya que F_X no depende linealmente de X.

Otro ejemplo es el siguiente (tomado del libro de Barry James): Si usamos la integral de Stieltjes, entonces la fórmula:

$$E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) \, dF(x)$$

puede no tener sentido si φ tiene un punto de discontinuidad en común con F. Esa es la razón por la que si utilizamos la integral de Stieltjes, debemos restringir φ a ser una función continua, y entonces por ejemplo φ no puede ser el indicador de un evento.

Por el contrario, la teoría de la integral de Lebesgue permite probar los teoremas sobre la esperanza de variables aleatorias con toda generalidad, y en forma sencilla y elegante.

Uno de los propósitos fundamentales de este apéndice es presentar una prueba de dos teoremas centrales de la teoría de Lebesgue: el teorema de convergencia monótona y el teorema de convergencia mayorada, que forman parte del programa de la asignatura Probabilidad y Estadística (para matemáticos).

Así mismo, probaremos que la definición de esperanza en términos de la integral de Stieltjes es equivalente a la que utiliza la integral de Lebesgue.

D.1. Funciones Medibles

Consideramos un conjunto Ω y una σ-álgebra $\mathcal{M}$ de subconjuntos de Ω. Al par $(\Omega, \mathcal{M})$ lo llamamos *espacio medible*. A los cojuntos de $\mathcal{M}$ los llamaremos *conjuntos medibles* (representará la clase de aquellos conjuntos a los que asignaremos medida o probabilidad).

En la interpretación probabilística, Ω es el *espacio muestral* (conjunto de posibles resultados de un experimento aleatorio) y $\mathcal{M}$ será la σ-álgebra $\mathcal{E}$ de los eventos (aquellas partes de Ω a las que les asignaremos probabilidad).

Las funciones con las que vamos a trabajar deberán satisfacer una condición técnica, a saber que podamos medir ciertos conjuntos asociados a la función.

Definición D.1.1 *Sea $(\Omega, \mathcal{M})$ un espacio medible y sea $f : \Omega \to \overline{\mathbb{R}}$ una función. Diremos que f es una* **función medible** *(respecto a la σ-álgebra $\mathcal{M}$) si para todo $\alpha \in \mathbb{R}$ el conjunto $\{f > \alpha\} = \{\omega \in \Omega : f(\omega) > \alpha\}$ es medible, es decir pertenece a $\mathcal{M}$.*

Si $(\Omega, \mathcal{E}, P)$ es un espacio de probabilidad, las funciones medibles sobre Ω (respecto a la σ-álgebra P) son precisamente las *variables aleatorias* definidas sobre Ω.

La noción de función medible puede formularse de varias maneras equivalentes. (En lo sucesivo, usaremos las notaciones abreviadas $\{f < \alpha\} = \{\omega \in \Omega : f(\omega) < \alpha\}$, etcétera).

Lema D.1.2 *Sea $f : \Omega \to \overline{\mathbb{R}}$ una función. Son equivalentes:*

i) f es medible.

ii) Para todo $\alpha \in \mathbb{R}$, $\{f \geq \alpha\}$ es medible.

iii) Para todo $\alpha \in \mathbb{R}$, $\{f < \alpha\}$ es medible.

iv) Para todo $\alpha \in \mathbb{R}$, $\{f \leq \alpha\}$ es medible.

Prueba: $i) \Rightarrow ii)$:

$$\{f \geq \alpha\} = \bigcap_{n\in\mathbb{N}} \{f > \alpha - 1/n\}$$

Como f es medible, cada uno de los conjuntos $\{f > \alpha - 1/n\}$ pertenece a $\mathcal{M}$, y como $\mathcal{M}$ es una σ-álgebra, es cerrada por intersecciones numerables. Concluimos que $\{f \geq \alpha\} \in \mathcal{M}$.

$ii) \Rightarrow iii)$: Notamos que $\{f < \alpha\} = \Omega - \{f \geq \alpha\}$, y como $\mathcal{M}$ es cerrada por complementos, $\{f < \alpha\} \in \mathcal{M}$.

$iii) \Rightarrow iv)$: Escribimos

$$\{f \leq \alpha\} = \bigcap_{n\in\mathbb{N}} \{f < \alpha + 1/n\}$$

y utilizamos que $\mathcal{M}$ es cerrada por intersecciones numerables.

$iv) \Rightarrow i)$: Notamos que $\{f > \alpha\} = \Omega - \{f \leq \alpha\}$, y utilizamos que $\mathcal{M}$ es cerrada por complementos. □

Proposición D.1.3 *Sean $f, g : \Omega \to \overline{\mathbb{R}}$ funciones medibles. Entonces:*
$\{f < g\} = \{\omega \in \Omega : f(\omega) < g(\omega)\}$ es medible.

Prueba: Notamos que

$$\{f < g\} = \bigcup_{q\in\mathbb{Q}} \{f < q < g\} = \bigcup_{q\in\mathbb{Q}} (\{f < q\} \cap \{q < g\})$$

y usamos que $\mathcal{M}$ es una σ-álgebra y que $\mathbb{Q}$ es numerable. □

El hecho de que la σ-álgebra $\mathcal{M}$ sea cerrada por operaciones conjuntísticas numerables, tendrá como consecuencia que la clase de funciones medibles será cerrada por las operaciones algebraicas, y por las operaciones de tomar supremo o límites. Más precisamente tenemos las siguientes propiedades:

Lema D.1.4 *Sean $f, g : \Omega \to \mathbb{R}$ funciones medibles Entonces:*

i) $f + k$ y kf son medibles para todo $k \in \mathbb{R}$.

ii) $f + g$ y $f - g$ son medibles.

iii) f^2 es medible.

iv) $f \cdot g$ es medible,

v) Si $g \neq 0$, f/g es medible.

Prueba: i): $\{f + k > \alpha\} = \{f > \alpha - k\}$ Si $k > 0$: $\{kf > \alpha\} = \{f > \alpha/k\}$ mientras que si $k < 0$: $\{kf > \alpha\} = \{f < alpha/k\}$

ii): $\{f + g > \alpha\} = \{f > \alpha - g\}$ y $\alpha - g$ es medible por i)

iii): Si $\alpha \geq 0$, $\{f^2 > \alpha\} = \{f > \sqrt{\alpha}\} \cup \{f < -\sqrt{\alpha}\}$ (sino $\{f^2 > \alpha\} = \Omega$).

iv): Se deja como ejercicio (por iii) basta ver que $1/g$ es medible) □

Observación: El lema se puede adaptar al caso en que f o g toman los valores $\pm\infty$. $f + g$ está bien definida, salvo cuando es de la forma $(+\infty) + (-\infty)$ o $(-\infty) + \infty$. Para definir $f \cdot g$, hay que utilizar las convenciones $0 \cdot (\pm\infty) = (\pm\infty) \cdot 0 = 0$

Lema D.1.5 *Sea $(f_n)_{n\in\mathbb{N}}$ una sucesión de funciones medibles. Entonces*

$$\sup_{n\in\mathbb{N}} f_n(x) \quad \inf_{n\in\mathbb{N}} f_n(x)$$

$$\liminf_{n\in\mathbb{N}} f_n(x) \quad \limsup_{n\in\mathbb{N}} f_n(x)$$

son medibles.

En particular si f_n converge, entonces:

$$f(x) = \lim_{n\to+\infty} f_n(x)$$

es medible.

Prueba: Notamos que

$$\{\sup_{n\in\mathbb{N}} f_n(x) > \lambda\} = \bigcup_{n\in\mathbb{N}} \{f_n > \lambda\}$$

Por lo que si cada f_n es medible, $\{f_n > \lambda\} \in \mathcal{M}\ \forall\ n \in N$, y en consecuencia como $\mathcal{M}$ es una σ-álgebra, $\{\sup_{n\in\mathbb{N}} f_n(x) > \lambda\} \in \mathcal{M}$. Esto prueba que $\sup_n f_n(x)$ es medible.

Del mismo modo, se prueba que $\text{ínf}_n\, f_n(x)$ es medible, ya que:

$$\{\inf_{n\in N} f_n(x) < \lambda\} \bigcup_{n\in\mathbb{N}} \{f_n < \lambda\}$$

Para probar que $\limsup f_n$ es medible, notamos que

$$\limsup f_n = \inf_k \sup_{k\geq n} f_n$$

Pero para cada k, $\sup_{k\geq n} f_n$ es medible por lo que ya probamos, y en consecuencia $\limsup f_n$ es medible. De modo análogo, de que

$$\liminf f_n = \sup_k \inf_{k\geq n} f_n$$

Se deduce que $\lim \text{ínf}\, f_n$ es medible. Finalmente notamos que si la sucesión (f_n) converge, entonces $\lim_{n\to+\infty} f_n(x) = \liminf f_n(x) = \limsup f_n(x)$, por lo que la función límite de las f_n es medible. □

Definición D.1.6 *Sea $\varphi : \mathbb{R} \to \mathbb{R}$ una función. Diremos que φ es medible Borel si es medible con respecto a la σ-álgebra de Borel $\mathcal{B}(\mathbb{R})$, generada por los intervalos. Es decir si para todo intervalo $(a, b]$, su pre-imagen por φ, $\varphi^{-1}((a, b])$ es un conjunto boreliano de la recta.*

Lema D.1.7 *Sean $(\Omega, \mathcal{M})$ un espacio medible y $f : \Omega \to \mathbb{R}$ una función. Entonces f es medible si y sólo si $f^{-1}(B) \in \mathcal{M}$ para todo $B \in \mathcal{B}(\mathbb{R})$.*

Prueba: Notamos que:

$$\mathcal{A} = \{B \subset R : f^{-1}(B) \in \mathcal{M}\}$$

es una σ-álgebra. Si f es medible, entonces $\mathcal{A}$ contiene a los intervalos. Por lo tanto contiene a toda la σ-álgebra de Borel (que es la menor σ-álgebra que contiene a los intervalos). □

Corolario D.1.8 *Si (Ω, P) es un espacio medible, $f : \Omega \to \mathbb{R}$ es medible y $\varphi : \mathbb{R} \to \mathbb{R}$ es medible Borel, entonces $\varphi \circ f : \Omega \to \mathbb{R}$ es medible.*

Prueba: Sea B un boreliano de la recta, entonces $\varphi^{-1}(B)$ es boreliano, y en consecuencia como f es medible:

$$(\varphi \circ f)^{-1}(B) = f^{-1}(\varphi^{-1}(B) \in \mathcal{M}$$

Como esto vale para todo B boreliano, concluimos que $\varphi \circ f$ es medible. □

Interpretación probabilística: Sea $(\Omega, \mathcal{E}, P)$ un espacio de probabilidad. Si $X : \Omega \to \mathbb{R}$ es una variable aleatoria, y $\varphi : \mathbb{R} \to \mathbb{R}$ es medible Borel, entonces $\varphi(X) = \varphi \circ X : \Omega \to \mathbb{R}$ es una variable aleatoria.

D.1.1. Funciones Simples

Definición D.1.9 *Llamamos función simple a una función medible $f : \Omega \to \mathbb{R}$ que toma un número finito de valores $\{c_1, c_2, \ldots, c_n\}$. Podemos representarla entonces como:*

$$f = \sum_{i=1}^{n} c_i I_{E_i} \tag{D.1}$$

donde $E_i = \{\omega \in \Omega : f(\omega) = c_i\}$, y I_{A_i} es el **indicador** *(o* **función característica**[1]*) del conjunto E_i, definido por:*

$$I_{E_i}(\omega) = \begin{cases} 1 & si \quad \omega \in E_i \\ 0 & si \quad \omega \notin E_i \end{cases}$$

[1]En la teoría de probabilidades el nombre de función característica suele usarse para otra cosa, por eso preferimos en estas notas el de indicador. A veces se usa la notación χ_E en lugar de I_E

En la interpretación probabilística, las funciones simples corresponden a las variables aleatorias discretas que toman sólo un número finito de valores.

El siguiente lema de aproximación por funciones simples, será de gran utilidad para la teoría de la integral:

Lema D.1.10 *Si $f : \Omega \to [0, +\infty]$ es una función medible no negativa, entonces existe una sucesión $\varphi_n(x)$ de funciones simples no negativas tales que*

$$\lim_{n\to+\infty} \varphi_n(x) = f(x) \; \forall \; x \in \Omega$$

Prueba: Para cada $n \in N$, definimos:

$$\varphi_n(x) = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} I_{E_{n,i}}(x) + nF_n$$

siendo

$$E_{n,i} = \left\{ \{x \in \Omega : \frac{i-1}{2^n} \leq f(x) < \frac{i}{2^n} \right\}$$

$$F_n = \{x \in \Omega : f(x) \geq n\}$$

Es decir que:

$$\varphi_n(x) = \begin{cases} \frac{i-1}{2^n} & \text{si} \quad \frac{i-1}{2^n} \leq f(x) < \frac{i}{2^n} \\ n & \text{si} \quad f(x) \geq n \end{cases}$$

Se prueba que $\varphi_n(x)$ tiene las propiedades del enunciado. □

D.2. Integral de Funciones Simples

Consideramos ahora un espacio de medida $(\Omega, \mathcal{M}, \mu)$ es decir un espacio medible, donde además está definida una medida (σ-aditiva) $\mu : \mathcal{M} \to [0, +\infty]$.

Si $f : \Omega \to \mathbb{R}$ es una función simple, representada por (D.1) definimos su integral de la siguiente manera:

$$\int_\Omega f \; d\mu = \sum_{i=1} c_i \; \mu(A_i)$$

En la interpretación probabilística, tenemos un espacio de probabilidad $(\Omega, \mathcal{E}, P)$ donde la probabilidad no es otra cosa que una medida que asigna a todo el espacio Ω medida 1 (o sea: $P(\Omega) = 1$).

Entonces la definición de integral de una función simple, no es otra cosa que nuestra definición de esperanza de una variable aleatoria discreta, escrita en el lenguaje de la teoría de la medida. Es decir, que si $X : \Omega \to \mathbb{R}$ es una variable aleatoria discreta, entonces

$$E[X] = \int_\Omega X \, dP$$

La integral de las funciones simples, tiene las siguientes propiedades: (que se demuestran exactamente como las propiedades de la esperanza de variables aleatorias discretas)

Proposición D.2.1 *1. linealidad: Si f y g son funciones simples:*

$$\int_\Omega (f+g) \, d\mu = \int_\Omega f \, d\mu + \int_\Omega g \, d\mu$$

Si f es una función simple, y k una constante:

$$\int_\Omega (kf) \, d\mu = k \int_\Omega f \, d\mu$$

2. Monotonía: si f y g son funciones simples y $f \leq g$, entonces:

$$\int_\Omega f \, d\mu \leq \int_\Omega g \, d\mu$$

3. Si f es una función simple, entonces

$$\left| \int_\Omega f \, d\mu \right| \leq \int_\Omega |f| \, d\mu$$

D.3. Integral de funciones no negativas

Definición D.3.1 *Sea $(\Omega, \mathcal{M}, \mu)$ un espacio de medida, y $f : \Omega \to [0, +\infty]$ una función medible no negativa. Definimos la integral de f de la siguiente manera:*

$$\int_\Omega f d\mu = \sup \left\{ \int_\Omega \varphi \, d\mu : 0 \leq \varphi \leq f, \ \varphi \text{ simple} \right\}$$

Una consecuencia inmediata de la definición es la siguiente:

Proposición D.3.2 *Si $f, g : \Omega \to [0, +\infty]$ son funciones simples no negativas tales que $f \leq g$, entonces*

$$\int_\Omega f(x) \, d\mu \leq \int_\Omega g(x) \, d\mu$$

Definición D.3.3 *Si $A \in \mathcal{M}$ es un conjunto medible, y $f : \Omega \to [0, +\infty]$ es una función medible no negativa, definimos la integral de f sobre E como:*

$$\int_\Omega f \, d\mu = \int_\Omega f \cdot I_A \, d\mu$$

Lema D.3.4 *Sea φ una función simple no negativa. Entonces la función $\lambda = \lambda_\varphi : \mathcal{M} \to [0, +\infty]$ definida por:*

$$\lambda(A) = \int_A \varphi \, d\mu$$

es una medida

Prueba: Supongamos que un conjunto medible A se representa como una unión disjunta numerable de una sucesión $(A_n)_{n\in\mathbb{N}}$ de conjuntos medibles:

$$A = \bigcup_{n\in\mathbb{N}} A_n$$

Queremos probar que:

$$\lambda(A) = \sum_{n=1}^{\infty} \lambda(A_n)$$

Como φ es una función simple, podremos representarla en la forma

$$\varphi = \sum_{i=1}^{N} c_i \; I_{E_i}$$

siendo E_i conjuntos medibles disjuntos.

Notamos que $\varphi(x) I_{A_n}(x)$ es una función simple, que toma el valor c_i en el conjunto $A_n \cup E_i$, es decir que su representación canónica es:

$$\varphi(x) I_{A_n}(x) = \sum_{i=1}^{N} c_i \; I_{E_i \cap A_n}$$

En consecuencia,

$$\lambda(A_n) = \sum_{i=1}^{N} c_i \; \mu(E_i \cap A_n)$$

Y por lo tanto

$$\sum_{n=1}^{\infty} \lambda(A_n) = \sum_{n=1}^{\infty} \sum_{i=1}^{N} c_i \; \mu(E_i \cap A_n)$$

Como en esta suma doble los términos $\mu(E_i \cap A_n)$ son no negativos, da lo mismo efectuar la suma en cualquier orden. En consecuencia,

$$\sum_{n=1}^{\infty} \lambda(A_n) = \sum_{i=1}^{N} \sum_{n=1}^{\infty} c_i \; \mu(E_i \cap A_n) = \sum_{i=1}^{N} c_i \sum_{n=1}^{\infty} \mu(E_i \cap A_n)$$

Ahora notamos que:

$$E_i \cap A = \bigcup_{n \in N} (E_i \cap A_n)$$

siendo esta unión disjunta. En consecuencia, como μ es una medida,

$$\mu(E_i \cap A) = \sum_{n=1}^{\infty} \mu(E_i \cap A_n)$$

y concluimos que:

$$\sum_{n=1}^{\infty} \lambda(A_n) = \sum_{i=1}^{N} c_i \ \mu(E_i \cap A) = \int_{\Omega} \varphi(x) \ I_A(x) \ d\mu = \int_A \varphi(x) \ d\mu$$

□

Teorema D.3.5 (Teorema de la Convergencia Monótona) [2] *Sea $f_n(x) : \Omega \to [0, +\infty]$ una sucesión creciente (o sea: $f_n(x) \leq f_{n+1}(x)$) de funciones medibles no negativas. Entonces,*

$$\int_{\Omega} \lim_{n \to +\infty} f(x) \ d\mu = \lim_{n \to +\infty} \int_{\Omega} f_n(x) \ dmu$$

Prueba: Sea

$$f(x) = \lim_{n \to +\infty} f_n(x)$$

Por la monotonía de la integral es claro que:

$$\int_{\Omega} f_n(x) \ d\mu \leq \int_{\Omega} f(x) \ d\mu$$

Y por lo tanto que:

$$\lim_{n \to +\infty} \int_{\Omega} f_n(x) \ d\mu \leq \int_{\Omega} f(x) \ d\mu$$

Por otra parte, sea φ una función simple tal que $\varphi \leq f$. Dado $\alpha \in (0, 1)$, consideramos los conjuntos (medibles)

$$A_n = \{x \in \Omega : f_n(x) \geq \alpha \varphi x\}$$

Entonces la sucesión $(A_n)_{n \in \mathbb{N}}$ es monótona creciente (o sea $A_n \subset A_{n+1}$) y

$$\Omega = \bigcup_{n \in N} A_n$$

[2] También conocido como teorema de Beppo Levi.

Además la función λ_φ definida en el lema anterior, es una medida, por lo tanto:

$$\lambda(\Omega) = \lim_{n\to+\infty} \lambda(A_n)$$

es decir,

$$\lim_{n\to+\infty} \int_{A_n} \varphi(x)\ d\mu = \int_\Omega \varphi(x)\ d\mu$$

Por otra parte, para cada $n \in N$,

$$\alpha \int_{A_n} \varphi(x)\ d\mu \leq \int_{A_n} f_n(x)\ d\mu \leq \int_\Omega f_n(x)\ d\mu$$

De modo que,

$$\alpha \int_\Omega \varphi(x)\ d\mu = \alpha \lim_{n\to+\infty} \int_{A_n} \varphi(x) d\mu \leq \lim_{n\to+\infty} \int_\Omega f_n(x)\ d\mu$$

Haciendo tender α a 1 deducimos que:

$$\int_\Omega \varphi(x)\ d\mu \leq \lim_{n\to+\infty} f_n(x)\ d\mu$$

y por lo tanto como esto vale para toda función simple φ con $0 \leq \varphi \leq f$, por la definición de integral, deducimos que:

$$\int_\Omega f(x)\ d\mu \leq \lim_{n\to+\infty} f_n(x)\ d\mu$$

□

Proposición D.3.6 (Linealidad de la integral) *Si $f, g : \Omega \to [0, +\infty]$ son funciones medibles no negativas y $\lambda_1, \lambda_2 \geq 0$ son números reales no negativos, entonces:*

$$\int_\Omega [\lambda_1 f(x) + \lambda_2 g(x)]\ d\mu = \lambda_1 \int_\Omega f(x)\ d\mu + \lambda_2 \int_\Omega g(x)\ d\mu$$

Prueba: Utilizamos el lema de aproximación por funciones simples: sabemos que existen una sucesión creciente $(f_n(x))$ de funciones simples que converge a $f(x)$, y una sucesión creciente $(g_n(x))$ de funciones simples que converge a $g(x)$. Entonces por la linealidad de la integral de funciones simples,

$$\int_\Omega [\lambda_1 f_n(x) + \lambda_2 g_n(x)]\ d\mu = \lambda_1 \int_\Omega f_n(x)\ d\mu + \lambda_2 \int_\Omega g_n(x)\ d\mu$$

Y el teorema de convergencia monótona implica entonces que:

$$\int_\Omega [\lambda_1 f(x) + \lambda_2 g(x)]\ d\mu = \lambda_1 \int_\Omega f(x)\ d\mu + \lambda_2 \int_\Omega g(x)\ d\mu$$

□

Teorema D.3.7 (Lema de Fatou) *Sea $f_n : \mathcal{M} \to [0, +\infty]$ una sucesión de funciones medibles no negativas. Entonces:*

$$\int_\Omega \liminf_{n\to+\infty} f_n(x)\, d\mu \leq \liminf_{n\to+\infty} \int_\Omega f_n(x)\, d\mu$$

Prueba: Llamemos

$$f(x) = \liminf_{n\to+\infty} f_n(x) = \sup_{k\in N} \left(\inf_{n\geq k} f_n(x) \right)$$

y consideremos la sucesión creciente de funciones no negativas:

$$g_k(x) = \inf_{n\geq k} f_n(x)$$

Entonces por el teorema de convergencia monótona:

$$\int_\Omega f(x)\, d\mu = \int_\Omega \lim_{k\to+\infty} g_k(x)\, d\mu = \lim_{k\to+\infty} \int_\Omega g_k(x)\, d\mu \tag{D.2}$$

Por otra parte si $n \geq k$, tenemos que

$$\int_\Omega g_k(x)\, d\mu \leq \int_\Omega f_n(x)\, d\mu$$

y en consecuentcia:

$$\int_\Omega g_k(x)\, d\mu \leq \liminf_{n\to+\infty} \int_\Omega f_n(x)\, d\mu$$

Y por lo tanto:

$$\lim_{k\to+\infty} \int_\Omega g_k(x)\, d\mu \leq \liminf_{n\to+\infty} \int_\Omega f_n(x)\, d\mu$$

En consecuencia utilizando (D.2), deducimos que:

$$\int_\Omega f(x) d\mu \leq \liminf_{n\to+\infty} \int_\Omega f_n(x)\, d\mu$$

□

D.4. Funciones Integrables

Si $f : \Omega \to \overline{\mathbb{R}}$ es una función medible, hacemos la descomposición:

$$f = f^+ - f^- \tag{D.3}$$

como diferencia de dos funciones medibles no negativas, siendo

$$f^+(x) = \begin{cases} f(x) & \text{si} \quad f(x) \geq 0 \\ 0 & \text{si} \quad f(x) < 0 \end{cases}$$

y

$$f^-(x) = \begin{cases} 0 & \text{si} \quad f(x) \geq 0 \\ -f(x) & \text{si} \quad f(x) < 0 \end{cases}$$

Notamos que:

$$|f| = f^+ + f^-$$

Definición D.4.1 *Diremos que una función medible $f : \Omega \to \overline{\mathbb{R}}$ es integrable si son finitas las integrables*

$$\int_\Omega f^+(x)\, d\mu$$

y

$$\int_\Omega f^-(x)\, d\mu$$

En ese caso, definimos la integral de f con respecto a μ en el espacio Ω por:

$$\int_\Omega f(x)\, d\mu = \int_\Omega f^+(x)\, d\mu + \int_\Omega f^-(x)\, d\mu$$

Observación: De la definición de función integrable, deducimos que f es integrable si y sólo si

$$\int_\Omega |f(x)| d\mu < +\infty$$

Además:

$$\left| \int_\Omega f(x) d\mu \right| \leq \int_\Omega |f(x)|\, d\mu$$

Proposición D.4.2 (Linealidad de la integral) *Si $f, g : \Omega \to \overline{\mathbb{R}}$ son funciones integrables y λ_1, λ_2 son números reales, entonces $\lambda_1 f + \lambda_2 g$ es integrable, y se tiene que:*

$$\int_\Omega [\lambda_1 f(x) + \lambda_2 g(x)]\, d\mu = \lambda_1 \int_\Omega f(x)\, d\mu + \lambda_2 \int_\Omega g(x)\, d\mu$$

Prueba: Primero probaremos que es posible sacar escalares de la integral: En efecto si $\lambda > 0$, tenemos que:

$$(\lambda f)^+ = \lambda f^+$$

$$(\lambda f)^- = \lambda f^-$$

Entonces es claro por la definición y la linealidad de la integral para funciones no negativas, que si f es integrable, λf también lo es y se verifica que:

$$\int_\Omega \lambda f\, d\mu = \int_\Omega (\lambda f)^+\, d\mu - \int_\Omega (\lambda f)^-\, d\mu =$$

$$= \lambda \int_\Omega f^+ \, d\mu - \lambda \int_\Omega f^- \, d\mu$$

$$= \lambda \int_\Omega f \, d\mu$$

Si $\lambda < 0$, notamos que:

$$(\lambda f)^+ = (-\lambda) f^-$$
$$(\lambda f)^- = (-\lambda) f^+$$

y de nuevo, vemos usando la definición y la linealidad de la integral para funciones no negativas, que si f es integrable, λf también lo es y se verifica que:

$$\int_\Omega \lambda f \, d\mu = \int_\Omega (\lambda f)^+ \, d\mu - \int_\Omega (\lambda f)^- \, d\mu =$$

$$= -\lambda \int_\Omega f^- \, d\mu + \lambda \int_\Omega f^+ \, d\mu$$

$$= \lambda \int_\Omega f \, d\mu$$

(El caso $\lambda = 0$ es trivial porque la integral de la función nula dá 0).

Ahora probaremos que la integral distribuye la suma: Para ello notamos que (D.3) proporciona una escritura de f como diferencia de dos funciones no negativas. Pero que si tenemos otra escritura de f como diferencia de dos funciones medibles no negativas:

$$f = f_1 - f_2$$

Entonces de $f^+ - f^- = f_1 - f_2$, deducimos $f^+ + f_2 = f_1 + f^-$, entonces por la linealidad de la integral para funciones no negativas:

$$\int_\Omega f^+ \, d\mu + \int_\Omega f_2 \, d\mu = \int_\Omega f_1 \, d\mu + \int_\Omega f^- d\mu$$

En consecuencia,

$$\int_\Omega f \, d\mu = \int_\Omega f_1 \, d\mu - \int_\Omega f_2 \, d\mu$$

Vale decir que si en lugar de (D.3), utilizáramos cualquier otra descomposición de f como diferencia de funciones medibles no negativas obtendríamos el mismo valor de la integral.

Hecha esta observación, notamos que

$$f + g = f^+ - f^- + g^+ - g^- = (f^+ + g^+) - (f^- + g^-)$$

y que esta última expresión proporciona una escritura de $f+g$ como diferencia de funciones no negativas. En consecuencia, por la observación anterior, y la linealidad de la integral para funciones no negativas:

$$\int_\Omega (f+g)\,d\mu = \int_\Omega (f^+ + g^+)\,d\mu - \int_\Omega (f^- + g^-)\,d\mu =$$
$$\int_\Omega f^+\,d\mu + \int_\Omega g^+\,d\mu - \int_\Omega f^- d\mu - \int_\Omega g^-\,d\mu =$$
$$= \int_\Omega f\,d\mu + \int_\Omega g\,d\mu$$

□

Teorema D.4.3 (De convergencia mayorada, de Lebesgue) *Sea $f_n(x) : \Omega \to \overline{\mathbb{R}}$ una sucesión de funciones integrables, que converge puntualmente a una función $f(x)$*

$$f(x) = \lim_{n\to+\infty} f_n(x)$$

y tal que existe una función integrable g de modo que $|f_n(x)| \leq g$ (en casi todo punto con respecto a la medida μ). Entonces

$$\lim_{n\to+\infty} \int_\Omega |f_n(x) - f(x)|\,d\mu = 0$$

En particular,

$$\lim_{n\to+\infty} \int_\Omega f_n(x)\,d\mu = \int_\Omega f(x)\,d\mu$$

Prueba: Sea $h_n(x)$ la sucesión de funciones medibles no negativas, definida por:

$$h_n(x) = 2g(x) - |f_n(x) - f(x)|$$

Entonces, por el lema de Fatou,

$$2\int_\Omega g(x)\,d\mu = \int_\Omega \lim h_n(x)\,d\mu \leq \liminf_{n\to+\infty} \int_\Omega f_n(x)\,d\mu$$

$$2\int_\Omega g(x)\,d\mu - \limsup_{n\to+\infty} \int_\Omega |f_n(x) - f(x)|\,d\mu$$

En consecuencia,

$$\limsup_{n\to+\infty} \int_\Omega |f_n(x) - f(x)|\,d\mu = 0$$

Entonces,

$$\left|\int_\Omega f_n(x)\,d\mu - \int_\Omega f(x)\,d\mu\right| \leq \int_\Omega |f_n(x) - f(x)|\,d\mu \to 0 \text{ cuando } n \to \infty$$

□

D.5. Equivalencia de las distintas definiciones de Esperanza

Sean como antes $(\Omega, \mathcal{E}, P)$ un espacio de probabilidad y $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria. La esperanza de X no es otra cosa que su integral de Lebesgue respecto a la medida P:

$$E[X] = \int_\Omega X \, d\mu$$

A la variable aleatoria X le podemos asociar la medida μ_X (o probabilidad), definida para los conjuntos borelianos de la recta por:

$$\mu_X(B) = P(X^{-1}(B))$$

μ_X se llama la distribución de probabilidades de X. Notamos que $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$, donde $\mathcal{B}(\mathbb{R})$ denota la σ-álgebra de Borel de la recta, es un espacio de probabilidad.

El siguiente lema afirma que es posible transformar las integrales respecto a P, en integrales respecto a μ_X. Por consiguiente μ_X contiene toda la información sobre X que es necesaria para calcular la esperanza de X, o más generalmente, de una función $\varphi(X)$ de X.

Lema D.5.1 *Sea $\varphi : \mathbb{R} \to \mathbb{R}$ una función medible Borel. Entonces se tiene que*

$$E[\varphi(X)] = \int_\Omega \varphi(X) \, d\mu = \int_\mathbb{R} \varphi(x) \, d\mu_X$$

en el siguiente sentido.

1. *Si φ es no negativa, la fórmula vale sin restricciones. (Notar que estas integrales siempre existen, aunque pueden ser infinitas)*
2. *Si φ es cualquiera, entonces $\varphi(X)$ es integrable con respecto a P si y sólo si $\varphi(x)$ lo es con respecto a μ_X y en este caso es válida dicha fórmula.*

Prueba: Primero consideramos el caso en que $\varphi : \mathbb{R} \to \mathbb{R}$ es una función boreliana simple, entonces:

$$\varphi(x) = \sum_{i=1}^{n} c_i \, I_{B_i}(x)$$

para ciertos conjuntos $B_i \subset \mathbb{R}$ borelianos, de modo que:

$$\int_\mathbb{R} \varphi(x) \, d\mu_X = \sum_{i=1}^{n} c_i \, \mu_X(B_i)$$

Por otra parte, notamos que $\varphi(X) : \mathcal{M} \to \mathbb{R}$ es una función simple que toma el valor c_i en el conjunto $X^{-1}(B_i)$, de modo que:

$$\int_\Omega \varphi(X)\, dP = \sum_{i=1}^{n} c_i\, P(X^{-1}(B_i))$$

Dado que por definición de μ_X, $\mu_X(B_i) = P(X^{-1}(B_i))$, ambas integrales coinciden.

Sea ahora $\varphi : \mathbb{R} \to \mathbb{R}$ una función boreliana no negativa. Y consideramos una sucesión creciente de funciones borelianas simples $\varphi_n : \mathbb{R} \to \mathbb{R}$ que converge a φ en forma creciente. Dado que para cada $n \in \mathbb{N}$ tenemos que:

$$\int_\Omega \varphi_n(X)\, dP = \int_{\mathbb{R}} \varphi_n(x)\, d\mu_X$$

El teorema de convegencia monótona, implica que:

$$\int_\Omega \varphi(X)\, dP = \int_{\mathbb{R}} \varphi(x)\, d\mu_X$$

Finalmente, consideremos una función boreliana $\varphi : \mathbb{R} \to \mathbb{R}$ cualquiera. Como $|\varphi|$ es no negativa, ya sabemos que:

$$\int_\Omega |\varphi(X)|\, dP = \int_{\mathbb{R}} |\varphi(x)|\, d\mu_X$$

En consecuencia, $\varphi(X)$ es integrable con respecto a P si y sólo si $\varphi(x)$ lo es con respecto a μ_X.
Finalmente, hagamos uso de la descomposición:

$$\varphi(x) = \varphi^+(x) - \varphi^-(x)$$

Entonces como φ^+ y φ^- son no negativas, tenemos que:

$$\int_\Omega \varphi^+(X)\, dP = \int_{\mathbb{R}} \varphi^+(x)\, d\mu_X$$

y que:

$$\int_\Omega \varphi^-(X)\, d\mu = \int_{\mathbb{R}} \varphi^-(x)\, d\mu_X$$

La linealidad de la integral implica entonces que:

$$\int_\Omega \varphi(X)\, dP = \int_{\mathbb{R}} \varphi(x)\, d\mu_X$$

□

Anteriormente definimos la esperanza utilizando integrales de Stieltjes respecto a la función de distribución de X. El siguiente teorema afirma que la definición de esperanza que dimos anteriormente coincide con la nueva definición.

En la demostración, utilizaremos la notación:

$$\int_a^b \varphi(x)\, d\mu_X = \int_{[a,b]} \varphi(x)\, d\mu_X$$

Teorema D.5.2 *Sea $\varphi : \mathbb{R} \to \mathbb{R}$ una función continua. Entonces se tiene que:*

$$\int_a^b \varphi(x)\, d\mu_X = \int_a^b \varphi(x)\, dF(x)$$

en el siguiente sentido:

1. *Si φ tiene soporte en un intervalo $[a, b]$ entonces, la fórmula es válida, y ambos mimebros son finitos.*
2. *Si φ es no negativa, la fórmula es válida sin restricciones (aunque ambas integrales pueden ser infinitas)*
3. *Si φ es de signo abitrario, entonces $\varphi(x)$ es integrable con respecto a μ_X si y sólo si*

$$\int_{-\infty}^{\infty} |\varphi(x)|\, dF(x) < +\infty$$

y en este caso, también es válida dicha fórmula.

Prueba: Supongamos primero que φ tiene soporte en un intervalo cerrado $[a, b]$. Consideremos una partición $\pi : a = x_0 < x_1 < \ldots < x_n = b$ del intervalo y elijamos puntos intermedios $\xi_i \in (x_i, x_{i+1})$.

Definamos la función simple $\varphi_\pi : [a, b] \to \mathbb{R}$ dada por:

$$\varphi_\pi(x) = \xi_i \text{ si } x \in (x_i, x_{i+1}]$$

Entonces:

$$S(\pi, F) = \sum_{i=1}^{n-1} \varphi(xi_i)[F(x_{i+1}) - F(x_i)] = \sum_{i=1}^{n-1} \varphi(\xi_i)\mu_X((x_i.x_{i+1}]) = \int_\Omega \varphi_\pi(x)\, d\mu$$

Ahora bien, como φ es uniformemente continua en $[a, b]$, deducimos que φ_π converge uniformenete a φ en $[a, b]$ cuando la norma de la partición π tiende a cero. En efecto, dado $\varepsilon > 0$, sea $\delta > 0$ el que corresponde a ε por la continuidad uniforme de φ en $[a, b]$. Entonces, si $x \in (x_i, x_{i+1}]$,

$$|\varphi_\pi(x) - \varphi(x)| = |\varphi(\xi_i) - \varphi(x)| < \varepsilon$$

si $|x_{i+1} - x_i| < \delta$.

Deducimos que:

$$\lim_{|\pi|\to 0} \int_a^b \varphi_\pi(x)\ d\mu_X = \int_a^b \varphi(x)\ d\mu_X$$

ya que

$$\left| \int_\Omega \varphi_\pi(X) d\mu_X - \int_\Omega \varphi(x)\ d\mu_X \right| \leq \int_a^b |\varphi_\pi(x) - \varphi_\pi(x)|\ d\mu_X < \varepsilon \mu_X([a,b]) \leq \varepsilon$$

Por definición de integral de Stieltjes esto dice que la integral

$$\int_a^b \varphi(x)\ dF(x)$$

existe, y coincide con

$$\int_a^b \varphi(x)\ d\mu_X$$

Para el caso general, en el que φ no tiene soporte compacto, consideremos cualquier sucesión decreciente $(a_n)_{n\in\mathbb{N}}$ tal que $a_n \to -\infty$, y cualquier sucesión creciente $(b_n)_{n\in\mathbb{N}}$ tal que $b_n \to +\infty$, y observemos que

$$\int_{a_n}^{b_n} \varphi_(x)\ d\mu_X = \int_{\mathbb{R}} \varphi(x) I_{[a_n,b_n]}(x)\ d\mu(x) \to \int_{\mathbb{R}} \varphi(x)\ d\mu_X$$

Por el teorema de convergencia monótona aplicado a $\varphi(x)I_{[a_n,b_n]}$, si φ es no negativa. En consecuencia,

$$\int_{\mathbb{R}} \varphi(x)\ d\mu_X = \int_{-\infty}^{\infty} \varphi(x)\ dF(x) \tag{D.4}$$

vale siempre que $\varphi(x)$ sea no negativa.

Cuando φ tiene cualquier signo, observamos primero que

$$\int_{\mathbb{R}} |\varphi(x)|\ d\mu = \int_{-\infty}^{\infty} |\varphi(x)|\ dF(x)$$

Lo que en particular, dice que $|\varphi(x)|$ es integrable con respecto a μ_X si y sólo si:

$$\int_{-\infty}^{\infty} |\varphi(x)|\ dF(x) < +\infty$$

Si esto sucede, podemos aplicar el teorema de convergencia mayorada a la sucesión $\varphi(x)I_{[a_n,b_n]}$ (que claramente está mayorada por $|\varphi(x)|$, y deducir que la fórmula (D.4) es cierta, también en este caso. □

D.5.1. Vectores Aleatorios

Las ideas anteriores pueden generalizarse facilmente a vectores aleatorios. Si $(\Omega, \mathcal{E}, P)$ es un espacio de probabilidad, un vector aleatorio no es otra cosa que una función medible $\Omega : X \to \mathbb{R}^n$.

Podemos definir la distribución de probabilidades de X como la medida μ_X, definida en la σ-álgebra de Borel de $\mathbb{R}^n$ por:

$$\mu_X(B) = P(X^{-1}(B)$$

Y si $\varphi : \mathbb{R}^n \to \mathbb{R}$ es una función medible Borel, entonces tendremos la fórmula (generalización del lema D.5.1:

$$E[\varphi(X)] = \int_\Omega \varphi(X)\ dP = \int_{\mathbb{R}^n} \varphi(x) d\mu_X$$

Apéndice E

Independencia

En este apéndice utilizaremos las herramientas de la teoría de la medida para probar algunas propiedades de las variables aleatorias independientes.

E.1. El teorema $\pi - \lambda$ de Dynkin

Para la prueba de algunos teoremas de la teoría de probabilidades (y de la teoría de la medida) se necesita un resultado técnico conocido como el teorema $\pi - \lambda$ de Dynkin. Para enunciarlo, necesitamos algunas definiciones previas:

Definición E.1.1 *Sea Ω un conjunto. Una clase $\mathcal{P}$ de subconjuntos de Ω se llamará un π-sistema si es cerrado bajo intersecciones finitas, o sea si $A, B \in \mathcal{P} \Rightarrow A \cap B \in \mathcal{P}$.*

Definición E.1.2 *Una clase $\mathcal{L}$ se subconjuntos de Ω se llama un λ-sistema si verifica las siguientes propiedades:*

λ_1) $\Omega \in \mathcal{L}$

λ_2) $A \in \mathcal{L} \Rightarrow A^c = \Omega - A \in \mathcal{L}$

λ_3) *Si (A_n) es una familia numerable disjunta y $A_n \in \mathcal{L}$, entonces $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{L}$*

Obs: Debido a la condición de que los conjuntos sean disjuntos en la condición λ_3), la definición λ-sistema es mucho más débil que la de σ-álgebra. Toda σ-álgebra es un λ-sistema pero la recíproca no es válida.

Algunas propiedades de los λ-sistemas

- $\emptyset \in \mathcal{L}$

- Si $A \subset B$, y $A, B \in \mathcal{L} \rightarrow B - A \in \mathcal{L}$.

  Prueba: $B - A = B \cap A^c = (B^c \bigcup A)^c$ y $B^c \cap A = \emptyset$.

- $\mathcal{L}$ es cerrado por uniones numerables crecientes. Si $A_n \in \mathcal{L}\ \forall\ n \in N$, y $A_1 \subset A_2 \subset \ldots \subset A_n \subset \ldots$, entonces $\bigcup_{n\in N} A_n \in \mathcal{L}$.

  Prueba:

$$\bigcup_{n\in N} A_n = A_1 \cup (A_2 - A_1) \cup (A_3 - A_2) \cup \ldots \cup (A_n - A_{n-1} \cup \ldots$$

- Si $\mathcal{L}$ es a la vez un λ-sistema y un π-sistema, entonces $\mathcal{L}$ es una σ-álgebra.

Notación: Si $\mathcal{P}$ es una familia de partes de Ω, notamos por $\sigma(P)$ la σ-álgebra generada por L.

Teorema E.1.3 (Teorema $\pi - \lambda$ de Dynkin) *Si $\mathcal{P}$ es un π-sistema, $\mathcal{L}$ es un λ-sistema, y $\mathcal{P} \subset \mathcal{L}$ entonces $\sigma(P) \subset \mathcal{L}$.*

Prueba: Sea $\mathcal{L}_0$ el λ-sistema generado por $\mathcal{P}$, esto es la intersección de todos los λ-sistemas que contienen a $\mathcal{P}$ (que es a su vez un λ-sistema). Notamos que en particular $\lambda L_0 \subset \mathcal{L}$. Afirmamos que $\mathcal{L}_0$ es un π-sistema. Para probar que $\mathcal{L}_0$ es un π-sistema, procedemos del siguiente modo: dado $A \in \mathcal{L}$, definimos

$$\mathcal{L}_A = \{B \subset \Omega : A \cap B \in \mathcal{L}_0\}$$

Afirmación 1: Si $A \in \mathcal{L}_0$, entonces $\mathcal{L}_A$ es un λ-sistema.

- $A \cap \Omega = A \in \mathcal{L}_0$ por hipótesis, luego $\Omega \in \mathcal{L}_A$.

- Si $B_1, B_2 \in \mathcal{L}_A$ y $B_1 \subset B_2$, entonces por definición $A \cap B_1, A \cap B_2 \in \mathcal{L}_0$. Ahora como $\mathcal{L}_0$ es un λ-sistema y $A \cap B_1 \subset A \cap B_2$, tenemos que $A \cap B_1 - A \cap B_2 = A \cap (B_1 - B_2) \in \mathcal{L}_0$. En consecuencia, $B_1 - B_2 \in \mathcal{L}_A$.

- Si (B_n) es una familia disjunta de conjuntos de $\mathcal{L}_A$ entonces $A \cap B_n$ es una familia disjunta de conjuntos de $\mathcal{L}_0$, y como

$$A \cap \left(\bigcup_{n\in\mathbb{N}} B_n \right) = \bigcap_{n\in N} (A \cap B_n) \in \mathcal{L}_0$$

  entonces

$$\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{L}_A$$

Afirmación 2: Si $A \in \mathcal{P}$, entonces $\mathcal{L}_0 \subset \mathcal{L}_A$.

Si $A \in \mathcal{P}$, entonces para cualquier $B \in \mathcal{P}$ tenemos que $A \cap B \in \mathcal{P}$, ya que $\mathcal{P}$ es por hipótesis un π-sistema. Deducimos que $\mathcal{P} \subset \mathcal{L}_A$. Luego por la afirmación 1, $\mathcal{L}_A$ es un λ-sistema que continene a $\mathcal{P}$, lo cual por la definición de $\mathcal{L}_0$ implica que $\mathcal{L}_0 \subset \mathcal{L}_A$.

Afirmación 3: Si $C \in \mathcal{L}_0$, entonces $\mathcal{L}_0 \subset \mathcal{L}_C$.

Para todo $A \in \mathcal{P}$, por la afirmación 2, tenemos que $\mathcal{L}_A \subset \mathcal{L}_0$. Luego si $C \in \mathcal{L}_0$, entonces $C \in \mathcal{L}_A$, que por simetría de la definición implica que $A \in \mathcal{L}_C$. Como esto vale para todo $A \in \mathcal{P}$, deducimos que $\mathcal{P} \subset \mathcal{L}_C$.

Por la afirmación 1, deducimos que $\mathcal{L}_C$ es un λ-sistema que contiene a $\mathcal{P}$, lo que por la definición de $\mathcal{L}_0$, implica que $\mathcal{L}_0 \subset \mathcal{L}_C$.

Finalmente sean $D, E \in \mathcal{L}_0$. Entonces por la afirmación 3, $D \in \mathcal{L}_0 \subset \mathcal{L}_E$. En consecuencia por definición de $\mathcal{L}_E$, $D \cap E \in \mathcal{L}_0$. Concluimos que $\mathcal{L}_0$ es un π-sistema.

Conclusión de la prueba: Como $\mathcal{L}_0$ es a la vez un π-sistema, y un λ-sistema, es una σ-álgebra. Como contiene a $\mathcal{P}$, deducimos que $\sigma(\mathcal{P}) \subset \mathcal{L}_0$. Y entonces, como $\mathcal{L}_0 \subset \mathcal{L}$, concluimos que $\sigma(\mathcal{P}) \subset \mathcal{L}$. □

E.2. Variables independientes

Si X e Y son dos variables aleatorias, recordamos que X e Y se dicen independientes si para cualquier par de intervalos $(a, b]$ y $(c, d]$ de la recta, los eventos $\{X \in (a, b]\}$ y $\{Y \in (c, d]$ son idenpendientes, es decir que:

$$P\{(X, Y) \in (a, b] \times (c, d]\} = P\{X \in (a, b]\} \times P\{Y \in (c, d]\}$$

Podemos interpretar esta fórmula como:

$$\mu_{(X,Y)}((a, b] \times (c, d]) = \mu_X((a, b])\mu_Y((c, d])$$

El siguiente lema afirma que una fórmula análoga es válida si sustituimos los intervalos por conjuntos borelianos de la recta:

Lema E.2.1 *Sean X e Y dos variables aleatorias. Entonces X e Y son independientes si y sólo si:*

$$P\{(X, Y) \in B_1 \times B_2\} = P\{X \in B_1\} \cdot P\{Y \in B_2\}$$

para cualquier par B_1, B_2 de conjuntos borelianos de la recta.

Prueba: Fijemos primero B_1, como siendo un intervalo $(a, b]$ de la recta, y consideremos la familia

$$\mathcal{L}_1 = \{B \subset \mathbb{R} : P\{(X, Y) \in (a, b] \times B\} = P\{X \in (a, b]\} \cdot P\{Y \in B\}\}$$

Afirmamos que $\mathcal{A}_1$ es un λ-sistema de subconjuntos de $\mathbb{R}$. Chequeamos las tres condiciones de la definición:

$\lambda_1)$ $\mathbb{R} \in \mathcal{L}_1$:

$$P\{(X,Y) \in (a,b] \times \mathbb{R}\} = PX \in (a,b] = P\{X \in (a,b]\} \cdot P\{Y \in \mathbb{R}\}$$

ya que $P\{Y \in \mathbb{R} = 1$.

$\lambda_2)$ $B \in \mathcal{L}_1 \Rightarrow B^c = \mathbb{R} - B \in \mathcal{L}_1$

En efecto,

$$\begin{aligned} P\{(X,Y) \in (a,b] \times B^c\} &= P\{(X,Y) \in (a,b] \times \mathbb{R}\} - P\{(X,Y) \in (a,b] \times B\} \\ &= P\{X \in (a,b]\} - P\{X \in (a,b]\}P\{Y \in B\} \\ &= P\{X \in (a,b]\}(1 - P\{Y \in B) \\ &= P\{X \in (a,b]\}P\{Y \in B^c\} \end{aligned}$$

$\lambda_3)$ Si (B_n) es una familia numerable disjunta y $B_n \in \mathcal{L}_1$, entonces $B = \bigcup_{n\in\mathbb{N}} B_n \in \mathcal{L}_1$

En efecto, utilizando que los B_n son disjuntos, tenemos que:

$$\begin{aligned} P\{(X,Y) \in (a,b] \times B\} &= P\{(X,Y) \in \bigcup_{n\in N} ((a,b] \times B_n)\} \\ &= \sum_{n\in N} P\{(X,Y) \in (a,b] \times B_n\} \\ &= \sum_{n\in\mathbb{N}} P\{X \in (a,b]\}P\{Y \in B_n\} \\ &= P\{X \in (a,b]\}\left(\sum_{n\in\mathbb{N}} P\{Y \in B_n\}\right) \\ &= P\{X \in (a,b]\}P\{Y \in B\} \end{aligned}$$

Notemos que no es posible probar que $\mathcal{L}_1$ sea una σ-álgebra, pues este argumento no funciona si los B_n no fueran disjuntos.

Por otra parte la familia $\mathcal{P}$ de los intervalos semiabiertos de la recta (contando como intervalo semiabierto al conjunto vacío $(a,a] = \emptyset$ es un π-sistema, y por la definición de variables aleatorias independientes, $\mathcal{P} \subset \mathcal{L}_1$.

El teorema $\pi - \lambda$ nos permite concluir entonces que $\sigma(\mathcal{P}) \subset \mathcal{L}_1$, es decir: que la σ-álgebra $\mathcal{B}(\mathbb{R})$ de los borelianos de la recta, está contenida en $\mathcal{L}_1$. Entonces, hemos probado

que la fórmula del enunciado, se verifica cuando B_1 es un intervalo semiabierto y B_2 un boreliano arbitrario.

Ahora, repetimos el argumento, fijando la otra variable. Para ello consideramos la familia:

$$\mathcal{L}_2 = \{B \subset \mathbb{R} : P\{(X,Y) \in B \times B_2\} = P\{X \in B\} \cdot P\{Y \in B_2\} : \forall\, B \in \mathcal{B}(\mathbb{R})\ \}$$

Repitiendo el argumento anterior, podemos probar que $\mathcal{L}_2$ es un λ-sistema, y por lo anteriormente probado, $\mathcal{L}_2$ contiene a la clase $\mathcal{P}$ de los intervalos semiabiertos. Nuevamente, por el teorema $\pi - \lambda$, $\mathcal{L}_2$ contiene a los borelianos. Pero esto significa precisamente, que la fórmula del enunciado es válida para B_1, B_2 borelianos arbitrarios de la recta.

□

Corolario E.2.2 *Sean X, Y variables aleatorias independientes, y sean $\varphi_1, \varphi_2 : \mathbb{R} \to \mathbb{R}$ funciones medibles Borel. Entonces: $\varphi_1(X)$ y $\varphi_2(Y)$ son variables aleatorias independientes.*

Estos resultados se generalizan a varias variables independientes.

E.3. Esperanza del producto de variables independientes

A modo de ilustración de la utilidad de los teoremas de paso al límite en la integral, demostraremos la siguiente propiedad:

Teorema E.3.1 *Si X e Y son variables aleatorias independientes con esperanza finita (esto es, integrables) entonces*

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

Prueba: Hacemos uso una vez más del método de aproximación por funciones simples. Supongamos pues primero que X e Y son no negativas, y sean (X_n) e (Y_n) variables aleatorias simples (discretas) tales que X_n converja a X en forma creciente, e Y_n converja en forma creciente a Y, dadas por la construcción del lema D.1.10. Notamos que como consecuencia de dicha construcción, si X e Y son independientes, X_n e Y_n resultan independientes. En consecuencia, como ya probamos que el resultado es cierto para variables discretas (proposición 3.2.9) , tenemos que

$$E[X_n \cdot Y_n] = E[X_n] \cdot E[Y_n]$$

Ahora, en virtud del teorema de convergencia monótona,

$$E[X_n] \to E[X]$$

$$E[Y_n] \to E[Y]$$

$$E[X_n \cdot Y_n] \to E[X \cdot Y]$$

Luego,

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

Esto establece el resultado para funciones no negativas. En el caso general, hacemos uso, una vez más de la descomposición:

$$X = X^+ - X^-$$

$$Y = Y^+ - Y^-$$

Entonces,

$$E[X \cdot Y] = E[(X^+ - X^-)(Y^+ - Y^-] = E[X^+Y^+ - X^-Y^+ - X^+Y^- + X^-Y^-] =$$

$$E[X^+Y^+] - E[X^-Y^+] - E[X^+Y^-] + E[X^-Y^-]$$

Pero como X e Y son independientes X^+, X^- son independientes de Y^+, Y^- respectivamente; en consecuencia:

$$E[X \cdot Y] = E[X^+]E[Y^+] - E[X^-]E[Y^+] - E[X^+]E[Y^-] + E[X^-]E[Y^-] =$$

$$(E[X^+] - E[X^-])(E[Y^+] - E[Y^-]) = E[X]E[Y]$$

□

La prueba de este teorema ilustra como los teoremas de paso al límite resultan útiles para generalizar las propiedades que conocemos para variables discretas, al caso de variables aleatorias continuas.

Apéndice F

Existencia de las Integrales de Riemann-Stieltjes

En esta apéndice, presentaremos una prueba del siguiente resultado fundamental de la teoría de la integral de Riemann-Stieltjes:

Teorema F.0.1 *Si F es una función creciente en un intervalo cerrado $[a,b]$ de la recta, y φ es una función continua en $[a,b]$, entonces la integral de Riemann-Stieltjes*

$$\int_a^b \varphi(x)\, dF(x)$$

existe

Recordamos que esta integral, se define como el límite conforme la norma $|\pi|$ de la partición tiende a cero, de las sumas:

$$S_\pi(\varphi.F) = \sum_{i=0}^{n-1} \varphi(\xi_i)(F(x_{i+1}) - F(x_i))$$

donde $\pi : a = x_0 < x_1 < \ldots x_n = b$ es una partición de $[a,b]$ y $\xi_i \in [x_i, x_{i+1}]$ es un punto intermedio.

Estas sumas son poco manejables para nuestros propósitos pues dependen de los puntos intermedios ξ_i variables. Por ello, las reemplazamos por sumas superiores e inferiores que son de más fácil manejo:

Para cada i $(0 \leq i \leq n-1)$, notamos:

$$m_i = \inf_{x \in [x_i, x_{i+1}]} \varphi(x)$$

$$M_i = \sup_{x \in [x_i, x_{i+1}]} \varphi(x)$$

y consideramos las sumas superiores U_π y las sumas inferiores L_π definidas por:

$$L_\pi(\varphi.F) = \sum_{i=0}^{n-1} M_i(F(x_{i+1}) - F(x_i))$$

$$U_\pi(\varphi.F) = \sum_{i=0}^{n-1} M_i(F(x_{i+1}) - F(x_i))$$

Es claro entonces que:

$$L_\pi(\varphi, F) \leq S_\pi(\varphi, F) \leq U_\pi(\varphi, F)$$

Las sumas superiores e inferiores, tienen la siguiente propiedad importante (de monotonía): Si π' es un refinamiento de π, entonces

$$L_{\pi'}(\varphi, F) \geq L_\pi(\varphi, F)$$

$$U_{\pi'}(\varphi, F) \leq U_\pi(\varphi, F)$$

(Las sumas superiores decrecen al afinar la partición, mientras que las inferiores crecen.)

Para demostrarla, es fácil observar que se verifica si π' es una partición obtenida de π agregando un punto. Por inducción, se obtiene el caso general, ya que si π' es un refinamiento de π, ello significa que se obtiene de π agregando finitos puntos.

De esta observación, se deduce lo siguiente: toda suma superior es mayor que cualquier suma inferior. Es decir que si π y π' son dos particiones arbitrarias, siempre se verifica que:

$$L_\pi(\varphi, F) \leq U_{\pi'}(\varphi, F)$$

Para demostrar esta afirmación, es suficiente notar que la partición $\pi'' = \pi \cup \pi'$ es un refinamiento común [1]

Entonces, utilizando la propiedad de monotonía,

$$L_\pi(\varphi, F) \leq L_{\pi''} \leq U_{\pi''} \leq U_{\pi'}$$

Lema F.0.2 *Dado $\varepsilon > 0$, existe $\delta > 0$ tal que si $|\pi| < \delta$, tenemos que*

$$0 \leq U_\pi(\varphi, F) - L_\pi(\varphi, F) < \varepsilon$$

[1] Es esta propiedad de las particiones, de que dos particiones siempre tienen un refinamiento común, hace de las particiones un *conjunto dirigido*. Así pués, $S_\pi(\varphi, F)$ es una *red* que converge a la integral de Stieltjes.

Prueba: Dado $\varepsilon > 0$, como φ es uniformemente continua en $[a, b]$, existirá un $\delta > 0$ tal que si $|x-y| < \delta$ con $x, y \in [a, b]$, se tiene que $|\varphi(x)-\varphi(y)| < \varepsilon$. Entonces, si π es cualquier partición de $[a, b]$ tal que $|\pi| < \delta$, tendremos que:

$$U_\pi(\varphi, F) - L_\pi(\varphi, F) = \sum_{i=0}^{n-1} (M_i - m_i)(F(x_{i+1}) - F(x_i))$$

$$\leq \sum_{i=0}^{n-1} \varepsilon(F(x_{i+1}) - F(x_i)) \leq \varepsilon(F(b) - F(a))$$

□

Hechas estas observaciones, estamos en condiciones de demostrar el teorema, para ello comencemos eligiendo una sucesión (π_n) de particiones de $(a, b]$ de modo que π_{n+1} sea un refinamiento de π_n, y que $|\pi_n| \to 0$. Por ejemplo, podemos elegir como π_n la partición uniforme de $[a, b]$ en 2^n partes de igual longitud.

Entonces, por la propiedad de monotonía la sucesión de sumas inferiores $L_{\pi_n}(\varphi, F)$ será monótona creciente, y además está acotada pues

$$L_{\pi_n} \leq \left(\sup_{x \in [a,b]} \varphi(x) \right) (F(b) - F(a))$$

En consecuencia, existe el límite

$$I = \lim_{n \to +\infty} L_{\pi_n}(\varphi, F)$$

En virtud del lema, también tendremos que:

$$I = \lim_{n \to +\infty} U_{\pi_n}(\varphi, F)$$

Dado $\varepsilon > 0$, sea $\delta > 0$ el que corresponde a ε de acuerdo al lema, y elijamos n tal que $|\pi_n| < \delta$, y

$$|L_{\pi_n} - I| < \varepsilon$$

$$|U_{\pi_n} - I| < \varepsilon$$

Afirmamos entonces que:

$$|S_\pi(\varphi, F) - I| < 2\varepsilon$$

En efecto,

$$S_\pi(\varphi, F) - I \leq U_\pi(\varphi, F) - U_{\pi_n} + U_{\pi_n} - I$$

$$\leq U_\pi(\varphi, F) - L_\pi(\varphi, F) + \varepsilon < 2\varepsilon$$

Similarmente,

$$\begin{aligned} S_\pi(\varphi, F) - I &\geq L_\pi(\varphi, F) - L_{\pi_n} + L_{\pi_n} - I \\ &\geq L_\pi(\varphi, F) - U_\pi(\varphi, F) - \varepsilon > -2\varepsilon \end{aligned}$$

En consecuencia,

$$\lim_{|\delta|\to 0} S_\pi(\varphi, F) = I$$

Una observación adicional nos será útil para demostrar el teorema de Helly sobre paso al límite en la integral de Stieltjes: este δ sólo depende de la continuidad uniforme de φ y de la magnitud de la variación $F(b) - F(a)$ de F en $[a, b]$ (La partición π_n sólo juega un rol auxiliar en el argumento, pero δ es independiente de n y por lo tanto de F mientras $F(b) - F(a)$ permanezca acotado). Esto nos proporciona el siguiente corolario (sobre convergencia uniforme de la integral de Stieltjes respecto de la función F):

Corolario F.0.3 *Sea $\varphi \in C[a, b]$. Dados $\varepsilon > 0$ y $C > 0$, existe un $\delta > 0$ (que depende de $\varepsilon > 0$ y C pero es independiente de F) tal que si F es cualquier función $F : [a, b] \to \mathbb{R}$ creciente tal que*

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

y π una partición de $(a, b]$ con puntos marcados tal que $|\pi| < \delta$ entonces

$$\left| \int_a^b \varphi(x)\, dF(x) - S_\pi(\varphi, F) \right| < \varepsilon$$

Apéndice G

Las Leyes Fuertes de Kolmogorov

En este apéndice expondremos la demostración de la ley fuerte de los grandes números de Kolmogorov.

G.1. La Desigualdad de Kolmogorov

La desigualdad de Kolmogorov es una generalización de la desigualdad de Chebyshev:

Proposición G.1.1 (Desigualdad de Kolmogorov) *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes tales que $E[X_k] = 0$ y $Var(X_k) < +\infty$ para $k = 1, 2, \ldots, n$. Pongamos:*

$$S_n = X_1 + X_2 + \ldots + X_n$$

Entonces para todo $\lambda > 0$,

$$P\left\{\max_{1\leq k\leq n} |S_k| \geq \lambda\right\} \leq \frac{1}{\lambda^2} Var(S_n) = \frac{1}{\lambda^2} \sum_{k=1}^{n} Var(X_k)$$

donde $S_k = X_1 + X_2 + \ldots + X_n$.

Prueba: Consideremos el evento:

$$A = \left\{\max_{1\leq k\leq n} S_k^2 \geq \lambda^2\right\}$$

Queremos obtener una cota para $P(A)$. Para ello lo descomponemos en eventos disjuntos, de acuerdo a cual es la primera vez que $S_k^2 \geq \lambda^2$:

$$A_1 = \{S_1^2 \geq \lambda^2\}$$

$$A_2 = \{S_1^2 < \lambda, S_2^2 \geq \lambda^2\}$$

y en general:

$$A_k = \{S_1^2 < \lambda^2, S_2^2 < \lambda^2, \ldots, S_{k-1}^2 < \lambda^2, S_k \geq \lambda^2\}$$

Entonces los A_k son disjuntos dos a dos, y

$$A = \bigcup_{k \in N} A_k$$

Luego,

$$I_A = \sum_{k=1}^{n} I_{A_n}$$

$$S_n^2 \geq S_n^2 I_A = \sum_{k=1}^{n} S_n^2 I_{A_k}$$

y tomando esperanza:

$$E[S_n^2] \geq \sum_{k=1}^{n} E[S_n^2 I_{A_k}] \tag{G.1}$$

Nos gustaria sustituir S_n por S_k en esta sumatoria. Para ello, notamos que:

$$S_n^2 = (S_n - S_k + S_k)^2 = (S_n - S_k)^2 + 2S_k(S_n - S_k) + S_k^2 \geq 2S_k(S_n - S_k) + S_k^2$$

Multiplicando por I_{A_k} y tomando esperanza tenemos que:

$$E[S_n^2 I_{A_k}] \geq E[S_k^2 I_{A_k}] + 2E[S_k(S_n - S_k) I_{A_k}]$$

Observamos ahora que $S_k I_{A_k}$ y $S_n - S_k$ son independientes (pues $S_k I_{A_k}$ depende de $X_1, X_2, \ldots, X_k$ y $S_n - S_k$ depende de $X_{k+1}, X_{k+2}, \ldots, X_n$). En consecuencia:

$$E[S_k(S_n - S_k) I_{A_k}] = E[S_k I_{A_k}] E[S_n - S_k] = 0$$

pues $E[S_n] = E[S_k] = 0$. En consecuencia:

$$E[S_n^2 I_{A_k}] \geq E[S_k^2 I_{A_k}]$$

Ahora en A_k, $S_k^2 \geq \lambda^2$. En consecuencia,

$$E[S_n^2 I_{A_k}] \geq E[\lambda^2 I_{A_k}] = \lambda^2 P(A_k)$$

Sustituyendo este resultado en la desigualdad (G.1), tenemos que:

$$E[S_n^2] \geq \lambda^2 \sum_{k=1}^{n} P(A_k) = \lambda^2 P(A)$$

Luego

$$P(A) \leq \frac{1}{\lambda^2} E[S_n^2] = \frac{1}{\lambda^2} \sum_{k=1}^{n} \text{Var}(X_k)$$

□

G.2. La Ley Fuerte de los Grandes Números

G.2.1. La Primera Ley Fuerte de Kolmogorov

Teorema G.2.1 (Primera ley fuerte de Kolmogorov) *Sea $(X_n)_{n\in N}$ una sucesión de variables aleatorias independientes con esperanza finita, y supongamos que:*

$$\sum_{n=1}^{\infty} \frac{Var(X_n)}{n^2} < +\infty \tag{G.2}$$

Entonces $(X_n)_{n\in N}$ verifica la ley fuerte de los grandes números, es decir:

$$\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n} \xrightarrow{c.s.} 0$$

Prueba: Podemos suponer sin perdida de generalidad que $E[X_n] = 0 \ \forall \ n \in \mathbb{N}$ (Sino cambiamos X_n por $X_n - E[X_n]$). Queremos probar que:

$$\frac{S_n}{n} \xrightarrow{c.s.} 0$$

donde $S_n = X_1 + X_2 + \ldots + X_n$. Definamos las "variables maximales diádicas":

$$M_n = \max_{2^n < k \leq 2^{n+1}} \frac{|S_k|}{k}$$

Basta probar que $M_n \to 0$ casi seguramente.

Vamos a probar esto en dos etapas:
Etapa 1: Probaremos que

$$\sum_{n=1} P\left\{ M_n > \frac{1}{m} \right\} < +\infty$$

para $m = 1, 2, \ldots$, utilizando la desigualdad de Kolmogorov.

Etapa 2: Probaremos que $M_n \to 0$ casi seguramente, utilizando el lema de Borel-Cantelli.

Etapa 1: Para probar la primera afirmación notamos que:

$$P\left\{\max_{2^n<k\leq 2^{n+1}} \frac{|S_k|}{k} > \frac{1}{m}\right\} \leq P\left\{\max_{2^n<k\leq 2^{n+1}} |S_k| > \frac{2^n}{m}\right\}$$

(ya que dividir por 2^n en lugar de k agranda el máximo)

$$\leq P\left\{\max_{1\leq k\leq 2^{n+1}} |S_k| > \frac{2^n}{m}\right\} \leq \left(\frac{m}{2^n}\right)^2 \sum_{k=1}^{2^{n+1}} \mathrm{Var}(X_k)$$

Definamos el evento $A_{m,n} = \left\{M_n \geq \frac{1}{m}\right\}$. Entonces

$$\sum_{n=1}^{\infty} P(A_{m,n}) \leq \sum_{n=1}^{\infty} \left(\frac{m^2}{4^n} \sum_{k=1}^{2^{n+1}} \mathrm{Var}(X_k)\right)$$

Cambiando el orden de la suma deducimos que:

$$\sum_{n=1}^{\infty} P(A_{m,n}) \leq m^2 \sum_{k=1}^{\infty} \left(\sum_{n:2^{n+1}\geq k} \frac{\mathrm{Var}(X_k)}{4^n}\right)$$

$$= m^2 \sum_{k=1}^{\infty} \mathrm{Var}(X_k) \left(\sum_{n:2^{n+1}\geq k} \frac{1}{4^n}\right)$$

Ahora bien, sumando la serie geométrica:

$$\sum_{n=j}^{\infty} \frac{1}{4^n} = \frac{4}{3}\,\frac{1}{4^j}$$

En consecuencia:

$$\sum_{n:2^{n+1}\geq k} \frac{1}{4^n} = \sum_{n=j(k)}^{\infty} \frac{1}{4^n}$$

donde $j(k)$ cumple:

$$2^{j(k)} < k \leq 2^{j(k)+1}$$

En consecuencia:

$$\sum_{n:2^{n+1}\geq k} \frac{1}{4^n} = \frac{4}{3}\,\frac{1}{4^{j(k)}} \leq \frac{4}{3}\,\frac{4}{k^2} = \frac{16}{3k^2}$$

(pues $2^{j(k)} \geq \frac{k}{2}$).

Por lo que sustituyendo, concluimos que:

$$\sum_{n=1}^{\infty} P(A_{m,n}) \leq \frac{16m^2}{3} \sum_{k=1} \frac{\text{Var}(X_k)}{k^2} < +\infty$$

por la hipótesis.

Etapa 2: Por el lema de Borel-Cantelli, concluimos que, fijado m con probabilidad 1, sólo ocurren finitos de los eventos $A_{n,m}$. Vale decir que si

$$A_{m,\infty} = \left\{ \omega \in \Omega : M_n(\omega) \geq \frac{1}{m} \text{ para infinitos } n \right\} = \bigcap_{k \in N} \bigcup_{n \geq k} A_{m,n}$$

entonces $P(A_{m,\infty} = 0)$. Y entonces si consideramos el evento:

$$A = \{ \omega \in \Omega : M_n(\omega) \not\longrightarrow 0 \} = \bigcup_{m \in N} \infty \bigcap_{k \in N} \bigcup_{n \geq k} A_{m,n}$$

por la σ-aditividad, tenemos que: $P(A) = 0$. Concluimos que $M_n \to 0$ con probabilidad 1. □

Corolario G.2.2 *La ley fuerte de los grandes números,*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n} \xrightarrow{c.s.} 0$$

es válida para toda sucesión $(X_n)_{n \to N}$ de variables aleatorias independientes y uniformemente acotadas.

Prueba: Supongamos que $|X_n| \leq c$. Entonces $\text{Var}(X_n) \leq E[X_n^2] \leq c^2$, y entonces la hipótesis (G.2) es satisfecha. □

G.2.2. Algunos Lemas Preparatorios

Nuestro siguiente objetivo será probar que la ley fuerte de los grandes números es válida sin la restricción de acotación uniforme. Para ello necesitaremos algunos lemas preparatorios:

Lema G.2.3 (Criterio de Integrabilidad) *Sea $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria. Entonces $E[|X|] < +\infty$ ("X es integrable") si y sólo si*

$$\sum_{n=1}^{\infty} P\{|X| > n\} < +\infty$$

Prueba: Pongamos

$$A_0 = \{\omega \in \Omega : X(\omega) = 0\}$$
$$A_n = \{\omega \in \Omega : n-1 < |X| \leq n\}$$
$$A_\infty = \{\omega \in \Omega : X(\omega) = \pm\infty\}$$

Los eventos A_n (con $n \in \mathbb{N} \cup \{\infty\}$) forman una partición del espacio Ω. Notemos así mismo que bajo cualquiera de las dos condiciones del enunciado X es finita con probabilidad 1, es decir A_∞ tiene probabilidad cero. En consecuencia, por la σ-aditividad de la integral (de Lebesgue) respecto del conjunto [1]:

$$E[|X|] = \sum_{n=0}^{\infty} \int_{A_n} |X| \, dP$$

y por lo tanto:

$$\sum_{n=1} \int_{A_n} (n-1) \, dP \leq E[|X|] \leq \sum_{n=1} \int_{A_n} n \, dP$$

(Notamos que el término correspondiente a $n = 0$ se anula). Es decir que:

$$\sum_{n=1} (n-1) P(A_n) \leq E[|X|] \leq \sum_{n=1} nP(A_n)$$

o sea, teniendo encuenta que los A_n forman una partición (y que por lo tanto sus probabilidades suman 1):

$$\left(\sum_{n=1} nP(A_n) \right) - 1 \leq E[|X|] \leq \sum_{n=1} nP(A_n)$$

Deducimos pues que:

$$E[|X|] < +\infty \Leftrightarrow \sum_{n=1}^{\infty} nP(A_n) < +\infty$$

Para escribir esto de otra forma (y obtener la conclusión del enunciado), introduzcamos los eventos:

$$B_n = \{\omega \in \Omega : |X(\omega)| > n\}$$

Entonces $A_n = B_{n-1} - B_n$ y como $B_n \subset B_{n-1}$ deducimos que:

$$P(A_n) = P(B_{n-1}) - P(B_n)$$

En consecuencia,

$$E[|X|] < +\infty \Leftrightarrow \sum_{n=1}^{\infty} n \{P(B_{n-1}) - P(B_n)\} < +\infty \tag{G.3}$$

[1] Aquí presentamos una prueba usando la integral Lebesgue. Son posibles pruebas alternativas, por ej. usando la integral de Stieltjes. Ver Barry James

Ahora notamos que "sumando por partes":

$$\sum_{n=1}^{N} n\left\{P(B_{n-1}) - P(B_n)\right\} = 1(P(B_0) - P(B_1)) + 2(P(B_1) - P(B_2)) + \ldots$$

$$+N(P(B_{N-1}) - P(B_N) = P(B_0) + P(B_1) + P(B_2) + \ldots + P(B_{N-1}) - NP(B_N)$$

Es decir que:

$$\sum_{n=1}^{N} n\left\{P(B_{n-1} - P(B_n)\right\} = \sum_{n=0}^{N-1} P(B_n) - NP(B_N) \tag{G.4}$$

Ahora probaremos el enunciado: Si $E[|X|]$ es finita, por la desigualdad de Markov:

$$P(B_N) \leq \frac{1}{N} E[|X|]$$

En conscuencia, de (G.4) y (G.3), deducimos que la serie de términos no negativos:

$$\sum_{n=1}^{\infty} P(B_N)$$

tiene sumas parciales acotadas, y es por lo tanto convergente. Esto prueba una de las implicaciones del enunciado. Para probar la otra, supongamos que dicha serie es convergente. Entonces, por (G.4):

$$\sum_{n=1}^{N} n\left\{P(B_{n-1}) - P(B_n)\right\} \leq \sum_{n=1}^{N-1} P(B_0)$$

y en consecucnia por (G.3), $E[|X|] < +\infty$. □

Lema G.2.4 *Sea X una variable aleatoria con esperanza finita, y pongamos para cada n, $A_n = \{\omega \in \Omega : -n \leq |X(\omega)| \leq n\}$.Entonces:*

$$K = \sum_{n=1}^{\infty} \frac{1}{n^2}\, E[X^2 I_{A_n}] < +\infty$$

Prueba: Necesitamos la siguiente propiedad:

$$\sum_{n=j}^{\infty} \frac{1}{n^2} \leq \frac{2}{j} \tag{G.5}$$

Para establecer esta fórmula, notemos que para cada $n \in N$:

$$\frac{1}{n^2} \leq \frac{1}{n(n-1)} = \frac{1}{n-1} - \frac{1}{n}$$

En consecuencia, sumando esta serie telescópica, obtenemos que:

$$\sum_{n=j}^{\infty} \frac{1}{n^2} = \frac{1}{j^2} + \sum_{n=j+1} \frac{1}{n^2} \leq \frac{1}{j^2} + \sum_{n=j+1} \left(\frac{1}{n-1} - \frac{1}{n} \right)$$

$$= \frac{1}{j^2} + \frac{1}{j} < \frac{2}{j}$$

Volviendo a la prueba del lema, para cada $j \in N$, consideramos el evento:

$$B_j = \{\omega \in \Omega : j - 1 < |X(\omega)| \leq j\}$$

y

$$B_0 = \{\omega \in \Omega : X(\omega) = 0\}$$

Entonces:

$$A_n = \bigcup_{j=0}^{n} B_j \text{ (unión disjunta)}$$

En consecuencia:

$$E[X^2 I_{A_n}] = \sum_{j=0}^{n} E[X^2 I_{B_j}]$$

y por lo tanto:

$$K = \sum_{n=1}^{\infty} \frac{1}{n^2} \, E[X^2 I_{A_n}] = \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{j=-0}^{n} E[X^2 I_{B_j}]$$

Cambiando el orden de la suma (cosa que está permitida, ya que es una serie de términos no negativos):

$$K = \sum_{j=1}^{\infty} \sum_{n=j}^{\infty} \frac{1}{n^2} E[X^2 I_{B_j}]$$

Utilizando entonces la propiedad (G.5), vemos que:

$$K \leq \sum_{j=1}^{\infty} \frac{2}{j} E[X^2 I_{B_j}]$$

Ahora bien, cuando ocurre el evento B_j, $X^2 \leq j|X|$. Deducimos que,

$$K \leq 2 \sum_{j=1}^{\infty} E[|X| I_{B_j}] \leq 2E[|X|] < +\infty$$

ya que los eventos (B_j) forman una partición de Ω. □

G.2.3. La Segunda Ley Fuerte de Kolmogorov

Teorema G.2.5 *Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con $E[|X_i|] < +\infty$. Sea $\mu = E[X_i]$ entonces*

$$\frac{X_1 + X_2 + \ldots X_n}{n} \xrightarrow{c.s.} \mu$$

cuando $n \to +\infty$.

La prueba se basa en el *método de truncamiento*. Definimos unas nuevas variables aleatorias Y_n por:

$$Y_n = \begin{cases} X_n & \text{si} \quad |X_n| \leq n \\ 0 & \text{si} \quad |X_n| > n \end{cases}$$

Lema G.2.6 *Supongamos que se cumplen las hipótesis del teorema G.2.5 Las variables truncadas Y_n tienen las siguientes propiedades:*

i)
$$lim_{n\to+\infty} E[Y_n] = \mu$$

ii)
$$\sum_{n=1}^{\infty} \frac{Var(Y_n)}{n^2} < +\infty$$

iii) Con probabilidad 1, dado $\omega \in \Omega$ existe un $n_0 = n_0(\omega)$ tal que $X_n(\omega) = Y_n(\omega)$ para $n \geq n_0$.

Prueba: i): Como las X_n son idénticamente distribuidas:

$$E[Y_n] = E[X_n I_{\{|X_n|\leq 1\}}] = E[X_1 I_{\{|X_n|\leq 1\}}]$$

Ahora bien la secuencia de variables aleatorias: $X_1 I_{\{|X_n|\leq 1\}}$ está acotada por $|X_1|$:

$$|X_1 I_{\{|X_n|\leq 1\}}| \leq |X_1|$$

que es integrable por hipótesis. En consecuencia, por el teorema de convergencia mayorada:

$$E[Y_n] \to E[X_1] = \mu$$

ii): Nuevamente, como las X_n son idénticamente distribuidas

$$\text{Var}(Y_n) = \text{Var}(X_1 I_{\{|X_1|\leq n\}})$$

y la conclusión se sigue del lema G.2.4 pues X_1 es integrable.

iii): Consideramos el evento

$$A = \{\omega \in \Omega : \exists n_0 = n_0(\omega) \text{ tal que } \forall n \geq n_0 : X_n(\omega) = Y_n(\omega)\}$$

Queremos ver que $P(A) = 1$. Para ello consideramos los eventos,

$$A_n = \{\omega \in \Omega : X_n(\omega) \neq Y_n(\omega)\}$$

Entonces:

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P\{X_n \neq Y_n\} = \sum_{n=1}^{\infty} P\{|X_n| > n\} = \sum_{n=1}^{\infty} P\{|X_1| > n\} < +\infty$$

por el criterio de integrabilidad (lema G.2.3). En consecuencia, por el lema de Borel-Cantelli, con probabilidad 1, sólo ocurre un número finito de los sucesos A_n, es decir que $P(A) = 1$. □

Corolario G.2.7 *Si consideramos el evento*

$$B = \left\{\omega \in \Omega : \lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} |X_k(\omega) - Y_k(\omega)| = 0\right\}$$

tenemos que $P(B) = 1$

En efecto, como $A \subset B$ (donde A es el evento definido en la prueba anterior), y $P(A) = 1$ deducimos que $P(B) = 1$.

Necesitaremos también un lema (ejercicio) de análisis I:

Lema G.2.8 *Sea* $(\mu_k)_{k\in\mathbb{N}}$ *una sucesión de números reales tales que* $\mu_k \to \mu$ *cuando* $k \to +\infty$*, y pongamos* $z_n = \frac{1}{n}\sum_{k=1}^{n} \nu_k$ *entonces* $z_n \to \mu$ *cuando* $n \to +\infty$*.*

Podemos ahora concluir la prueba de la segunda ley fuerte de Kolmogorov (teorema G.2.5): consideramos el evento

$$C = \left\{\omega \in \Omega : \frac{X_1(\omega) + X_2(\omega) + \ldots + X_n(\omega)}{n} \to \mu \text{ cuando } n \to +\infty\right\}$$

Y consideramos también el evento:

$$D = \left\{\omega \in \Omega : \frac{Y_1(\omega) + Y_2(\omega) + \ldots + Y_n(\omega)}{n} - \overline{\mu} \to 0 \text{ cuando } n \to +\infty\right\}$$

siendo $\mu_k = E(Y_k)$ y $\overline{\mu} = -\frac{\mu_1+\mu_2+\ldots+\mu_n}{n}$.

En virtud del lema G.2.6, ii), vemos que las variables truncadas Y_n verifican las hipótesis de la primera ley fuerte de Kolmogorov (teorema G.2.1), en consecuencia $P(D) = 1$. Ahora bien, en virtud del lema G.2.8:

$$\frac{\mu_1 + \mu_2 + \ldots + \mu_n}{n} \to \mu$$

y en consecuencia: $B \cap D \subset C$. Pero como, $P(B) = P(D) = 1$, deducimos que $P(C) = 1$.

Esto concluye la prueba de la segunda ley fuerte de Kolmogorov.

Nota: Una demostración alternativa del teorema (G.2.5), que no depende de la desigualdad de Kolmogorov, se da en el artículo de N. Etemadi [Ete81].

Apéndice H

Compacidad para la convergencia en distribución

La siguiente condición, donde se pide que esto valga uniformemente en n, nos permitirá evitar la pérdida de masa en el infinito:

Definición H.0.1 *Sea (F_n) una sucesión de funciones de distribución. Diremos que (F_n) es **ajustada** [1] si dado $\varepsilon > 0$ existe $M_\varepsilon > 0$ tal que*

$$\limsup_{n\to+\infty} 1 - F_n(M_\varepsilon) + F(-M_\varepsilon) \leq \varepsilon$$

Si X_n es una sucesión de variables aleatorias con función de distribución F_n, esto es equivalente a decir que la sucesión (X_n) está acotada en probabilidad en el sentido de la proposición 8.1.8.

H.1. El Principio de Selección de Helly

Veremos en esta sección un teorema de compacidad para la convergencia en distribución.

Teorema H.1.1 *Supongamos que $(F_n)_{n\in\mathbb{N}}$ es una sucesión de funciones de distribución. Entonces existe una subsucesión F_{n_k} y una función $F : \mathbb{R} \to \mathbb{R}$ creciente y continua por la derecha, tal que*

$$\lim_{k\to+\infty} F_{n_k}(x) = F(x)$$

para todo punto de continuidad x de F.

[1] **tight** en inglés

Observación H.1.2 *La función límite F puede no ser una función de distribución. Por ejemplo si $a+b+c=1$, y*

$$F_n(x) = aI_{[n,+\infty)}(x) + bI_{[-n,+\infty)} + cG(x)$$

donde G es alguna función de distribución, entonces

$$F_n(x) \to F(x) = b + cG(x) \text{ cuando } n \to +\infty$$

y tenemos que

$$\lim_{x\to-\infty} F(x) = b, \qquad \lim_{x\to+\infty} F(x) = b + c = 1 - a$$

Luego se produce un fenómeno de "escape de masa al infinito".

Prueba: Utilizando el método diagonnal de Cantor (y la numerabilidad de los racionales), podemos construir una subsucesión F_{n_k} de F_n tal que

$$\lim_{k\to+\infty} F_{n_k}(q) = G(q)$$

exista para todo $q \in \mathbb{Q}$ (es decir todo q racional).

La función G puede no ser continua por la derecha, pero si definimos

$$F(x) = \text{ínf}\{G(q) : q \in \mathbb{Q}, q > x\}$$

obtenemos una función continua por la derecha pues

$$\begin{aligned}\lim_{x_n\downarrow x} F(x_n) &= \text{ínf}\{G(q) : q \in \mathbb{Q}, q > x_n \text{ para algún } n\} \\ &= \text{ínf}\{G(q) : q \in \mathbb{Q}, q > x\} = F(x)\end{aligned}$$

Para completar la prueba, consideremos un punto x de continuidad de F, y elijamos números racionales r_1, r_2, s tales que $r_1 < r_2 < x < s$ y

$$F(x) - \varepsilon < F(r_1) \le F(r_2) \le F(x) \le F(x) < F(x) + \varepsilon$$

Como $F_{n_k}(r_2) \to G(r_2) \ge G(r_1)$ y $F(n_k)(s) \to G(s) \le F(s)$, se deduce que si $k \ge k_0(\varepsilon)$,

$$F(x) - \varepsilon < F_{n_k}(r_2) \le F_{n_k}(x) < F_{n_k}(s) < F(x) + \varepsilon$$

luego $F_{n_k}(x) \to F(x)$. □

Teorema H.1.3 (Teorema de Prokhorov) *Supongamos que (F_n) es una sucesión de funciones de distribución. Entonces son equivalentes:*

i) (F_n) es ajustada.

ii) Para cualquier subsucesión (F_{n_k}) tal que

$$F_{n_k}(x) \to F(x)$$

para todo punto de continuidad de F siendo F continua por la derecha (como en el principio de selección de Helly), se tiene que F es una función de distribución, es decir que

$$F(-\infty) = 0, \quad F(+\infty) = 1 \tag{H.1}$$

Prueba: Supongamos primero que (F_n) es ajustada, y sea F_{n_k} una subsucesión que verifica ii). Elijamos $r < -M_\varepsilon$ y $s > M_\varepsilon$ puntos de continuidad de F, emtpmces

$$\begin{aligned} 1 - F(s) + F(r) &= \lim_{k\to+\infty} 1 - F_{n_k}(s) + F_{n_k}(r) \\ &\leq \limsup_{n\to+\infty} 1 - F_n(M_\varepsilon) + F_n(-M_\varepsilon) \leq \varepsilon \end{aligned}$$

Deducimos que:

$$\limsup_{x\to+\infty} 1 - F(x) + F(-x) \leq \varepsilon$$

y como ε es arbitrario. se deduce que F que se verifica (H.1).

Para probar el recíproco, supongamos que (F_n) no es ajustada. Entonces hay un $\varepsilon > 0$ y una subsucesión F_{n_k} tal que

$$1 - F_{n_k}(k) + F_{n_k}(-k) \geq \varepsilon$$

Utilizando el principio de selección de Helly (y pasando a una subsucesión) podemos suponer que $F_{n_k}(x) \to F(x)$ en los puntos de continuidad de F (donde F es continua por la derecha). Sean $r < 0 < s$ puntos de continuidad de F, entonces

$$1 - F(s) + F(r) = \lim_{k\to+\infty} 1 - F_{n_k}(s) + F_{n_k}(r) \geq \liminf_{k\to+\infty} 1 - F_{n_k}(k) + F_{n_k}(-k) \geq \varepsilon$$

Haciendo que $s \to +\infty$ y que $r \to +\infty$ deducimos que

$$1 - F(+\infty) + F(-\infty) \geq \varepsilon$$

Luego F no puede ser una función de distribución. □

H.2. Una versión más general del Teorema de Continuidad de Paul Levy

Teorema H.2.1 *Sea $(F_n)_{n\in\mathbb{N}}$ una sucesión de distribuciones de probabilidad, y sean*

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx}\, dF_n(x)$$

las correspondientes funciones características. Entonces

i) Si F_n converge débilmente a una distribución F, entonces

$$\varphi_n(t) \to \varphi(t) \quad \forall\, t \in \mathbb{R}$$

donde φ es la función característica de F.

ii) Recíprocamente, si

$$\varphi_n(t) \to \varphi(t) \quad \forall t \in \mathbb{R}$$

donde $\varphi(t)$ es una función continua en $t = 0$, entonces existe una distribución de probabilidad F tal que F_n converge débilmente a F.

Prueba: La afirmación i) es una consecuencia del corolario 9.2.3 aplicado a $\varphi(t) = e^{itx}$.

Para probar la afirmación recíproca ii), vamos a mostrar que la sucesión de funciones de distribución $(F_n)_{n\in\mathbb{N}}$ es ajustada. Esto será una consecuencia de la continuidad de $\varphi(t)$ en $t = 0$

Si $x \in \mathbb{R}$ y $\delta > 0$ entonces

$$1 \leq 2\left(1 - \frac{\operatorname{sen}(\delta x)}{\delta x}\right) = \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \cos(tx)\, dt \quad \text{si } |\delta x| > 2$$

Podemos consider variables aleatorias X_n con distribución F_n, tomar $x = X_n$ y tomar esperanzas para obtener

$$\begin{aligned} P\{|\delta X_n| > 2\} &\leq \frac{1}{\delta}\int_{-\delta}^{\delta} E[1 - \cos(tX_n)]\, dt \\ &= \frac{1}{\delta}\int_{-\delta}^{\delta} \operatorname{Re}[E(1 - \exp(itX_n)]\, dt \\ &= \frac{1}{\delta}\int_{-\delta}^{\delta} \operatorname{Re}[(1 - E[\exp(itX_n)]]\, dt \\ &= \frac{1}{\delta}\int_{-\delta}^{\delta} \operatorname{Re}[1 - \varphi_n(t)]\, dt \end{aligned}$$

Como

$$|\operatorname{Re}[1 - \varphi_n(t)]| \leq |1 - \varphi_n(t)]| \leq +|\varphi_n(t)| \leq 2,$$

por el teorema de convergencia mayorada vemos que

$$P\{|\delta X_n| > 2\} \to \frac{1}{\delta}\int_{-\delta}^{\delta} \operatorname{Re}[1 - \varphi(t)]\, dt$$

Entonces, dado $\varepsilon > 0$, por la continuidad de $\varphi(t)$ en $t = 0$, podemos encontrar un $\delta > 0$ tal que

$$|1 - \varphi(t)| < \varepsilon \quad \text{si } |t| < \delta$$

y por lo tanto

$$P\{|\delta X_n| > 2\} \to \frac{1}{\delta}\int_{-\delta}^{\delta} |1-\varphi(t)|\, dt \leq 2\varepsilon$$

Esto implica que la sucesión de funciones de distribución F_n es ajustada.

Continuemos entonces la demostración de la afirmación ii). Como F_n es ajustada, por el teorema de Prokhorov H.1.3, existen una subsucesión (F_{n_k}) y una función de distribución F tales que $F_{n_k}(x) \to F(x)$ si $x \in C(F)$ donde $C(F)$ es el conjunto de puntos de continuidad de F. Entonces por la primera parte del teorema $\varphi_n(t) \to \varphi_F(t)$, y por la unicidad de la función característica, deducimos que $\varphi_F = \varphi$. Además esto implica que la sucesión (F_n) tiene un único punto de acumulación F para la convergencia en distribución (es decir: no puede haber dos subsucesiones de F_n que converjan a distribuciones distintas).

Este último hecho implica que $F_n(x) \to F(x)$ para todo $x \in C(F)$. En efecto, si suponemos que no vale para alún $x_0 \in C(F)$, existirían un $\varepsilon > 0$ y una subsucesión (F_{n_k}) de (F_n) (no necesariamente la misma que consideramos antes), tales que

$$|F_{n_k}(x_0) - F(x_0)| > \varepsilon \tag{H.2}$$

Como (F_n) es ajustada, (F_{n_k}) también lo es. Luego, de nuevo por el teorema de Prokhorov, existe una subsucesión $(F_{n_{k_j}})$ de F_n tal que F_n converge en distribución a alguna distribución de probabilidades, que por lo que dijimos antes tiene que ser necesariamente F. Entonces $F_{n_{k_j}}(x_0) \to F(x_0)$, y esto contradice (H.2).Este absurdo provino de suponer que $F_n(x_0) \not\to F(x_0)$, por lo que $F_{n_k}(x_0) \to F(x_0)$. Esto vale para todo $x_0 \in C(F)$.

□

Bibliografía

Referencias Básicas del curso

[Álv15] Miguel Ángel García Álvarez. *Introducción a la teoría de la probabilidad I - Primer Curso)*. Fondo de Cultura Económica, 2015.

[Ash08] Robert B Ash. *Basic probability theory*. Courier Corporation, 2008. URL: http://www.math.uiuc.edu/~r-ash/BPT.html.

[Fel57] William Feller. An introduction to probability theory and its applications. 1957.

[Jam02] Barry R James. *Probabilidade: Um Curso em Nível Intermediário, 2a. edição*. 2002.

[Ren78] Alfred Renyi. *Teoría de Probabilidades*. Reverté, 1978.

[Roz73] Yu. Rozanov. *Procesos Aleatorios*. Editorial Mir, 1973.

[San55] Luis A. Santaló. *La Probabilidad y sus aplicaciones*. Ed. lberoamerieana, Buenos Aires., 1955.

[Yoh] Victor Yohai. *Notas del curso Probabilidades y Estadística*. URL: http://mate.dm.uba.ar/~vyohai/Notas%20de%20Probabilidades.pdf.

Combinatoria y Probabilidad Elemental

[Wil04] Miguel R Wilhelmi. *Combinatoria y probabilidad*. Grupo de Investigación en Educación Estadística, Universidad de Granada, 2004.

Referencias Básicas sobre estadística

[Ash07] Robert B Ash. *Lectures on Statistics*. 2007. URL: http://www.math.uiuc.edu/~r-ash/Stat.html.

[Jam19] Sreenivasa Rao Jammalamadaka. *Essential Statistics with Python and R.* University of California, Santa Barbara., 2019. URL: https://escholarship.org/uc/item/03w0n5g3.

[Moo69] F. Mood, A. Graybill. *Introducción a la Teoría de la Estadística.* Aguilar, editor, 1969.

Libros avanzados sobre probabilidad:

Nota: Se recomiendan estos libros para quienes hayan cursado análisis real y quieran profundizar en estos temas.

[Álv15] Miguel Ángel García Álvarez. *Introducción a la teoría de la probabilidad II- Segundo Curso.* Fondo de Cultura Económica, 2015.

[Bil79] Patrick Billingsley. *Probability and Measure.* John Willey &Sons, 1979.

[Dur19] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Artículos Originales Citados

[Ber41] Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

[CG42] Esseen Carl-Gustav. On the liapunoff limit of error in the theory of probability. *Arkiv for matematik, astronomi och fysik, A: 1–19*, 1942.

[Chi22] Calvin Wooyoung Chin. A short and elementary proof of the central limit theorem by individual swapping. *The American Mathematical Monthly*, 129(4):374–380, 2022.

[Eis17] Bennett Eisenberg. A very short proof that the sum of independent normal random variables is normal. *The College Mathematics Journal*, 48(2):137, 2017.

[Ete81] Nasrollah Etemadi. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):119–122, 1981.

[GP97] Henryk Gzyl and Jose Luis Palacios. The weierstrass approximation theorem and large deviations. *The American mathematical monthly*, 104(7):650–653, 1997.

[Tro59] HF Trotter. An elementary proof of the central limit theorem. *Archiv der Mathematik*, 10(1):226–234, 1959.

Referencias de Interés Histórico

[Ber12] Sergei Bernstein. Démonstration du théoreme de weierstrass fondée sur le calcul des probabilities. *Comm. Soc. Math. Kharkov*, 13:1–2, 1912.

[Bor09] M Émile Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27(1):247–271, 1909.

[Fra17] P Francesco. Cantelli. sulla probabilità come limite della frequenza. *Rom. Acc. L. Rend.(5)*, 26(1):39–45, 1917.

[McD05a] D. McDonald. The local limit theorem: A historical perspective. *Journal of the Iranian Statistical Society*, 4:73–86, 2005.

[McD05b] DR McDonald. The local limit theorem: a historical perspective. 2005.

Libros de Análisis Real

[Duo03] Javier Duoandikoetxea. Lecciones sobre las series y transformadas de fourier. 2003. URL: http://www.ugr.es/acanada/docencia/matematicas/analisisdefourier/Duoandikoetxeafourier.pdf.

[FC09] Adan J Corcho Fernandez and Marcos Petrucio de A Cavalcante. *Introducao a analise harmonica e aplicacoes*. IMPA, 2009. URL: https://impa.br/wp-content/uploads/2017/04/27CBM_11.pdf.

[FK75] AN Kolmogorov-SV Fomín and AN Kolmogorov. Elementos de la teoría de funciones y del análisis funcional. *Editorial Mir. Moscú*, 1975.

[HN01] John K. Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Company, 2001.

[WZ77] Richard Lee Wheeden and Antoni Zygmund. *Measure and integral*, volume 26. Dekker New York, 1977.

Otra bibliografía consultada para la elaboración de estas notas

[CJ82] Richard Courant and Fritz John. Introducción al cálculo y al análisis matemático: Vol. i. 1982.

Referencias relacionadas con paseos al azar y ecuaciones diferenciales

[Law10] Gregory F Lawler. *Random walk and the heat equation*, volume 55. American Mathematical Soc., 2010.

[LL10] Gregory F Lawler and Vlada Limic. *Random walk: a modern introduction*, volume 123. Cambridge University Press, 2010.

[Ros] Julio D Rossi. Tug-of-war games and pdes. course in maxwell centre for analysis and nonlinear pdes. edimburg. scotland. may 2010. URL: http://mate.dm.uba.ar/~jrossi/ToWandPDEs(2).pdf.

Otros artículos sobre temas mencioandos en estas notas

[BF02] Verónica Becher and Santiago Figueira. An example of a computable absolutely normal number. *Theoretical Computer Science*, 270(1-2):947–958, 2002.

[GS07] Andrew Granville and Kannan Soundararajan. Sieving and the erdős–kac theorem. In *Equidistribution in number theory, an introduction*, pages 15–27. Springer, 2007. URL: http://arxiv.org/abs/math/0606039.

[Sie17] Waclaw Sierpinski. Démonstration élémentaire du théorème de M. Borel sur les nombres absolument normaux et détermination effective d'une tel nombre. *Bulletin de la Société Mathématique de France*, 45:125–132, 1917.

Aplicaciones y Ejemplos de Datos Reales

[Ige20] Oluwatobiloba Ige. *Markov Chain epidemic models and parameter estimation.* PhD thesis, Phd. Thesis- Marshall University, 2020.

[KGS13] P Ravi Kumar, Alex KL Goh, and Ashutosh Kumar Singh. Application of markov chain in the pagerank algorithm. *Pertanika Journal of Science and Technology*, 21:541–554, 2013.

[LJ20] G Lakshmi and M Jyothi. Application of markov process for prediction of stock market performance. *International Journal of Recent Technology and Engineering*, 8(6):1516–1519, 2020.

[LLLT04] Sanboh Lee, HY Lee, IF Lee, and CY Tseng. Ink diffusion in water. *European journal of physics*, 25(2):331, 2004.

[MRR13] Cameron Appel Max Roser and Hannah Ritchie. Human height. *Our World in Data*, 2013. URL: https://ourworldindata.org/human-height.

Índice alfabético