Estimación, Probabilidad e Inferencia

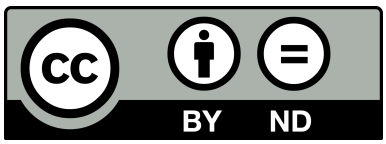
Construyendo modelos bajo incerteza

SERGIO DAVIS IRARRÁZABAL

Centro de Investigación en la Intersección
de Física de Plasmas, Materia y Complejidad (P^2mc)

Comisión Chilena de Energía Nuclear

Acerca de este libro



Prefacio

La teoría de la probabilidad, junto con la estadística, parecen ser algunas de las áreas de la matemática más usadas en la práctica, y simultáneamente son vistas por sus practicantes y estudiantes como algunas de las áreas más confusas y difíciles de asimilar en la Ciencia. ¿Cuál es la lógica y los principios que le dan coherencia a la idea de probabilidad? ¿Debo coleccionar un arsenal de técnicas a la espera de que puedan resultar útiles o existe una manera sistemática de construir la técnica que necesito *a la medida*?

Este libro surge a partir de un conjunto de apuntes propios diseñados para distintos cursos y coloquios de probabilidad bayesiana, teoría de información y simulación Monte Carlo, e intenta proponer un tratamiento autocontenido —casi podríamos decir *from scratch*— de las herramientas de la probabilidad y la inferencia, considerando esta última como el proceso de razonamiento bajo información incompleta. La probabilidad la entenderemos como una extensión de la lógica clásica, es decir, desde lo que se ha venido a llamar la interpretación *bayesiana* de la probabilidad. Sin embargo, hay una diferencia importante que nos separa de la tradición bayesiana: a diferencia de los tratamientos usuales cuyo punto de partida son **Los axiomas de Cox**, un punto central en este libro es que considera la idea de expectación (entendida como la estimación de una cantidad bajo conocimiento incompleto) como concepto fundamental, siendo la existencia de la probabilidad *deducida a partir de ésta*. El enfoque es entonces fuertemente centrado en el cálculo de expectaciones y en las identidades matemáticas que éstas satisfacen.

Entenderemos la estadística como un caso particular de la inferencia, en que se utilizan observaciones de una muestra para concluir propiedades de una población mayor. De esta forma, aunque no es nuestro foco principal, se deducirán algunos resultados clásicos como el método de mínimos cuadrados, el método de máxima verosimilitud, el concepto de histograma entre otros.

Público objetivo: Es mi aspiración que el contenido de este libro pueda ser cubierto en mayor o menor profundidad en un curso de un semestre para estudiantes de últimos años de carreras STEM[(1)]. Para la lectura de este libro es requisito únicamente un conocimiento de cálculo en varias variables y algo de álgebra lineal. El tratamiento matemático no es excesivamente formal, dada la formación en Física del autor, sin embargo hay demostraciones importantes expuestas de forma simplificada. En particular no se usa el lenguaje de la teoría de la medida.

Organización del libro: En el **Capítulo 1** comenzamos revisando las clases de razonamiento que llamaremos deducción e inferencia, e introducimos la idea de modelos actualizables bajo nueva información, para luego concentrarnos en el proceso deductivo, abordando elementos de la lógica y la deducción en el **Capítulo 2**. A continuación revisamos en el **Capítulo 3** algunos conceptos matemáticos como la delta de Dirac, la función escalón de Heaviside, la aproximación de Laplace, funciones generadoras y otros, conceptos necesarios para los capítulos posteriores. Una vez teniendo estas herramientas, en el **Capítulo 4** volvemos a la lógica pero esta vez usándola para describir variables discretas y continuas. En el **Capítulo 5** proponemos las reglas de la operación estimación, y a partir de ella deducimos la existencia de la probabilidad, para conectar nuestros desarrollos en el **Capítulo 6** con la formulación estándar de la probabilidad bayesiana. Continuamos en el **Capítulo 7** describiendo las propiedades de las distribuciones de probabilidad continuas y discretas, definiendo conceptos como media, varianza, distribución acumulada entre otros, para luego llegar a la definición y las propiedades de la distribución normal en el **Capítulo 8**.

En el **Capítulo 9** se muestra la aplicación del teorema de Bayes para el ajuste de parámetros de un modelo en función de datos observados. Luego de pasar por herramientas avanzadas en el **Capítulo 10** continuamos con el **Capítulo 11**, donde vemos criterios para elegir entre dos modelos y entramos en el área de la teoría de la información. Es en este capítulo donde se introduce el concepto de entropía. Inmediatamente después, el **Capítulo 12** introduce el principio de máxima entropía de Jaynes como herramienta para la construcción o actualización de modelos.

Finalmente en el **Capítulo 13** tratamos algunas aplicaciones a procesos donde el tiempo aparece de manera explícita, y revisamos algunos elementos del formalismo para estudiar caminatas al azar (*random walks*), para cerrar en el **Capítulo 14** introduciendo algunas ideas de métodos computacionales útiles en inferencia, como la generación de variables pseudoaleatorias y la simulación Monte Carlo.

(1) STEM: Ciencia, Tecnología, Ingeniería y Matemáticas (de la sigla en inglés).

Convenciones y notación: Escribiremos las integrales, tanto definidas como indefinidas, con el diferencial a la izquierda, como en

$$Z = \int_0^\infty \mathrm{d}x f(x).$$

En el caso de integrales multidimensionales sobre n variables $x_1, x_2, \ldots, x_n$, en lugar de la forma usual

$$Z = \int_\Omega \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \; f(x_1, \ldots, x_n)$$

donde Ω es el dominio de integración, usaremos la forma compacta

$$Z = \int_\Omega \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}),$$

es decir, $\mathrm{d}\boldsymbol{x}$ representa un elemento de volumen en n dimensiones.

Para denotar una familia de funciones $f(\boldsymbol{x})$ que depende de un conjunto de parámetros $\boldsymbol{\theta}$ escribiremos $f(\boldsymbol{x}; \boldsymbol{\theta})$. Para derivadas en varias dimensiones usaremos el operador nabla ∇, de forma que el gradiente de $f(\boldsymbol{x})$ es

$$\nabla f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right).$$

Usaremos la notación $a := b$ para indicar que a es igual a b *por definición*, mientras que $a \stackrel{?}{=} b$ indica que no sabemos aún si a es igual a b. Indicaremos que una expresión se evalúa en un punto particular usando paréntesis o corchetes con subíndices, como en

$$\left(x + \frac{\partial}{\partial x}(x^3) \right)_{x=a} = a + 3a^2.$$

Ocasionalmente usaremos la «notación flecha» $\mapsto$ para referirnos a una función como un *mapa* de una o más variables a una expresión. Por ejemplo, para referirnos a la función raíz cuadrada, en lugar de escribir $f(x) = \sqrt{x}$, lo cual nos obliga a inventar un nombre f, podremos referirnos a la función anónima

$$x \mapsto \sqrt{x},$$

donde el símbolo x es una *variable muda*. Cuando se trata de funciones de un argumento, podremos usar también la «notación punto» como en $g(5, \bullet)$ para referirnos a la función $z \mapsto g(5, z)$.

Usaremos marcas y símbolos en **púrpura** para anotar los desarrollos o demostraciones, como en

$$\begin{aligned} & (ax + b)^2 - b^2 \\ \hookrightarrow\ & a^2x^2 + 2axb + \cancel{b^2} - \cancel{b^2} \\ \hookrightarrow\ & a^2x^2 + 2axb \\ \hookrightarrow\ & ax(ax + 2b) \qquad \blacksquare \end{aligned}$$

donde el símbolo $\blacksquare$ reemplaza al clásico QED que indica el fin de una demostración[(2)].

[(2)] De la expresión en latín *quod erat demonstrandum*, o «que era lo que se quería demostrar».

Como estudiante, las fórmulas destacadas en recuadros deberían resultarle familiares —o al menos conocer de su existencia y dónde encontrarlas — al final de una lectura completa del libro. Similarmente los teoremas y definiciones se encuentran destacados y en un índice en la página XI.

Aspectos técnicos de esta edición: La tipografía del libro fue realizada con LaTeX bajo la distribución TeX Live (`https://tug.org/texlive`) en el sistema operativo Debian GNU/Linux (`https://debian.org`). Las figuras fueron generadas usando la librería Matplotlib (`https://matplotlib.org`) para Python y el paquete TikZ (`https://ctan.org/pkg/pgf`) de LaTeX. Se recomienda leer este documento `PDF` en un software habilitado para seguir hipervínculos, los cuales son usados en las referencias a ecuaciones, bibliografía y glosario.

Ubicación en la red: La versión más reciente de este libro está disponible en `http://sdavis.cl/epi/epi.pdf` de forma permanente.

Errata: La edición que está leyendo probablemente aún tiene muchos errores. Si desea reportarlos, puede hacerlo a `sergdavis@gmail.com` con el asunto «Errata EPI».

Piénselo muchas veces antes de imprimir este libro en papel.

Agradecimientos: En primer lugar quisiera agradecer a mi familia, gracias a cuyo esfuerzo pude tener la educación y el tiempo para eventualmente llegar a pensar en estas ideas.

Además de tal esfuerzo este libro se ha beneficiado de incontables conversaciones enriquecedoras con colegas, alumnas y alumnos durante los últimos diez años. En particular quisiera agradecer a Gonzalo Gutiérrez, Carlos Esparza, Yasmín Navarrete, Diego Contreras, Diego González, Humberto Loguercio, Joaquín Peralta, Ariel Caticha, Abiam Tamburrini, Felipe Moreno, Vivianne Olguín, Haridas Umpierrez, Constanza Farías, Ignacio Tapia, Boris Maulén, y muchas otras personas.

También me gustaría agradecer el apoyo del Centro de Investigación en la Intersección de Física de Plasmas, Materia y Complejidad (P^2mc) de la Comisión Chilena de Energía Nuclear (CCHEN).

SERGIO DAVIS
El Bosque, Santiago, Chile.
1 de marzo de 2022 *(versión 1.00)*

Índice general

FIGURAS

DEFINICIONES

TEOREMAS

RECUADROS

Lista de Notaciones

$f(\boldsymbol{x};\boldsymbol{\theta})$	Familia de funciones de $\boldsymbol{x}$ parametrizada por $\boldsymbol{\theta}$	IV
$\Theta(x)$	Función escalón de Heaviside evaluada en x	25
$\text{rect}(x;a,b)$	Función rectangular evaluada en x	27
$\arg\min_x f(x)$	Valor de x que hace que $f(x)$ sea mínimo	158
$\arg\max_x f(x)$	Valor de x que hace que $f(x)$ sea máximo	112
$\binom{n}{k}$	Coeficiente binomial para k elementos de un total de n	44
$a := b$	a se define como b	IV
$a \stackrel{?}{=} b$	Por determinar si a es igual a b	IV
$z \mapsto f(z)$	Función que toma z y retorna $f(z)$	IV
$f(x_0, \bullet)$	Función de un argumento, equivalente a $z \mapsto f(x_0, z)$	IV
$\delta(n,m)$	Delta de Kronecker de n y m	64
$\delta(x)$	Delta de Dirac evaluada en x	28
$\mathcal{A}[f]$	Funcional $\mathcal{A}$ evaluado en la función f	59
$\Big(A(x,y)\Big)_{x=a,y=b}$	Expresión $A(x,y)$ evaluada en $x=a, y=b$	IV
∇	Nabla, operador diferencial vectorial	IV
$\mathbb{H}$	Matriz hessiana	131
$\mathbb{J}_{uv}$	Matriz jacobiana	38
$\mathcal{J}_{uv}$	Determinante de la matriz $\mathbb{J}_{uv}$	38

$\mathbb{T}$	Verdadero	10
$\mathbb{F}$	Falso	10
$\neg A$	Negación de A	10
$A \vee B$	Disyunción de A y B	11
$A \wedge B$	Conjunción de A y B	11
$A \Rightarrow B$	A implica B	13
$A \Leftrightarrow B$	A es equivalente a B	15
$\mathrm{Q}(A)$	Función indicador de la proposición A	63
$\boldsymbol{X}$	Vector de variables	80
$\boldsymbol{\theta}$	Vector de parámetros	101
ω	Función de prueba	31
$X \sim M$	Variable x descrita por el modelo M	101
$\varnothing$	Estado de conocimiento no informativo	146
$\langle X \rangle_I$	Expectación (estimación) de X en el estado de conocimiento I	73
$P(A\|I)$	Probabilidad de A en el estado de conocimiento I	84
$\mathcal{L}_D(\boldsymbol{\theta})$	Función log-verosimilitud	146
$\mathcal{S}_A(I)$	Entropía de Shannon de A en el estado de conocimiento I	190
$\mathcal{S}(I_0 \to I)$	Entropía relativa desde I_0 hasta I	192
$D_{KL}(p\|\|q)$	Divergencia de Kullback-Leibler desde q hacia p	185

CAPÍTULO 1

Deducción e inferencia

> To teach how to live without certainty and yet without being paralysed by hesitation is perhaps the chief thing that philosophy, in our age, can do for those who study it.
>
> Bertrand Russell

Día a día observamos el mundo y de acuerdo a esas observaciones tomamos decisiones, desde las más nimias a las más trascendentes. Quisiéramos creer por supuesto que somos racionales, esto es, que tomamos decisiones de manera informada, sin prejuicios y sin que las emociones nos quiten el control. Sin embargo, excepto en casos muy particulares, nuestras decisiones son tomadas sin tener acceso a toda la información necesaria. En realidad navegamos a través de las actividades cotidianas en una permanente nube de incerteza, y a pesar de esto logramos funcionar adecuadamente, guiados por algo que podríamos llamar *intuición informada*[(1)].

Contrastemos ésto con el ideal del razonamiento infalible, alejado de toda emocionalidad, por ejemplo como podría ocurrir en el pensamiento matemático, en el cual nada es más cierto que un teorema pues una vez demostrado permanecerá cierto por siempre, independiente de emociones y prejuicios.

Como ejemplos de ambos casos, podemos pensar por un lado en el tipo de razonamiento que lleva a demostrar un resultado en matemáticas, y por otro lado el que lleva a resolver un crimen. ¿Existe diferencia entre ambos? ¿Cuál es esa diferencia? El propio Sherlock Holmes (Doyle 1960), brillante detective protagonista de las novelas de Arthur Conan Doyle, usa una metodología racional bastante cercana a la lógica pura, basada en la eliminación sistemática de posibilidades para resolver los más extraordinarios crímenes.

(1) Sobre las distintas definiciones de intuición y su conexión con la racionalidad, ver el libro de Bunge (2013).

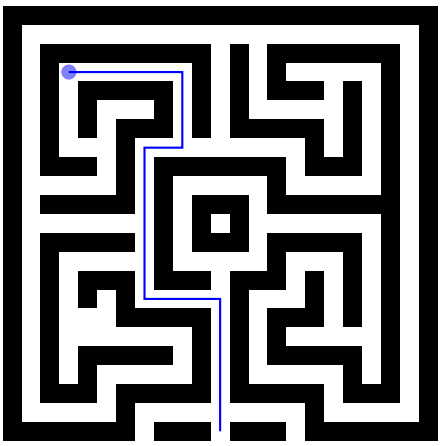


Figura 1.1: Un laberinto del cual podemos encontrar la salida usando la lógica. [figura adaptada de Hustad (2021)]

Al tipo de razonamiento matemático le llamaremos *deducción lógica* o simplemente **deducción**, mientras que nos referiremos al razonamiento de un detective, incluido Sherlock, como **inferencia**.

1.1 — Deducción: todas las cartas sobre la mesa

Con deducción nos referimos al procedimiento lógico con el cual vamos desde un conjunto de **premisas**, que forman nuestro punto de partida y que son supuestas ciertas de antemano, hasta una consecuencia de ellas, que llamaremos **conclusión**. Si las premisas se consideran ciertas, la deducción *asegura* que la conclusión es cierta.

> **Importante:** Ni siquiera la deducción puede determinar que algo es cierto de manera absoluta, sólo puede determinar que algo se sigue (es deducible) de ciertas premisas. Siempre es necesario un conjunto de premisas.

Algunos ejemplos donde tenemos toda la información, y podemos en principio encontrar una solución al problema usando el razonamiento deductivo, son: dibujar el camino con el que se escapa de un laberinto, como el que se muestra en la **Figura 1.1**, evaluar una expresión algebraica conociendo los valores de sus variables, resolver un puzzle *sudoku* o calcular una derivada simbólica de forma recursiva usando la *regla de la cadena*.

En matemática y a veces en física se emplean los siguientes nombres formales, que utilizaremos también en este libro.

Axioma o Postulado: Una afirmación fundamental que se supone cierta, sin necesitar justificación, de manera de deducir a partir de ella otras afirmaciones, es decir forma parte de las premisas de una deducción.

Teorema: Una afirmación que puede deducirse a partir de uno o más axiomas u otras afirmaciones demostradas previamente, es decir, es la conclusión de una deducción.

Lema: Un teorema que es demostrado como un paso intermedio para llegar a una posterior demostración más importante.

Corolario: Una afirmación que es consecuencia directa de una o más afirmaciones (ya sean axiomas, lemas o teoremas). También es una conclusión en el contexto de una deducción, aunque una mucho más inmediata que un lema o teorema.

1.2 — Inferencia: la mejor apuesta con lo que sabemos

Por supuesto, quisiéramos siempre poder usar la deducción, ya que ésta asegura la validez de la conclusión una vez supuesta la validez de las premisas. Sin embargo, en la gran mayoría de los casos esto es imposible y debemos optar por otros métodos de razonamiento.

Veamos algunas clases de situaciones en las cuales la deducción está fuera de nuestro alcance. En primer lugar, consideremos el caso cuando a pesar de tener toda la información necesaria, la solución del problema pasa por explorar un árbol gigantesco de posibilidades. Por ejemplo, cuando jugamos una partida de ajedrez o Go, aunque sabemos las reglas del juego y la meta a alcanzar, el número de posibilidades a evaluar es astronómico. Otro ejemplo similar es el cálculo de una integral simbólica, para el cual no existe un algoritmo predefinido y debe procederse a través de una búsqueda de patrones, similar a la exploración del árbol de movidas de una partida de ajedrez. Finalmente, otro problema matemático que no puede ser resuelto de forma deductiva es la descomposición de un entero suficientemente grande en un producto de números primos. Aunque no existe un algoritmo útil en la práctica[(2)], en principio este problema podría ser resuelto por fuerza bruta, ya que sabremos «de sólo mirarla» cuando una posible descomposición es la correcta. Sin embargo una vez más el problema es que el árbol de búsqueda es demasiado grande para ser manejado.

[(2)] No existe un algoritmo *clásico*, es decir no cuántico, suficientemente eficiente. En computación cuántica existe el algoritmo de Shor.

Además de estos casos existen otras situaciones donde no podemos usar la deducción, ya sea porque no poseemos la información de entrada suficiente, o porque no tenemos un criterio bien definido para decidir entre múltiples propuestas de solución. Esto se da, por ejemplo, en problemas de **predicción** de un evento en el futuro, tales como el saber en qué lugar ocurrirá el próximo terremoto de magnitud mayor que 8.5. A la inversa, tampoco podemos usar la deducción en problemas de **retrodicción**, esto es, reconstrucción de un evento en el pasado como la arqueología o incluso la ciencia forense. Pre-

guntas más abstractas como las referentes a alguna propiedad del Universo, por ejemplo la naturaleza de la energía oscura, o preguntas acerca de una estructura conceptual como podría serlo una teoría científica, tales como si existe una posible teoría que unifique la física cuántica y la relatividad general, también quedan fuera del alcance de la deducción, como también ocurre con otras preguntas de naturaleza filosófica, por ejemplo la existencia del libre albedrío.
En resumen, podríamos estar en una de tres situaciones:

(a) Existe un algoritmo bien definido, pero no es posible usarlo en la práctica por limitaciones de cálculo.

(b) No existe un algoritmo que asegure obtener la solución, y aunque es posible decidir si una propuesta de solución es la correcta, no es posible explorar todo el espacio de soluciones.

(c) Se tiene información incompleta, ya sea porque no hay datos de entrada suficientes o porque no se puede validar una propuesta de solución de forma satisfactoria.

Para las situaciones (a), (b) y (c), consideraremos el razonamiento mediante la inferencia en lugar de la deducción lógica. Llamaremos inferencia al procedimiento con el cual obtenemos **la mejor respuesta dada la información que poseemos**, siendo el objetivo de los primeros capítulos de este libro el llegar a descubrir las reglas de dicho procedimiento. El costo que pagamos al no tener toda la información es que se pierde la certeza de la deducción: ya nada puede asegurar que nuestras conclusiones por medio de la inferencia serán ciertas: sólo podremos afirmarlas con un cierto *grado de plausibilidad*. Esto es, la inferencia nos entrega un conjunto de posibilidades que podemos ordenar desde la más razonable hasta la menos razonable dada la información usada.

El razonamiento de Sherlock Holmes y otros detectives de la ficción es un caso de inferencia, como lo discute Kadane (2009), a pesar de que el mismo Sherlock lo denomine como deducción: en todos sus casos se consideran algunas de las hipótesis plausibles que explican lo ocurrido —¡hipótesis que surgen de su intuición o imaginación, de hecho!— las que luego se van eliminando conforme se obtienen nuevas *evidencias*. Un heredero legítimo del razonamiento detectivesco de Sherlock Holmes es presentado elegantemente en la serie de televisión Dr. House[(3)] (Irwin y Jacoby 2008). En ella, Gregory House, un brillante especialista en diagnóstico médico, resuelve junto a su equipo casos aparentemente inexplicables, usando el método llamado de *diagnóstico diferencial*, un proceso de eliminación sistemática que pasa por realizar exámenes, recoger la historia médica del paciente y cualquier otra evidencia relevante —médica o no—, con el fin de llegar al diagnóstico correcto. Cada pieza de información recogida hace más o menos plausibles las

[(3)] Los creadores de Dr. House admiten abiertamente que la inspiración para Gregory House fue Sherlock Holmes, con quien comparte muchas características como la capacidad analítica, el abuso de sustancias y una visión de la vida centrada en la racionalidad por sobre la empatía.

distintas explicaciones a los síntomas, las que literalmente van siendo anotadas y eliminadas del pizarrón hasta que se confirma un diagnóstico.

El razonamiento de Holmes o de House esencialmente busca explicaciones, pero además de éste existe otra categoría por excelencia donde se realiza inferencia: la que se centra en la predicción de eventos futuros y tiene que ver fundamentalmente con el azar y los **fenómenos aleatorios** (también denominados fenómenos estocásticos).

Consideraremos los fenómenos aleatorios como casos donde es imposible en la práctica adquirir la información necesaria porque hay detalles microscópicos del fenómeno que están permanentemente fuera de nuestro control, lo cual evita que puedan ser tratados de forma determinista. Por ejemplo, en el caso del lanzamiento de una moneda, si pudiéramos conocer la distribución de masa de ésta, su vector velocidad inicial con completa exactitud, y a la vez conocer todo el detalle de las corrientes de aire que la rodea y del campo gravitacional terrestre, podríamos en principio calcular el resultado del lanzamiento como cara o sello con completa exactitud.

> **Importante:** El que sea imposible adquirir la información para alcanzar determinismo puede deberse a limitaciones prácticas de medición experimental o del procesamiento de datos, pero también existe la posibilidad de limitaciones intrínsecas a nuestro actuar impuestas por las leyes de la física, como las que interpretamos a partir de la física cuántica. Este es un punto que está lejos de poder tener una respuesta definitiva, dado nuestro escaso entendimiento actual de los fundamentos de la teoría cuántica.

1.3 — Inferencia como modelos actualizables

Si bien existe un método bien establecido y exacto —la lógica— para realizar deducciones cuando se tiene toda la información, la pregunta natural es si existe un método único u óptimo para realizar inferencias cuando no tenemos acceso a la deducción. ¿Podemos unificar el razonamiento de Holmes, House y el razonamiento estadístico utilizado para la predicción, entre muchos otros, en un único método o conjunción armoniosa de métodos posible de ser aprendida, por ejemplo, durante una carrera científica?

La visión que este libro apoya es que la respuesta es afirmativa, y es la que se ha denominado razonamiento bayesiano. Este es un tipo de inferencia que se basa en el concepto de *probabilidad* —que nosotros desarrollaremos en capítulos posteriores— entendida como un grado de plausibilidad que asignamos a una hipótesis de forma subjetiva, aunque siempre basados en la información disponible.

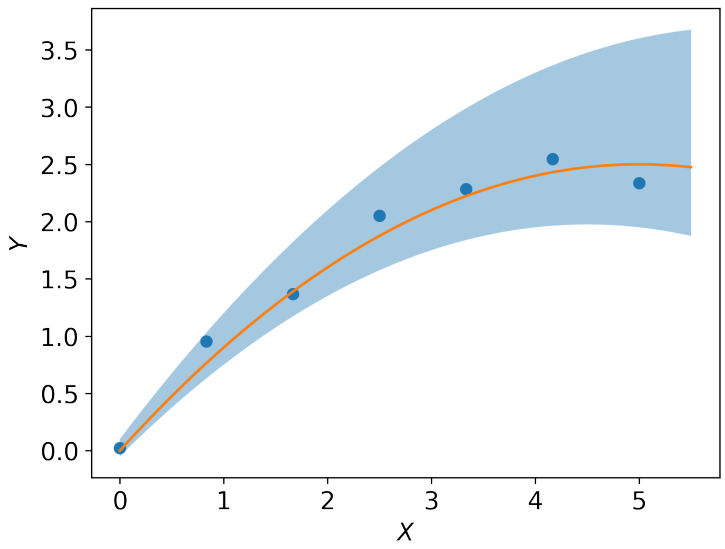


Figura 1.2: Un modelo para la dependencia $Y = f(X)$ de dos variables X e Y bajo incerteza. En este caso el modelo incluye tanto el valor más probable de Y para cada X (curva naranja) como la banda (en azul) de posibles curvas compatibles con los datos (círculos azules).

En nuestro recorrido por la inferencia comenzaremos introduciendo, por ahora sin una base matemática, dos conceptos fundamentales. Un **estado de conocimiento**, que denominaremos genéricamente con el símbolo I, representa el conjunto de nuestras premisas bajo las cuales intentamos hacer inferencia, ya sea para explicar un hecho ocurrido, predecir un evento futuro, o en general intentar contestar una pregunta.

Usando la información contenida en nuestro estado de conocimiento I podemos construir un **modelo**, que denotaremos con el símbolo M, el cual es una representación abreviada o un «resumen» que captura ciertos aspectos de la realidad que nos interesa describir. La meta de la inferencia es entonces tomar un estado de conocimiento I y a partir de éste construir el mejor modelo $M(I)$, de tal manera que para el mismo conocimiento este modelo óptimo sea el mismo. De esta forma estamos consiguiendo que los modelos sean *objetivos*, esto es, independientes de cualquier peculiaridad o idiosincrasia de quien los construye. En otras palabras, dos personas que manejan la misma información sobre un aspecto de la realidad, deberán llegar al mismo modelo de éste. Un ejemplo de modelo para la relación entre dos variables se muestra en la **Figura 1.2**.

Un modelo M será entonces capaz de dar *respuestas* correctas a ciertas preguntas que hacemos sobre la realidad, y por supuesto fallará al contestar otras preguntas, ya que por diseño todo modelo es incompleto[(4)]. Lo interesante de la idea de modelos es que éstos podrían contestar correctamente preguntas sobre las cuales no hemos incorporado información.

Un buen modelo es «compacto», captura la esencia de un fenómeno y para esto incorpora la información I más importante para la descripción de éste, dejando afuera todos los detalles irrelevantes. Como en el clásico chiste de primeros años en Física, suponer «una vaca esférica de masa m en el vacío...» es un modelo perfectamente válido para describir el impacto que produce en

[(4)] Si pudiera reproducir cada aspecto de la realidad, claramente ya no se trataría de un modelo sino de una copia exacta de la realidad.

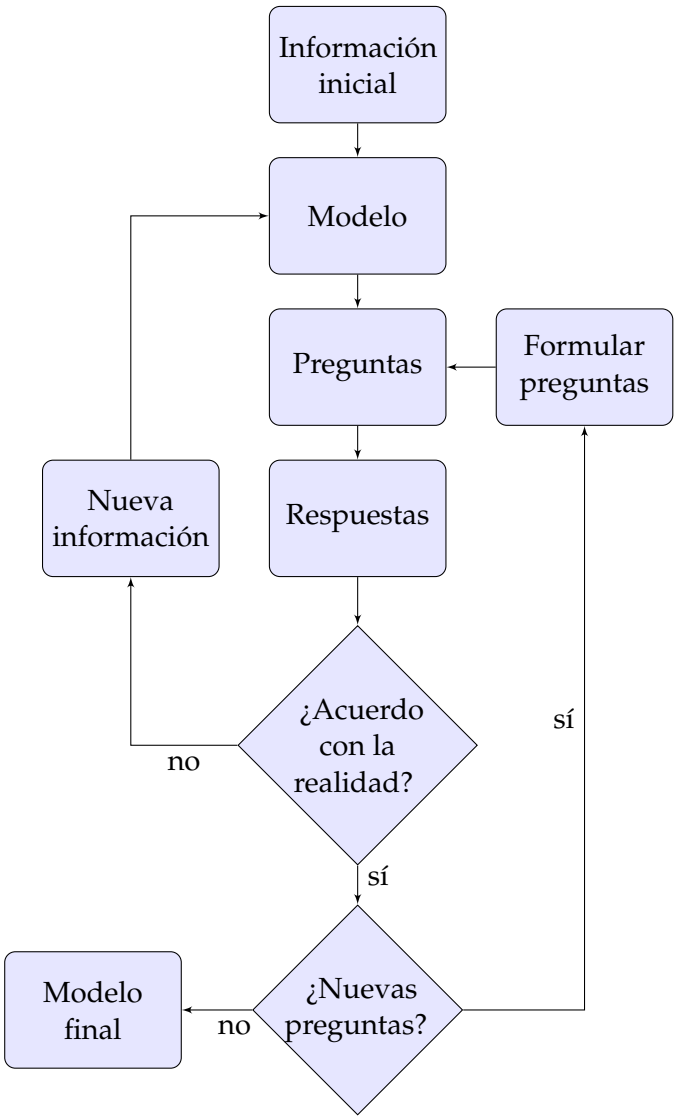


Figura 1.3: Esquema de funcionamiento de un procedimiento de inferencia en base a construcción y refinamiento de modelos. Un modelo, una vez construido, vive en este ciclo de refinamientos sucesivos hasta que decidimos que representa suficientemente bien los aspectos de la realidad que nos interesan.

el piso la vaca al caer desde un segundo piso, pero no para entender por qué este año la producción de leche fue menor que el año pasado.

Si un modelo M da una respuesta R que no está de acuerdo con la realidad, hemos adquirido nueva información que no está incluida en I, y tenemos la oportunidad de incorporarla, en cuyo caso pasamos de un estado de conocimiento I a uno diferente, I', en el que construimos un modelo actualizado $M' = M(I')$. Por supuesto, si la información nueva es irrelevante, podríamos preferir no incluirla en el modelo.

Así entramos en un ciclo de *refinamientos sucesivos* de un modelo, en el que vamos iterativamente incorporando nueva información, y nos mantenemos en este ciclo «desafiando» al modelo a contestar nuevas preguntas y contrastarlas con la realidad, hasta que decidimos que el modelo captura lo suficiente, como se muestra en la **Figura 1.3**.

En casos excepcionales, un modelo puede ser tan exitoso que logra establecerse como una *ley de la naturaleza*, a tal punto que llegamos a veces a confundir a los elementos que se definen en tal modelo con aspectos de la realidad. Pensemos por ejemplo en la masa m de un objeto, la cual medimos siempre indirectamente, ya sea observando la inercia del objeto al intentar ser acelerado, u observando cómo es afectado por (o como él afecta) la gravitación. Sin esas teorías físicas —que no son más que modelos que alcanzaron el *status* de leyes de la naturaleza— no existiría el concepto de masa[(5)]. Teniendo en cuenta esto, sólo nos queda preguntarnos: ¿es la masa real?

El efecto que nos hace confundir los modelos con la realidad que representan queda muy vívidamente reflejado en la frase «el mapa no es el territorio», y también en la llamada *falacia de la proyección mental* descrita por Jaynes (2003). En esta última, juicios sobre nuestros modelos son transformados en juicios sobre la realidad. Por ejemplo, en el caso de la física cuántica, pasamos de «la teoría no nos permite describir simultáneamente la posición y el momento de una partícula» a «la Naturaleza prohíbe que la posición y el momento de una partícula existan simultáneamente».

Considerando cada uno de los eslabones que unen los elementos de la **Figura 1.3**, vemos que una teoría de inferencia necesita varios componentes, en forma de reglas.

Recuadro 1.1 — Elementos de una teoría de inferencia
Una teoría de inferencia debería poseer reglas para

(1) Construir un modelo a partir de información inicial,

(2) Contestar preguntas acerca de la realidad usando el modelo,

(3) Comparar las respuestas que da el modelo con la realidad,

(4) Actualizar un modelo cuando hay desacuerdo con la realidad.

El cómo llegar a estas reglas sin ambigüedades será el objetivo de los siguientes capítulos, y para esto en primer lugar revisaremos los fundamentos de la lógica proposicional a continuación.

(5) El significado de la masa en el entendimiento moderno de la física es mucho más complejo. Así lo hace ver Frank Wilczek (2008), quien afirma que la masa es una propiedad emergente.

CAPÍTULO 2

Lógica proposicional

> How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?
>
> Sherlock Holmes, The Sign of Four

Comenzamos nuestro recorrido revisando los elementos básicos de la lógica proposicional, y desarrollando ejemplos de las reglas válidas que se usan para avanzar desde un conjunto de *premisas* hasta una *conclusión*. Posteriormente retomaremos el lenguaje de la lógica para conectarlo con la matemática tradicional a través de las *funciones indicador*, que veremos en el **Capítulo 4**.

El primer paso para razonar usando la lógica es escribir todas las afirmaciones conocidas, y las que se desea probar o refutar, como proposiciones lógicas, y conectar éstas usando operaciones lógicas que expresan líneas de razonamiento del tipo

Si esto es cierto, entonces se sigue que esto otro debe ser cierto.

2.1 — PROPOSICIONES LÓGICAS

Llamaremos *proposición lógica*, o simplemente proposición, a cualquier afirmación que, en principio, puede ser declarada verdadera o falsa. Algunos ejemplos de proposiciones válidas son

$A =$ "La capital de Francia es Paris", (2.1a)

$B =$ "El cuadrado de 6 es 49", (2.1b)

$C =$ "Está lloviendo ahora", (2.1c)

$D =$ "La teoría de la Relatividad General de Einstein es correcta", (2.1d)

$E =$ "La Ilíada y la Odisea fueron escritas por un único autor.". (2.1e)

Podemos suponer a las etiquetas A, B, C, D, E en el ejemplo anterior como *variables booleanas*, esto es, que sólo pueden tomar uno de dos valores: verdadero o falso.

Denotaremos simbólicamente el valor verdadero como $\mathbb{T}$ y el valor falso como $\mathbb{F}$, con lo que podemos decir $A \in \{\mathbb{T}, \mathbb{F}\}$. Con el conocimiento general que tenemos referente a geografía y aritmética, podemos decir que la proposición A es cierta, mientras que la proposición B es falsa, por lo tanto, simbólicamente escribiremos

$$A = \mathbb{T},$$
$$B = \mathbb{F}.$$

Con esto estamos especificando el *valor de verdad* de A y de B. En esta notación, dos expresiones lógicas conectadas con $=$ tienen el mismo valor de verdad.

Respecto a las proposiciones C, D y E, vemos que determinar su valor de verdad no es tan sencillo como en el caso de A y B pero es *en principio* posible. La proposición C requiere especificar dónde se sitúa geográficamente quien intenta evaluarla, y a qué instante en específico se refiere «ahora». Por otro lado, el consenso en la comunidad de la física es que la proposición D es cierta, sin embargo entenderla requiere un contexto altamente técnico, mientras que la proposición E es directamente entendible pero no existe un consenso claro entre los historiadores.

De la misma manera como ocurre en el caso de las identidades matemáticas, por ejemplo

$$(a + b)^2 = (a^2 + 2ab + b^2),$$

donde la igualdad es cierta para cualquier valor que tomen las variables a y b, existen *identidades lógicas* que son válidas para cualquier valor de verdad de las proposiciones involucradas. Por ejemplo, la regla

La negación de la negación de una proposición es la misma proposición.

es una identidad lógica, que podremos escribir de manera simbólica una vez introducidas las *operaciones lógicas*, a continuación.

2.2 — Operaciones lógicas

Para llevar a cabo el proceso de deducción se necesita combinar las distintas proposiciones lógicas que forman parte de nuestras premisas con el fin de llegar a una conclusión, válidamente establecida. Para esto, existen las operaciones lógicas básicas: *negación*, *conjunción* y *disyunción*, que juntas constituyen un *álgebra* para las proposiciones lógicas.

Escribiremos $\neg A$ para referirnos a la **negación** de A, conocida también como la operación `not`. Su definición se entrega a través de su *tabla de verdad*, que se muestra en la **Tabla 2.1** (izquierda), y su efecto es simplemente transformar $\mathbb{T}$ en $\mathbb{F}$ y viceversa.

A	$\neg A$
$\mathbb{T}$	$\mathbb{F}$
$\mathbb{F}$	$\mathbb{T}$

A	B	$A \wedge B$	$A \vee B$
$\mathbb{F}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{F}$
$\mathbb{F}$	$\mathbb{T}$	$\mathbb{F}$	$\mathbb{T}$
$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$
$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$

Tabla 2.1: Las operaciones de la lógica: Negación ($\neg$), conjunción ($\wedge$) y disyunción ($\vee$).

La única identidad lógica que podemos construir sólo usando la negación es la que mencionamos anteriormente, que nos dice que la negación de la negación de A es A,

$$\neg(\neg A) = A. \tag{2.2}$$

La operación negación es una operación lógica *unaria*, es decir, recibe un sólo argumento, mientras que existen dos operaciones *binarias*, esto es, que reciben dos argumentos, las cuales formarán la base de nuestra álgebra booleana. La **conjunción** entre A y B, también conocida como `and`, y que denominaremos $A \wedge B$, representa la afirmación

A y B son ambas ciertas.

Por otro lado la **disyunción** entre A y B, también conocida como `or`, y que denominaremos $A \vee B$, representa la afirmación

A es cierta o B es cierta, o ambas son ciertas.

La conjunción y la disyunción se encuentran definidas a través de sus tablas de verdad en la **Tabla 2.1** (derecha). Ambas son operaciones conmutativas,

$$A \wedge B = B \wedge A, \tag{2.3a}$$
$$A \vee B = B \vee A, \tag{2.3b}$$

y asociativas,

$$A \wedge (B \wedge C) = (A \wedge B) \wedge C = A \wedge B \wedge C, \tag{2.4a}$$
$$A \vee (B \vee C) = (A \vee B) \vee C = A \vee B \vee C, \tag{2.4b}$$

por lo que en una secuencia de conjunciones o disyunciones se puede omitir los paréntesis. De la misma forma, conjunción y disyunción son distributivas una sobre la otra, esto es

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C), \tag{2.5a}$$
$$A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C). \tag{2.5b}$$

Hasta aquí podemos imaginar que la negación se comporta como el análogo del inverso aditivo $a \mapsto -a$, la conjunción como el análogo de la multiplicación $a, b \mapsto a \cdot b$ y la disyunción como el análogo de la adición, $a, b \mapsto a + b$.

De hecho podemos establecer que $\mathbb{F}$ es el elemento neutro de la disyunción, ya que

$$A \vee \mathbb{F} = A, \tag{2.6}$$

y $\mathbb{T}$ el elemento neutro de la conjunción, dado que

$$A \wedge \mathbb{T} = A. \tag{2.7}$$

Sin embargo, las siguientes propiedades muestran un comportamiento diferente. La *ley del tercero excluido* es la identidad

$$\boxed{A \vee (\neg A) = \mathbb{T},} \tag{2.8}$$

que nos indica que A debe ser cierta, o debe ser falsa, y no hay otras alternativas. Veámoslo caso a caso: si $A = \mathbb{F}$, entonces $\neg A = \mathbb{T}$ y tenemos $\mathbb{F} \vee \mathbb{T} = \mathbb{T}$, mientras que si $A = \mathbb{T}$, se tiene $\neg A = \mathbb{F}$ y entonces $\mathbb{T} \vee \mathbb{F} = \mathbb{T}$. En ambos casos hay exactamente un factor que es verdadero, haciendo la expresión completa verdadera.

> **Importante:** En el caso en que tenemos una identidad del tipo (2.8), donde una expresión lógica es igual a $\mathbb{T}$, podemos afirmar simplemente dicha proposición lógica, es decir, podemos escribir directamente
>
> $$A \vee (\neg A).$$

Las *leyes de De Morgan*[(1)] nos muestran cómo se distribuye la negación sobre la conjunción o la disyunción,

$$\neg(A \wedge B) = (\neg A) \vee (\neg B), \tag{2.9a}$$
$$\neg(A \vee B) = (\neg A) \wedge (\neg B), \tag{2.9b}$$

de manera que el efecto no es sólo distribuir la negación sino que la conjunción cambia a disyunción, y viceversa. Podemos por tanto usar las leyes de De Morgan para eliminar una de las operaciones binarias en favor de la otra.

Una regla mnemotécnica útil para (2.9a) es recordar que la única manera de que $A \wedge B$ sea falso es que A sea falso o que B sea falso. De la misma manera, una regla útil para recordar (2.9b) es que la única manera de que $A \vee B$ sea falso es que tanto A como B sean falsas.

Si negamos ambos lados de (2.8) y usamos (2.9b), obtenemos el llamado *principio de no contradicción*,

$$\boxed{A \wedge (\neg A) = \mathbb{F},} \tag{2.10}$$

que nos dice que es imposible que A sea cierta y a la vez sea falsa. De nuevo, viéndolo caso a caso tenemos $\mathbb{T} \wedge \mathbb{F} = \mathbb{F}$ si A es verdadera, mientras que obtenemos $\mathbb{F} \wedge \mathbb{T} = \mathbb{F}$ si A es falsa.

(1) Gracias a las leyes de De Morgan es posible construir la lógica sólo con dos de las tres operaciones, por ejemplo, sólo con negación y disyunción.

A	B	$A \Rightarrow B$
$\mathbb{F}$	$\mathbb{F}$	$\mathbb{T}$
$\mathbb{F}$	$\mathbb{T}$	$\mathbb{T}$
$\mathbb{T}$	$\mathbb{F}$	$\mathbb{F}$
$\mathbb{T}$	$\mathbb{T}$	$\mathbb{T}$

Tabla 2.2: Implicación lógica ($\Rightarrow$). La expresión $A \Rightarrow B$ es sólo falsa cuando A es cierta y B es falsa.

Podemos definir una tercera operación lógica binaria, la *implicación lógica* (o simplemente implicación) $A \Rightarrow B$, que leeremos como A *implica* B, en función de las operaciones conjunción y disyunción como

$$A \Rightarrow B := (\neg A) \vee B. \tag{2.11}$$

La implicación sólo es falsa cuando A es cierta pero B es falsa, como se ve en la **Tabla 2.2**, y por tanto representa la afirmación

Si A es cierta entonces B es cierta.

Notemos que es posible reescribir $A \Rightarrow B$ de acuerdo a

$$\begin{aligned} & A \Rightarrow B \\ & \hookrightarrow (\neg A) \vee B \\ \text{usando (2.2)} \quad & \hookrightarrow (\neg A) \vee (\neg[\neg B]) \\ \text{usando (2.9a)} \quad & \hookrightarrow \neg(A \wedge \neg B) \end{aligned} \tag{2.12}$$

es decir, la implicación puede entenderse como la negación de la frase «A ocurre pero B no». Notemos también que la implicación $A \Rightarrow B$ es verdadera para $A = \mathbb{F}$ independientemente del valor de B, propiedad que se conoce como el *principio de explosión*: «de una falsedad se sigue cualquier cosa»[(2)].

Ejemplo 2.2.1. *La afirmación*

$$x > 5 \Rightarrow x > 3 \tag{2.13}$$

es siempre cierta, para cualquier valor de x. De hecho escribiendo (2.13) como

$$(x \leq 5) \vee (x > 3) \tag{2.14}$$

es claro que si x es mayor que 5, sabemos que automáticamente x es mayor que 3 y luego (2.13) es cierta por la segunda proposición en (2.14). Por otro lado, si x es menor o igual que 5, la primera proposición en (2.14) es cierta y luego (2.13) es cierta sin importar si x es mayor, menor o igual a 3.

[(2)] En latín, *ex falso quodlibet*.

> **Importante:** ¡La implicación lógica no necesariamente significa una relación de causa y efecto! Del hecho que B ocurra *sin falta* cuando A ocurre no se sigue que A sea la causa de B. Bien podría ser que A y B son causados por un tercero C, o que no tenga sentido ningún tipo de relación causal entre ellos.

También es posible reescribir $A \Rightarrow B$ como

$$\begin{aligned} & A \Rightarrow B \\ \hookrightarrow\ & (\neg A) \vee B \\ \hookrightarrow\ & B \vee (\neg A) \\ \hookrightarrow\ & (\neg B) \Rightarrow (\neg A) \end{aligned} \tag{2.15}$$

con lo que vemos que si A implica B, entonces puede concluirse que la negación de B implica la negación de A. Es importante notar que si $A \Rightarrow B$ es cierto, de allí **no se sigue** que $\neg A \Rightarrow \neg B$.

La manera correcta de sacar conclusiones a partir de una implicación está plasmada en dos reglas de deducción, cuyo origen es muy anterior a la formulación simbólica de la lógica.

Recuadro 2.1 — Dos reglas de deducción clásicas

***Modus ponens*:** Si R es cierto y $R \Rightarrow S$ es cierto, entonces S es cierto.

Esto es, $(R \Rightarrow S) \wedge R = \mathbb{T}$ se reduce a $S = \mathbb{T}$.

***Modus tollens*:** Si S es falso y $R \Rightarrow S$ es cierto, entonces R es falso.

Esto es, $(R \Rightarrow S) \wedge (\neg S) = \mathbb{T}$ se reduce a $R = \mathbb{F}$.

Adicionalmente, utilizando la implicación podemos definir tres términos ampliamente usados en matemáticas.

***A* es condición necesaria para *B*:** B es sólo cierta si se cumple A.

Es decir,

$$B \Rightarrow A.$$

De acuerdo al *modus tollens*, también se cumple

$$\neg A \Rightarrow \neg B.$$

***A* es condición suficiente para *B*:** Si A sea cierta entonces B es cierta.

Es decir,

$$A \Rightarrow B.$$

***A* si y sólo si *B*:** Si A se cumple, entonces B se cumple, y viceversa.

Esto es,

$$(A \Rightarrow B) \wedge (B \Rightarrow A).$$

En este caso también diremos que A es equivalente a B, usando la notación

$$A \Leftrightarrow B.$$

Ejemplo 2.2.2. *Veamos algunos ejemplos para ilustrar estas tres definiciones.*

(a) Tener 18 años o más es una condición necesaria para poder tener licencia de conducir. Si se tiene licencia de conducir, entonces se tiene 18 años o más. Esta no es una condición suficiente, porque además de tener 18 años o más se requiere haber aprobado ciertas pruebas psicomotoras.

(b) Que esté lloviendo es condición suficiente para que la tierra esté mojada. Notemos que no es condición necesaria, ya que por ejemplo si el patio fue regado también la tierra estará mojada, aunque no haya llovido.

(c) Tener nota igual o superior a la nota mínima (por ejemplo 4.0) es condición necesaria y suficiente para aprobar un curso. Podemos decir que el curso se aprueba si y sólo si se obtiene una nota igual o superior a la mínima, y también podremos decir que aprobar el curso es equivalente a tener nota igual o superior a la mínima.

2.3 — MÉTODOS DE DEMOSTRACIÓN

Ahora que conocemos y dominamos las operaciones de la lógica, veamos algunos métodos utilizados para demostrar la verdad o falsedad de una proposición a partir de un conjunto de premisas.

2.3.1 La regla de resolución

Consideremos dos proposiciones R_1 y R_2, dadas por

$$A \vee C, \qquad (R_1)$$
$$B \vee (\neg C), \qquad (R_2)$$

donde la primera contiene una proposición C y la segunda su negación, $\neg C$. Reescribiendo R_1 y R_2 como implicaciones, tenemos

$$(\neg C) \Rightarrow A, \qquad (R_1)$$
$$C \Rightarrow B. \qquad (R_2)$$

Si ahora suponemos que tanto R_1 como R_2 son ciertas, entonces tenemos dos casos. Cuando C es cierta, R_2 nos dice que B debe ser cierta, mientras que si C es falsa, R_1 nos dice que A debe ser cierta. El resultado final, independiente de C, es que si R_1 y R_2 son ciertas entonces $A \vee B$ debe ser cierta. Tenemos entonces el siguiente teorema, denominado la *regla de resolución*.

Teorema 2.1 — Regla de resolución
Si para tres proposiciones A, B y C se cumple la identidad

$$(A \vee C) \wedge (B \vee (\neg C)) = \mathbb{T}, \tag{2.16}$$

entonces también se cumple la identidad

$$A \vee B = \mathbb{T}, \tag{2.17}$$

independiente del valor de verdad de C.

Este resultado puede ser entendido como si C cancelara $\neg C$ de una manera particular:

$$\begin{aligned}(A \vee \cancel{C}) \wedge (B \vee \cancel{(\neg C)}) \\ \hookrightarrow A \vee B.\end{aligned}$$

Ahora generalizaremos esto a dos proposiciones lógicas compuestas ${R_1}'$ y ${R_2}'$, dadas por

$$A_1 \vee C \vee A_2, \tag{${R_1}'$}$$
$$B_1 \vee (\neg C) \vee B_2, \tag{${R_2}'$}$$

Como podemos redefinir $A := (A_1 \vee A_2)$ y $B := (B_1 \vee B_2)$, y escribir

$$A_1 \vee C \vee A_2 = (A_1 \vee A_2) \vee C = A \vee C, \tag{${R_1}'$}$$
$$B_1 \vee (\neg C) \vee B_2 = (B_1 \vee B_2) \vee (\neg C) = B \vee (\neg C), \tag{${R_2}'$}$$

la regla de resolución asegura que si

$$(A_1 \vee C \vee A_2) \wedge (B_1 \vee (\neg C) \vee B_2) = \mathbb{T} \tag{2.18}$$

es una identidad, entonces

$$A_1 \vee A_2 \vee B_1 \vee B_2 = \mathbb{T} \tag{2.19}$$

también es una identidad. La regla de resolución encapsula todo el proceso de deducción en la lógica proposicional, y puede ser usada como una regla única para demostrar teoremas comenzando desde un conjunto de axiomas. De hecho, las reglas del *modus ponens* y el *modus tollens* son casos particulares de la regla de resolución. El primero es la aplicación de resolución en la expresión

$$R \wedge (R \Rightarrow S) = (\cancel{R} \vee \mathbb{T}) \wedge (\cancel{\neg R} \vee S) = \mathbb{T} \vee S = S, \tag{2.20}$$

mientras que el segundo es la aplicación de resolución en

$$(\neg S) \wedge (R \Rightarrow S) = (\cancel{\neg S} \vee \mathbb{T}) \wedge (\neg R \vee \cancel{S}) = \mathbb{T} \vee \neg R = \neg R. \tag{2.21}$$

2.3.2 Inducción matemática

Se desea demostrar $A_n = \mathbb{T}$ para todo entero $n \geq 1$, es decir,

$$A_1 \wedge A_2 \wedge A_3 \wedge \ldots = \mathbb{T}. \tag{2.22}$$

La técnica de inducción matemática para demostrar (2.22) consiste en lo siguiente.

(1) Se demuestra que A_1 es cierta.
(2) Se demuestra que A_n es cierta suponiendo que A_{n-1} lo es. Esto es, se demuestra que la implicación $(A_{n-1} \Rightarrow A_n)$ es cierta para cualquier entero $n \geq 1$.

Si ambos pasos (1) y (2) son exitosos entonces (2.22) es cierta, esto es, A_n es cierta para todo entero $n \geq 1$. Para ver por qué este argumento de inducción matemática es correcto, basta aplicar *modus ponens* iterativamente comenzando desde A_1,

$$\begin{aligned}
&\text{de } A_1 \wedge (A_1 \Rightarrow A_2) = \mathbb{T} \text{ se sigue que } A_2 = \mathbb{T},\\
&\text{de } A_2 \wedge (A_2 \Rightarrow A_3) = \mathbb{T} \text{ se sigue que } A_3 = \mathbb{T},\\
&\text{de } A_3 \wedge (A_3 \Rightarrow A_4) = \mathbb{T} \text{ se sigue que } A_4 = \mathbb{T},\\
&\text{de } A_4 \wedge (A_4 \Rightarrow A_5) = \mathbb{T} \text{ se sigue que } A_5 = \mathbb{T},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots
\end{aligned}$$

Veamos un ejemplo concreto de cómo utilizar esta técnica para demostrar una propiedad de la operación derivada.

Ejemplo 2.3.1. *Demostremos por inducción matemática que*

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^n) = nx^{n-1} \tag{2.23}$$

para $n \geq 1$. Para esto, en primer lugar verificamos que (2.23) es cierta para $n = 1$, ya que

$$\frac{\mathrm{d}x}{\mathrm{d}x} = 1.$$

El siguiente paso es suponer *que (2.23) es cierta para $n - 1$, es decir, suponer que*

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^{n-1}) = (n-1)x^{n-2}, \tag{2.24}$$

y a partir de esta suposición, probar que (2.23) es cierta para n. Esto lo conseguimos separando $x^n = x \cdot x^{n-1}$ y usando la regla de Leibniz para desarrollar

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^n) = \frac{\mathrm{d}}{\mathrm{d}x}(x \cdot x^{n-1}) = x^{n-1} + x(n-1)x^{n-2} = nx^{n-1}. \tag{2.25}$$

2.3.3 Prueba por contradicción

Nuestro problema es demostrar que, a partir de n premisas $A_1, A_2, \ldots, A_n$ se sigue una conclusión B, y para realizar este tipo de demostración, supondremos en primer lugar el conjunto de todas premisas, llamémoslo R, como cierto,

$$R := A_1 \wedge A_2 \wedge \ldots \wedge A_n = \mathbb{T}, \tag{2.26}$$

para luego agregar como cierta la **negación de la conclusión**, $\neg B$, esperando con ello demostrar que esta conjunción ampliada de $n+1$ proposiciones lleva a una contradicción. Es decir, buscamos probar que

$$A_1 \wedge A_2 \wedge \ldots \wedge A_n \wedge (\neg B) = \mathbb{F}. \tag{2.27}$$

Este razonamiento es correcto, dado que sustituyendo (2.26) en (2.27) tenemos

$$\begin{aligned} & R \wedge (\neg B) = \mathbb{F} \\ \hookrightarrow\ & \mathbb{T} \wedge (\neg B) = \mathbb{F} \\ \hookrightarrow\ & B = \mathbb{T} \end{aligned}$$

Para lograr probar (2.27) son válidas todas las reglas de la lógica que hemos visto, en particular puede usarse repetidamente *modus ponens*, *modus tollens*, resolución o las identidades como (2.8) y (2.10).

Ejemplo 2.3.2. *Demostremos la regla del modus ponens: si R es cierto, y $R \Rightarrow S$ es cierto, entonces se sigue que S es cierto. Tenemos dos premisas,*

$$\begin{aligned} A_1 &= R, \\ A_2 &= (R \Rightarrow S), \end{aligned}$$

y nuestra conclusión a demostrar, S. Formamos entonces la expresión

$$\begin{aligned} & R \wedge (R \Rightarrow S) \wedge (\neg S) \\ \hookrightarrow\ & R \wedge ([\neg R] \vee S) \wedge (\neg S) \\ \hookrightarrow\ & ((R \wedge [\neg R]) \overset{\mathbb{F}}{\vee} (R \wedge S)) \wedge (\neg S) \\ \hookrightarrow\ & (R \wedge S) \wedge (\neg S) \\ \hookrightarrow\ & R \wedge (S \wedge (\neg S)) \overset{\mathbb{F}}{=} \mathbb{F}, \end{aligned} \tag{2.28}$$

que por tanto es una contradicción, y hemos mostrado que S se sigue de $R \wedge (R \Rightarrow S)$.

Ejemplo 2.3.3. *Considere las siguientes afirmaciones:*

(1) Si el Ministro sabía de los abusos, entonces mintió al pueblo y debe renunciar.

(2) Si el Ministro no sabía de los abusos, entonces fue incompetente y debe renunciar.

Demuestre que de (1) y (2) se sigue que el Ministro debe renunciar[3].
Solución: *Definiremos las proposiciones*

$$S = \text{“El Ministro sabía de los abusos”}, \tag{2.29a}$$
$$M = \text{“El Ministro mintió al pueblo”}, \tag{2.29b}$$
$$I = \text{“El Ministro fue incompetente”}, \tag{2.29c}$$
$$R = \text{“El Ministro debe renunciar”}, \tag{2.29d}$$

con las que podemos escribir nuestras premisas como

$$S \Rightarrow M, \tag{2.30a}$$
$$M \Rightarrow R, \tag{2.30b}$$
$$(\neg S) \Rightarrow I, \tag{2.30c}$$
$$I \Rightarrow R. \tag{2.30d}$$

y la negación de la conclusión como $\neg R$. Debemos probar entonces la identidad

$$(S \Rightarrow M) \wedge (M \Rightarrow R) \wedge ([\neg S] \Rightarrow I) \wedge (I \Rightarrow R) \wedge (\neg R) = \mathbb{F}. \tag{2.31}$$

Ahora simplemente desarrollamos sustituyendo la definición de implicación y distribuyendo las operaciones lógicas,

$$\begin{aligned}
&(S \Rightarrow M) \wedge (M \Rightarrow R) \wedge ([\neg S] \Rightarrow I) \wedge (I \Rightarrow R) \wedge (\neg R) \\
&\quad \hookrightarrow (\neg S \vee M) \wedge (\neg M \vee R) \wedge (S \vee I) \wedge (\neg I \vee R) \wedge (\neg R) \\
&\quad \hookrightarrow (\neg S \vee M) \wedge (\neg M \vee R) \wedge (S \vee I) \wedge ((\neg I \wedge \neg R) \vee \cancelto{\mathbb{F}}{(R \wedge \neg R)}) \\
&\quad \hookrightarrow (\neg S \vee M) \wedge (\neg M \vee R) \wedge (S \vee I) \wedge (\neg I) \wedge (\neg R) \\
&\quad \hookrightarrow (\neg S \vee M) \wedge ((\neg M \wedge \neg R) \vee \cancelto{\mathbb{F}}{(R \wedge \neg R)}) \wedge (S \vee I) \wedge (\neg I) \\
&\quad \hookrightarrow (\neg S \vee M) \wedge (\neg M) \wedge (\neg R) \wedge (S \vee I) \wedge (\neg I) \\
&\quad \hookrightarrow (((\neg S) \wedge (\neg M)) \vee \cancelto{\mathbb{F}}{(M \wedge \neg M)}) \wedge (\neg R) \wedge (S \vee I) \wedge (\neg I) \\
&\quad \hookrightarrow (\neg S) \wedge (\neg M) \wedge (\neg R) \wedge (S \vee I) \wedge (\neg I) \\
&\quad \hookrightarrow (\neg S) \wedge (\neg M) \wedge (\neg R) \wedge ((S \wedge \neg I) \vee \cancelto{\mathbb{F}}{(I \wedge \neg I)}) \\
&\quad \hookrightarrow (\neg S) \wedge (\neg M) \wedge (\neg R) \wedge S \wedge (\neg I) \\
&\quad \hookrightarrow \cancelto{\mathbb{F}}{((\neg S) \wedge S)} \wedge (\neg M) \wedge (\neg R) \wedge (\neg I) \\
&\quad \hookrightarrow \mathbb{F} \wedge (\neg M) \wedge (\neg R) \wedge (\neg I) = \mathbb{F},
\end{aligned} \tag{2.32}$$

con lo que hemos llegado a una contradicción. Notemos que la contradicción se produce, en la penúltima línea, porque se forma la expresión $(\neg S) \wedge S$ que es falsa por el principio de no contradicción. Luego podemos interpretar el significado de la demostración como el hecho que hay dos caminos de deducción separados, suponer $\neg R$ lleva, por un lado, a demostrar S y por otro lado lleva a demostrar $\neg S$.

[3] Cualquier parecido con la realidad es solamente coincidencia

La reconstrucción en lenguaje natural de la demostración diría más o menos así:

> *"Supongamos que el Ministro no debe renunciar. Entonces, no fue incompetente, y entonces sabía de los abusos. Por otro lado, si no debe renunciar, entonces no mintió, y si no mintió entonces no sabía de los abusos. Como sabía y no sabía de los abusos se ha llegado a una contradicción, y la suposición inicial de que el Ministro no debe renunciar es falsa."*

Pasaremos ahora a definir dos conceptos, comenzando por el de proposiciones mutuamente excluyentes.

Definición 2.1 — Proposiciones mutuamente excluyentes
Diremos que dos proposiciones A y B son mutuamente excluyentes si se cumple

$$A \Rightarrow (\neg B), \tag{2.33}$$

que es equivalente a

$$B \Rightarrow (\neg A). \tag{2.34}$$

Es decir, A y B no pueden ser ciertas a la vez,

$$\neg(A \wedge B).$$

Además del ejemplo trivial de las proposiciones A y $\neg A$, podemos pensar en las proposiciones A=«n es par» y B=«n es impar», o por ejemplo

$$A = \text{«Mi casa queda 100 km al norte de Coquimbo»},$$
$$B = \text{«Mi casa queda 100 km al sur de Coquimbo»}.$$

De manera similar, también podemos definir el concepto de conjunto exhaustivo de proposiciones, o simplemente proposiciones exhaustivas.

Definición 2.2 — Proposiciones exhaustivas
Diremos que las proposiciones $A_1, A_2, \ldots, A_n$ forman un conjunto exhaustivo si se cumple,

$$A_1 \vee A_2 \vee \ldots \vee A_n, \tag{2.35}$$

es decir, al menos una de ellas debe ser verdadera. Dicho de otra manera, no es posible que todas sean falsas, es decir,

$$(\neg A_1) \wedge (\neg A_2) \wedge \ldots \wedge (\neg A_n) = \mathbb{F}. \tag{2.36}$$

a	*b*	*d*	*e*
1	*c*	**2**	*f*
r	**4**	*x*	**3**
s	*t*	*y*	*z*

Figura 2.1: Puzzle *sudoku* de 4×4 donde las casillas desconocidas (en azul) están etiquetadas con nombres de variables, cada una de éstas tomando valores $\in \{1,2,3,4\}$. La solución se discute en el **Ejemplo 2.3.4**.

En términos sencillos, un conjunto exhaustivo de proposiciones cubre todas las posibilidades, sin dejar ninguna fuera. Por ejemplo, el conjunto de las doce proposiciones

$$\begin{aligned} A_1 &= \text{«El mes actual es enero»},\\ A_2 &= \text{«El mes actual es febrero»},\\ A_3 &= \text{«El mes actual es marzo»},\\ &\vdots\\ A_{12} &= \text{«El mes actual es diciembre»}, \end{aligned}$$

es un conjunto exhaustivo.

Un ejemplo por excelencia del uso de la regla de Sherlock Holmes («cuando se ha eliminado lo imposible, lo que queda debe ser la verdad») es el puzzle denominado *sudoku*. Ahí efectivamente se pueden evaluar todas las alternativas y eliminar todas excepto una, que resulta ser la solución. En la **Figura 2.1** se muestra un puzzle *sudoku* de 4×4, cuya solución en términos de proposiciones lógicas veremos en el siguiente ejemplo.

Ejemplo 2.3.4 (Sudoku). *Para resolver el puzzle en la **Figura 2.1**, escribamos las restricciones iniciales para las 12 proposiciones lógicas que describen las posibles soluciones, donde por ejemplo a_3 = «la casilla a contiene el valor 3». Eliminaremos como falsas las proposiciones que son prohibidas por las cuatro casillas dadas del puzzle, para esto reconociendo las restricciones que impone cada casilla ocupada, como se muestra en (2.37) a continuación.*

$$\mathbf{1} \quad \Rightarrow \quad (\neg a_1) \wedge (\neg b_1) \wedge (\neg c_1) \wedge (\neg r_1) \wedge (\neg s_1) \wedge (\neg f_1) = \mathbb{T}, \tag{2.37a}$$

$$\mathbf{4} \quad \Rightarrow \quad (\neg r_4) \wedge (\neg s_4) \wedge (\neg t_4) \wedge (\neg b_4) \wedge (\neg c_4) \wedge (\neg x_4) = \mathbb{T}, \tag{2.37b}$$

$$\mathbf{2} \quad \Rightarrow \quad (\neg d_2) \wedge (\neg e_2) \wedge (\neg f_2) \wedge (\neg c_2) \wedge (\neg x_2) \wedge (\neg y_2) = \mathbb{T}, \tag{2.37c}$$

$$\mathbf{3} \quad \Rightarrow \quad (\neg x_3) \wedge (\neg y_3) \wedge (\neg z_3) \wedge (\neg e_3) \wedge (\neg f_3) \wedge (\neg r_3) = \mathbb{T}. \tag{2.37d}$$

a	*b*	*d*	*e*
1	**3**	**2**	**4**
2	**4**	**1**	**3**
s	*t*	*y*	*z*

Figura 2.2: Estado del puzzle *sudoku* de la **Figura 2.1** luego de aplicar las restricciones en (**2.38**).

*Enseguida escribiremos los valores posibles para cada una de las doce casillas desconocidas en la forma de conjuntos exhaustivos de proposiciones, para luego ir eliminando los valores prohibidos de acuerdo a (**2.37**),*

$$a_1^{\mathbb{F}} \vee a_2 \vee a_3 \vee a_4 = \mathbb{T}, \tag{2.38a}$$
$$b_1^{\mathbb{F}} \vee b_2 \vee b_3 \vee b_4^{\mathbb{F}} = \mathbb{T}, \tag{2.38b}$$
$$c_1^{\mathbb{F}} \vee c_2^{\mathbb{F}} \vee c_3 \vee c_4^{\mathbb{F}} = \mathbb{T}, \tag{2.38c}$$
$$d_1 \vee d_2^{\mathbb{F}} \vee d_3 \vee d_4 = \mathbb{T}, \tag{2.38d}$$
$$e_1 \vee e_2^{\mathbb{F}} \vee e_3^{\mathbb{F}} \vee e_4 = \mathbb{T}, \tag{2.38e}$$
$$f_1^{\mathbb{F}} \vee f_2^{\mathbb{F}} \vee f_3^{\mathbb{F}} \vee f_4 = \mathbb{T}, \tag{2.38f}$$
$$r_1^{\mathbb{F}} \vee r_2 \vee r_3^{\mathbb{F}} \vee r_4^{\mathbb{F}} = \mathbb{T}, \tag{2.38g}$$
$$s_1^{\mathbb{F}} \vee s_2 \vee s_3 \vee s_4^{\mathbb{F}} = \mathbb{T}, \tag{2.38h}$$
$$t_1 \vee t_2 \vee t_3 \vee t_4^{\mathbb{F}} = \mathbb{T}, \tag{2.38i}$$
$$x_1 \vee x_2^{\mathbb{F}} \vee x_3^{\mathbb{F}} \vee x_4^{\mathbb{F}} = \mathbb{T}, \tag{2.38j}$$
$$y_1 \vee y_2^{\mathbb{F}} \vee y_3^{\mathbb{F}} \vee y_4 = \mathbb{T}, \tag{2.38k}$$
$$z_1 \vee z_2 \vee z_3^{\mathbb{F}} \vee z_4 = \mathbb{T}. \tag{2.38l}$$

Sólo con esta información ya hemos establecido $c_3 = \mathbb{T}$, $f_4 = \mathbb{T}$, $r_2 = \mathbb{T}$ *y* $x_1 = \mathbb{T}$, *determinando los valores* $c = 3$, $f = 4$, $r = 2$ *y* $x = 1$ *que nos completan las dos filas centrales del puzzle, como se ve en la **Figura 2.2**.*

Nos quedan 8 incógnitas por determinar, las correspondientes a las filas 1 y 4, y el siguiente proceso es escribir las restricciones debido a las dos filas centrales. Si vamos de izquierda a derecha columna a columna, tenemos

$$(a_4 \wedge s_3) \vee (a_3 \wedge s_4) = \mathbb{T}, \tag{2.39a}$$
$$(b_2 \wedge t_1) \vee (b_1 \wedge t_2) = \mathbb{T}, \tag{2.39b}$$
$$(d_3 \wedge y_4) \vee (d_4 \wedge y_3) = \mathbb{T}, \tag{2.39c}$$
$$(e_1 \wedge z_2) \vee (e_2 \wedge z_1) = \mathbb{T}. \tag{2.39d}$$

Como $s_4 = \mathbb{F}$ por (2.38h), entonces (2.39a) implica $a_4 \wedge s_3 = \mathbb{T}$, es decir, $a = 4$ y $s = 3$. De la misma manera, $b_1 = \mathbb{F}$ por (2.38b), luego (2.39b) implica $b_2 \wedge t_1 = \mathbb{T}$, es decir, $b = 2$ y $t = 1$. Nuevamente, $y_3 = \mathbb{F}$ por (2.38k), entonces (2.39c) implica $d_3 \wedge y_4 = \mathbb{T}$, es decir, $d = 3$ y $y = 4$. Finalmente, $e_2 = \mathbb{F}$ por (2.38e), entonces (2.39d) implica $e_1 \wedge z_2 = \mathbb{T}$, es decir, $e = 1$ y $z = 2$, completando el puzzle.

PROBLEMAS

Problema 2.1. *Si $A, B \in \{\mathbb{T}, \mathbb{F}\}$, ¿cuántas operaciones unarias $f(A)$ existen? ¿Cuántas operaciones binarias $g(A, B)$ existen? Entre las binarias, cuántas de ellas son conmutativas?*

Problema 2.2. *Demuestre que si $A \Rightarrow B$ y $B \Rightarrow C$, entonces $A \Rightarrow C$.*

Problema 2.3. *Demuestre que $(x = y) \Leftrightarrow (x \geq y) \wedge (x \leq y)$.*

Problema 2.4. *Demuestre que R se sigue de las premisas*

$$\begin{gathered} \neg P \vee Q, \\ Q \Rightarrow \big[(\neg R) \wedge (\neg P)\big], \\ P \vee R. \end{gathered}$$

Problema 2.5. *Defina las proposiciones lógicas «E pertenece a la unión de A y B» y «E pertenece a la intersección entre A y B» en términos de las proposiciones que indican pertenencia a A y pertenencia a B. Según su resultado, ¿cuál es la conexión de $\cap$ y $\cup$ con $\vee$ y $\wedge$?*

Problema 2.6. *Demuestre que $A \Rightarrow (B \wedge \neg B)$ es equivalente a $\neg A$.*

Problema 2.7. *Formalice la frase*

> *Carlos viene a la fiesta siempre y cuando Daniela no venga, pero si Daniela viene, entonces Beatriz no viene.*

escribiéndola como una proposición lógica usando los símbolos $\neg$, $\wedge$ y $\vee$.

Problema 2.8. *Formalice la frase*

> *Si juegas y estudias pasarás los exámenes, mientras que si juegas y no estudias no los pasarás. Por lo tanto si juegas, o estudias y pasas, o no estudias y no pasas.*

escribiéndola como un argumento con premisas y una conclusión y usando los símbolos $\neg$, $\wedge$ y $\vee$. ¿Es la deducción correcta?

Problema 2.9. *En una expedición, Lara Croft encuentra tres cajas cerradas, cada una con una inscripción como sigue.*

Caja 1: *«El oro no se encuentra aquí.»*

Caja 2: *«El oro no se encuentra aquí.»*

Caja 3: *«El oro está en la caja 2.»*

Sólo una de estas inscripciones es cierta, las otras dos son falsas. ¿Cuál caja tiene el oro?

Problema 2.10. *Demuestre usando inducción matemática que, para* $x > -1$*, se cumple*

$$(1+x)^n \geq 1 + nx \tag{2.40}$$

para todos los enteros $n \geq 1$*.*

Problema 2.11. *Demuestre usando inducción matemática que*

$$\sum_{k=1}^{n} (2k-1) = n^2 \tag{2.41}$$

para $n \geq 1$*.*

Problema 2.12. *Demuestre usando inducción matemática que* $5^n - 1$ *es divisible por 4 para todo entero positivo n.*

CAPÍTULO 3

Herramientas matemáticas

> It's dangerous to go alone! Take this.
>
> The Legend of Zelda

Antes de seguir con nuestro recorrido debemos adquirir algunos conceptos y técnicas como parte del vocabulario matemático necesario para expresar las ideas de los siguientes capítulos. En primer lugar, introduciremos algunas funciones y funciones generalizadas que utilizaremos para expresar tipos particulares de conocimiento acerca de variables continuas y discretas.

3.1 — FUNCIÓN ESCALÓN DE HEAVISIDE

La función escalón de Heaviside Θ es una función que sólo depende del signo de su argumento, tomando el valor 0 para argumentos negativos, y 1 cuando su argumento es cero o un número positivo. Esto es, $\Theta(x) \in \{0, 1\}$, como se ve en la **Figura 3.1**.

Definición 3.1 — Función escalón de Heaviside
La *función escalón* de Heaviside $\Theta(x)$ se define como

$$\Theta(x) := \begin{cases} 1 \text{ si } x \geq 0, \\ \\ 0 \text{ si } x < 0, \end{cases} \tag{3.1}$$

bajo la convención donde la función escalón en cero es igual a 1.

Usando $x - x_0$ como argumento de (3.1) podemos fijar en x_0 el punto donde ocurre el salto discontinuo,

$$\Theta(x - x_0) = \begin{cases} 1 \text{ si } x \geq x_0, \\ \\ 0 \text{ si } x < x_0. \end{cases} \tag{3.2}$$

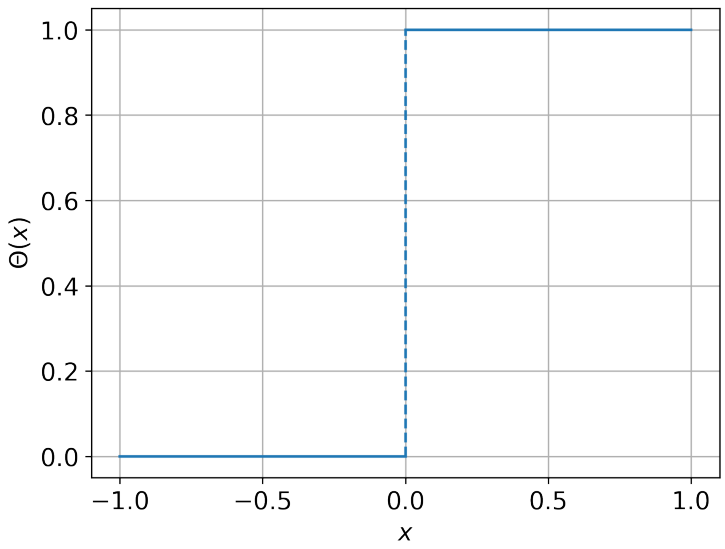


Figura 3.1: Función escalón de Heaviside.

mientras que usando $x_0 - x$ como argumento, tenemos una versión «reflejada» de (3.2) en torno a $x = x_0$,

$$\Theta(x_0 - x) = \begin{cases} 1 \text{ si } x \leq x_0, \\ \\ 0 \text{ si } x > x_0. \end{cases} \tag{3.3}$$

De la definición (3.1) es directo ver que la función escalón no es par ni impar, sino que transforma como

$$\Theta(-x) = 1 - \Theta(x) \qquad \text{para } x \neq 0. \tag{3.4}$$

Un uso inmediato de la función escalón es «truncar» una suma o integral. Por ejemplo, para fijar el límite superior usamos

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{d}x f(x)\Theta(b-x) &= \int_{-\infty}^{b} \mathrm{d}x f(x)\cancelto{1}{\Theta(b-x)} + \int_{b}^{\infty} \mathrm{d}x f(x)\cancelto{0}{\Theta(b-x)} \\ &= \int_{-\infty}^{b} \mathrm{d}x f(x), \end{aligned} \tag{3.5}$$

mientras que para fijar el límite inferior, usamos

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{d}x f(x)\Theta(x-a) &= \int_{-\infty}^{a} \mathrm{d}x f(x)\cancelto{0}{\Theta(x-a)} + \int_{a}^{\infty} \mathrm{d}x f(x)\cancelto{1}{\Theta(x-a)} \\ &= \int_{a}^{\infty} \mathrm{d}x f(x). \end{aligned} \tag{3.6}$$

Para truncar una suma discreta se tiene de la misma manera,

$$\sum_{n=0}^{\infty} a_n \Theta(m-n) = \sum_{n=0}^{m} a_n, \tag{3.7a}$$

$$\sum_{n=0}^{\infty} a_n \Theta(n-m) = \sum_{n=m}^{\infty} a_n. \tag{3.7b}$$

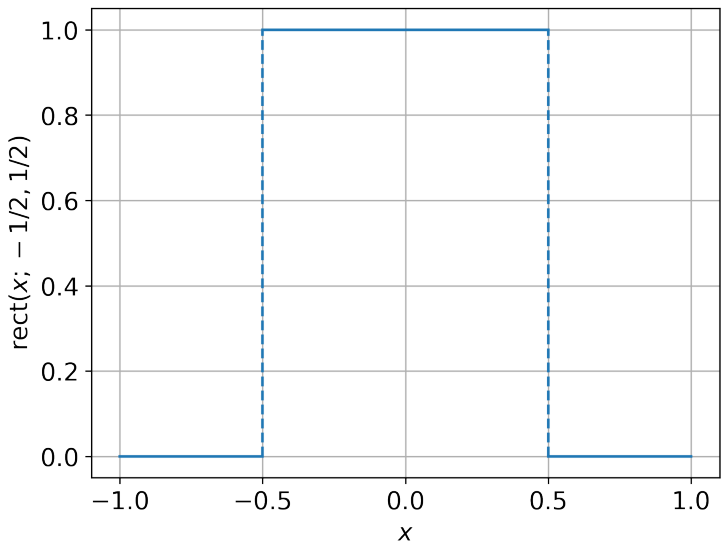


Figura 3.2: Función rectangular $\mathrm{rect}(\bullet; -\frac{1}{2}, \frac{1}{2})$.

Si usamos (3.5) reemplazando $f(x)$ por $f(x)\Theta(x-a)$ podemos fijar simultáneamente los límites superior e inferior,

$$\int_{-\infty}^{\infty} \mathrm{d}x f(x)\Theta(b-x)\Theta(x-a) = \int_{-\infty}^{b} \mathrm{d}x f(x)\Theta(x-a) = \int_{a}^{b} \mathrm{d}x f(x). \quad (3.8)$$

lo cual nos lleva a definir el producto de funciones escalón como una nueva función discontinua, la *función rectangular*, que se muestra en la Figura 3.2.

Definición 3.2 — Función rectangular
Definimos la función rectangular en el intervalo $[a, b]$ como

$$\mathrm{rect}(x; a, b) := \Theta(b-x)\Theta(x-a) = \begin{cases} 1 \text{ si } a \leq x \leq b, \\ 0 \text{ en caso contrario.} \end{cases} \quad (3.9)$$

Usando la función rectangular es posible escribir (3.8) cómodamente como

$$\int_{-\infty}^{\infty} \mathrm{d}x f(x)\,\mathrm{rect}(x; a, b) = \int_{a}^{b} \mathrm{d}x f(x). \quad (3.10)$$

Esta función rectangular posee la siguiente interesante propiedad.

Lema 3.1.

$$\mathrm{rect}(x; a, b) = \Theta(b-x) - \Theta(a-x). \quad (3.11)$$

Su demostración es directa, y pasa por evaluar caso a caso el comportamiento de la función.

Demostración. Analizamos los tres casos posibles para x.

Caso con $x > b$: Tenemos $\Theta(b-x) = 0$, y como $x > b$ implica $x > a$, también tenemos $\Theta(a-x) = 0$, luego

$$\Theta(b-x) - \Theta(a-x) = 0.$$

Caso con $x \leq a$: Tenemos $\Theta(a-x) = 1$, y como $x \leq a$ implica $x < b$, también tenemos $\Theta(b-x) = 1$, luego

$$\Theta(b-x) - \Theta(a-x) = 0.$$

Caso con $a < x \leq b$: Tenemos $\Theta(b-x) = 1$ y $\Theta(a-x) = 0$ luego

$$\Theta(b-x) - \Theta(a-x) = 1$$

En más dimensiones, también podemos usar la función escalón para cambiar el dominio de integración. Por ejemplo, para integrar sólo sobre los puntos en Ω tal que una función arbitraria $g(\boldsymbol{x})$ toma un valor menor que G, usamos

$$\int_{\Omega} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \Theta(G - g(\boldsymbol{x})) = \int_{\boldsymbol{x}\in\Omega,\, g\leq G} d\boldsymbol{x} f(\boldsymbol{x}). \tag{3.12}$$

En el Capítulo 4 veremos que las funciones escalón y rectangular pertenecen a una familia de funciones llamadas *funciones indicador*, cuya generalidad nos permitirá unificar (3.5), (3.6), (3.10) y (3.12) en una sola propiedad de cambio de dominio de integración.

3.2 — Delta de Dirac

La derivada de la función escalón $\Theta(x)$ es una cantidad discontinua en $x = 0$, conocida como la *delta de Dirac* y denotada por $\delta(x)$. Esta cantidad no es considerada como una función en el sentido usual sino que es una *función generalizada*[(1)] es decir, es el límite de una secuencia de funciones, que representa una cantidad infinitamente concentrada en un punto. Su definición es la siguiente.

[(1)] Estas funciones generalizadas son también conocidas como *distribuciones*, aunque para nosotros ése es un término reservado que emplearemos para referirnos a modelos en inferencia.

Definición 3.3 — Delta de Dirac
La delta de Dirac $\delta(x)$ es la derivada de la función escalón respecto a su argumento,

$$\delta(x) := \frac{\mathrm{d}}{\mathrm{d}x}\Theta(x) = \begin{cases} +\infty \text{ si } x = 0, \\ \\ 0 \text{ en caso contrario.} \end{cases} \tag{3.13}$$

Aunque gráficamente es claro, como se ve en la **Figura 3.1**, que la derivada de $\Theta(x)$ tiene esas propiedades, ya que $\Theta(x)$ es plana en todos lados excepto en $x = 0$, veámoslo un poco más formalmente. Para esto, construimos la derivada de la función escalón como el límite

$$\delta(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Theta(x) = \lim_{\Delta\to 0}\left[\frac{\Theta(x+\frac{\Delta}{2}) - \Theta(x-\frac{\Delta}{2})}{\Delta}\right], \tag{3.14}$$

y usando la propiedad (3.4), reescribimos

$$\begin{aligned}\Theta\Big(x+\tfrac{\Delta}{2}\Big) - \Theta\Big(x-\tfrac{\Delta}{2}\Big) &= 1 - \Theta\Big(-\tfrac{\Delta}{2}-x\Big) - \Big[1-\Theta\Big(\tfrac{\Delta}{2}-x\Big)\Big] \\ &= \Theta\Big(\tfrac{\Delta}{2}-x\Big) - \Theta\Big(-\tfrac{\Delta}{2}-x\Big).\end{aligned} \tag{3.15}$$

Ahora, utilizando (3.15) y la propiedad (3.11) de la función rectangular podemos reescribir la cantidad en el corchete del lado derecho de (3.14) como

$$\begin{aligned}\frac{1}{\Delta}\left[\Theta\Big(x+\tfrac{\Delta}{2}\Big) - \Theta\Big(x-\tfrac{\Delta}{2}\Big)\right] &= \frac{1}{\Delta}\,\mathrm{rect}\left(x; -\tfrac{\Delta}{2}, \tfrac{\Delta}{2}\right) \\ &= \begin{cases} \frac{1}{\Delta} \text{ si } -\frac{\Delta}{2} < x \leq \frac{\Delta}{2} \\ \\ 0 \text{ en caso contrario.}\end{cases}\end{aligned} \tag{3.16}$$

Al tomar el límite $\Delta \to 0$, tenemos que el intervalo $-\frac{\Delta}{2} < x \leq \frac{\Delta}{2}$ colapsa alrededor de $x = 0$, mientras que $\frac{1}{\Delta} \to \infty$, luego se tiene que

$$\delta(x) = \lim_{\Delta\to 0}\left[\frac{1}{\Delta}\,\mathrm{rect}\left(x; -\tfrac{\Delta}{2}, \tfrac{\Delta}{2}\right)\right] = \begin{cases} +\infty \text{ si } x = 0 \\ \\ 0 \text{ en caso contrario.}\end{cases} \tag{3.17}$$

A partir de (3.17) es fácil ver que la delta de Dirac es par,

$$\delta(-x) = \delta(x), \tag{3.18}$$

ya que $\mathrm{rect}(-x; -a, a) = \mathrm{rect}(x; -a, a)$ para todo $a > 0$.

La propiedad más útil de la delta de Dirac es que su integral junto a cualquier función simplemente evalúa dicha función en cero. Esto es,

$$\boxed{\int_{-\infty}^{\infty} \mathrm{d}x f(x)\delta(x) = f(0)} \tag{3.19}$$

para toda función $f(x)$.

Demostración. Para una cantidad $L > 0$ que luego llevaremos al límite $L \to \infty$, integramos por partes

$$\begin{aligned}\int_{-L}^{L} \mathrm{d}x f(x)\delta(x) &= \int_{-L}^{L} \mathrm{d}x f(x)\frac{\mathrm{d}}{\mathrm{d}x}\Theta(x) \\ &= f(x)\Theta(x)\Big|_{-L}^{L} - \int_{-L}^{L} \mathrm{d}x\Theta(x)\frac{\mathrm{d}f}{\mathrm{d}x} \\ &= f(L) - 0 - \int_{0}^{L} \mathrm{d}x\left(\frac{\mathrm{d}f}{\mathrm{d}x}\right) \\ &= \cancel{f(L)} - \cancel{f(L)} + f(0) = f(0)\end{aligned} \tag{3.20}$$

y esto es válido para cualquier valor de L, incluso para $L \to \infty$

En particular, si $f(x)$ es la función constante igual a 1 para todo x, tenemos

$$\int_{-\infty}^{\infty} \mathrm{d}x\delta(x) = 1. \tag{3.21}$$

En n dimensiones, si $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ usaremos la notación compacta $\delta(\boldsymbol{x})$ para representar el producto

$$\delta(\boldsymbol{x}) := \prod_{i=1}^{n} \delta(x_i), \tag{3.22}$$

tal que se cumple

$$\int \mathrm{d}\boldsymbol{x} f(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{x}_0) = f(\boldsymbol{x}_0). \tag{3.23}$$

De la misma manera se tiene que, integrando en toda la región Ω donde está definido $\boldsymbol{x}$,

$$\int_{\Omega} \mathrm{d}\boldsymbol{x}\delta(\boldsymbol{x} - \boldsymbol{x}_0) = 1. \tag{3.24}$$

> **Importante:** No siempre la integral de la delta de Dirac es igual a 1. En general, si integramos $\delta(\boldsymbol{x} - \boldsymbol{x}_0)$ en un dominio $U \subset \Omega$ particular, el resultado depende de si $\boldsymbol{x}_0$ está contenido o no en U.

Más precisamente,

$$\int_{U} \mathrm{d}\boldsymbol{x}\delta(\boldsymbol{x} - \boldsymbol{x}_0) = \begin{cases} 1 \text{ si } \boldsymbol{x}_0 \in U, \\ \\ 0 \text{ en caso contrario.} \end{cases} \tag{3.25}$$

Para demostrar propiedades de la delta de Dirac, se usa lo siguiente. Si queremos demostrar una identidad de la forma

$$A[\delta; \boldsymbol{x}] = B[\delta; \boldsymbol{x}], \tag{3.26}$$

donde A y B son expresiones que involucran a la delta de Dirac y son a la vez funciones de $\boldsymbol{x}$, entonces basta demostrar que

$$\int_\Omega \mathrm{d}\boldsymbol{x}\omega(\boldsymbol{x})A[\delta;\boldsymbol{x}] = \int_\Omega \mathrm{d}\boldsymbol{x}\omega(\boldsymbol{x})B[\delta;\boldsymbol{x}] \tag{3.27}$$

para toda función de prueba $\omega(\boldsymbol{x})$. Por ejemplo, es directo demostrar la identidad

$$\boxed{f(\boldsymbol{x})\delta(\boldsymbol{x}-\boldsymbol{x}_0) = f(\boldsymbol{x}_0)\delta(\boldsymbol{x}-\boldsymbol{x}_0),} \tag{3.28}$$

si uno ve que para toda función de prueba $\omega(\boldsymbol{x})$,

$$\begin{aligned}\int_\Omega \mathrm{d}\boldsymbol{x}\omega(\boldsymbol{x})\Big[f(\boldsymbol{x})\delta(\boldsymbol{x}-\boldsymbol{x}_0)\Big] &= \omega(\boldsymbol{x}_0)f(\boldsymbol{x}_0)\\ &= \int_\Omega \mathrm{d}\boldsymbol{x}\omega(\boldsymbol{x})\Big[f(\boldsymbol{x}_0)\delta(\boldsymbol{x}-\boldsymbol{x}_0)\Big].\end{aligned} \tag{3.29}$$

Utilicemos esta técnica para demostrar otra de las propiedades útiles de la delta de Dirac, la propiedad de reescalamiento, correspondiente a

$$\boxed{\delta(\alpha x) = \frac{\delta(x)}{|\alpha|}} \tag{3.30}$$

para $\alpha \neq 0$.

Demostración. Consideremos primero el caso $\alpha > 0$. Integramos una función de prueba $\omega(x)$ con la delta del lado izquierdo

$$\int_{-\infty}^{\infty} \mathrm{d}x\omega(x)\delta(\alpha x) \underset{u=\alpha x}{=} \frac{1}{\alpha}\int_{-\infty}^{\infty} \mathrm{d}u\omega(u/\alpha)\delta(u) = \frac{\omega(0)}{\alpha}. \tag{3.31}$$

Como

$$\frac{\omega(0)}{\alpha} = \int_{-\infty}^{\infty} \mathrm{d}x\omega(x)\frac{\delta(x)}{\alpha} \tag{3.32}$$

hemos demostrado que

$$\int_{-\infty}^{\infty} \mathrm{d}x\omega(x)\left[\delta(\alpha x)\right] = \int_{-\infty}^{\infty} \mathrm{d}x\omega(x)\left[\frac{\delta(x)}{\alpha}\right], \tag{3.33}$$

para cualquier función $\omega(x)$, y de aquí se sigue que

$$\delta(\alpha x) = \frac{\delta(x)}{\alpha}, \qquad \alpha > 0. \tag{3.34}$$

El caso $\alpha < 0$ se obtiene directamente, por la paridad de la delta de Dirac según (3.18) y la propiedad recién demostrada. Como $\alpha < 0$, podemos escribir $\alpha = -|\alpha|$ y luego

$$\delta(-|\alpha|x) = \delta(|\alpha|x) = \frac{\delta(x)}{|\alpha|}. \tag{3.35}$$

Juntando ambos casos vemos que (3.30) es cierta.

3.2.1 Representaciones de la delta de Dirac

Hemos dicho que la delta de Dirac es, más que una función, el límite de una secuencia de funciones. Más precisamente, queremos decir con esto que la delta de Dirac puede representarse por un número, en principio infinito, de distintos límites[(2)] de funciones, límites que poseen la forma general

$$\delta(x) = \lim_{\varepsilon\to 0}\left[\frac{1}{\varepsilon}\eta\left(\frac{x}{\varepsilon}\right)\right] \tag{3.36}$$

para toda función $\eta(u)$ tal que $\eta(+\infty) = \eta(-\infty) = 0$ y

$$\int_{-\infty}^{\infty} du\eta(u) = 1. \tag{3.37}$$

Sin un exceso de rigurosidad, veamos por qué esto es cierto. Teniendo en cuenta que para $\varepsilon > 0$ muy pequeño (¡pero no cero!) se tiene que

$$\eta\left(\frac{x}{\varepsilon}\right) \approx \begin{cases} \eta(0) \text{ si } x = 0, \\ \eta(-\infty) = 0 \text{ si } x < 0, \\ \eta(+\infty) = 0 \text{ si } x > 0, \end{cases} \tag{3.38}$$

calculamos la integral del límite en (3.36) con una función de prueba, obteniendo

$$\begin{aligned} \int_{-\infty}^{\infty} dx\omega(x) \lim_{\varepsilon\to 0}\left[\frac{1}{\varepsilon}\eta\left(\frac{x}{\varepsilon}\right)\right] &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{-\infty}^{\infty} dx\omega(x)\eta\left(\frac{x}{\varepsilon}\right) \\ \text{usando (3.38)} \quad &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\omega(0)\int_{-\infty}^{\infty} dx\eta\left(\frac{x}{\varepsilon}\right) \\ (\text{usando } u = x/\varepsilon) \quad &= \lim_{\varepsilon\to 0}\omega(0)\int_{-\infty}^{\infty} du\eta(u) = \omega(0), \end{aligned} \tag{3.39}$$

de lo cual se sigue (3.36). Por ejemplo, la *función gaussiana*

$$\phi(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2) \tag{3.40}$$

con $\int_{-\infty}^{\infty} du\phi(u) = 1$ nos lleva la siguiente representación de la delta de Dirac,

$$\delta(x) = \lim_{\varepsilon\to 0}\frac{1}{\sqrt{2\pi\varepsilon}}\exp\left(-\frac{x^2}{2\varepsilon^2}\right). \tag{3.41}$$

Por otro lado, la función

$$\eta(u) = \frac{\sin(u)}{\pi u} \tag{3.42}$$

nos dice que

$$\delta(x) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon\pi}\left(\frac{\varepsilon}{x}\right)\sin\left(\frac{x}{\varepsilon}\right) = \lim_{\varepsilon\to 0}\frac{1}{\pi x}\sin\left(\frac{x}{\varepsilon}\right). \tag{3.43}$$

[(2)] A estos límites que conducen a la delta de Dirac se les denomina aproximaciones a la identidad.

Haciendo uso de la integral

$$\int_{-a}^{a} \mathrm{d}k \exp(\mathrm{i}kx) = \int_{-a}^{a} \mathrm{d}k \cos(kx) + \mathrm{i}\cancelto{0}{\int_{-a}^{a} \mathrm{d}k \sin(kx)} = \frac{2}{x}\sin(ax), \tag{3.44}$$

vemos que

$$\begin{aligned}\delta(x) &= \lim_{\varepsilon\to 0} \frac{1}{\pi x}\sin\left(\frac{x}{\varepsilon}\right) \\ &= \lim_{\varepsilon\to 0} \frac{1}{2\pi}\int_{-1/\varepsilon}^{1/\varepsilon} \mathrm{d}k \exp(\mathrm{i}kx) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(\mathrm{i}kx).\end{aligned} \tag{3.45}$$

por lo tanto, obtenemos una representación integral de la delta de Dirac como

$$\boxed{\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(\mathrm{i}kx).} \tag{3.46}$$

Esta identidad puede extenderse a n dimensiones simplemente como

$$\begin{aligned}\delta(\boldsymbol{x}) = \prod_{i=1}^{n}\delta(x_i) &= \prod_{i=1}^{n}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k_i \exp(\mathrm{i}k_i x_i)\right] \\ &= \left(\frac{1}{2\pi}\right)^n \int \mathrm{d}\boldsymbol{k} \prod_{i=1}^{n}\exp(\mathrm{i}k_i x_i) \\ &= \frac{1}{(2\pi)^n}\int \mathrm{d}\boldsymbol{k}\exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}).\end{aligned} \tag{3.47}$$

3.2.2 Composición de la delta de Dirac

Una delta de Dirac cuyo argumento es una función arbitraria de la variable x, digamos $f(x)$, puede escribirse en términos de la delta de Dirac sobre la propia variable x, usando la identidad

$$\boxed{\delta(f(x)) = \sum_{x_0}\frac{\delta(x-x_0)}{|f'(x)|}} \tag{3.48}$$

donde $f'(x) = \dfrac{\mathrm{d}f}{\mathrm{d}x}$ es la derivada de f y donde x_0 son los puntos tales que $f(x_0) = 0$.

Demostración. Consideremos una función de prueba $\omega(x)$. Desarrollamos la integral

$$\int_{-\infty}^{\infty} \mathrm{d}x\omega(x)\delta(f(x)) = \sum_{x_0}\int_{x_0-\Delta}^{x_0+\Delta} \mathrm{d}x\omega(x)\delta(f(x)), \tag{3.49}$$

donde x_0 son los puntos tales que $f(x_0) = 0$, que de hecho son los únicos puntos que contribuyen a la integral, y donde Δ es un número positivo suficientemente pequeño. Expandiendo $f(x)$ a primer orden

en torno a x_0 en el argumento de la delta de Dirac del lado derecho, tenemos

$$\begin{aligned}\sum_{x_0}\int_{x_0-\Delta}^{x_0+\Delta} dx\omega(x)\delta(f(x)) \\ &= \sum_{x_0}\int_{x_0-\Delta}^{x_0+\Delta} dx\omega(x)\delta(\cancel{f(x_0)}^{0} + [x-x_0]f'(x_0)) \\ &= \sum_{x_0}\int_{x_0-\Delta}^{x_0+\Delta} dx\omega(x)\delta([x-x_0]f'(x_0)) \end{aligned} \tag{3.50}$$

ya que $f(x_0) = 0$ por la definición de x_0. Ahora, haciendo uso de la propiedad (3.30), tenemos

$$\begin{aligned}\sum_{x_0}\int_{x_0-\Delta}^{x_0+\Delta} dx\omega(x)\delta([x-x_0]f'(x_0)) &= \sum_{x_0}\int_{x_0-\Delta}^{x_0+\Delta} dx\omega(x)\frac{\delta(x-x_0)}{|f'(x_0)|} \\ &= \sum_{x_0}\frac{\omega(x_0)}{|f'(x_0)|}\underbrace{\int_{x_0-\Delta}^{x_0+\Delta} dx\delta(x-x_0)}_{=1} \\ &= \sum_{x_0}\frac{\omega(x_0)}{|f'(x_0)|}. \end{aligned} \tag{3.51}$$

Hemos demostrado entonces que

$$\begin{aligned}\int_{-\infty}^{\infty} dx\omega(x)\delta(f(x)) &= \sum_{x_0}\frac{\omega(x_0)}{|f'(x_0)|} \\ &= \int_{-\infty}^{\infty} dx\omega(x)\left[\sum_{x_0}\frac{\delta(x-x_0)}{|f'(x)|}\right] \end{aligned} \tag{3.52}$$

para cualquier función $\omega(x)$, por lo tanto (3.48) es cierta.

Podemos ver de inmediato que la identidad (3.30) es un caso particular de (3.48) con $f(x) = ax$ y por tanto $x_0 = 0$. En más dimensiones, el equivalente de la propiedad (3.48) es

$$\boxed{\delta(f(\boldsymbol{x})) = \sum_{x_0}\frac{\delta(\boldsymbol{x}-\boldsymbol{x}_0)}{|\nabla f|},} \tag{3.53}$$

donde $\nabla f = \dfrac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$ y $\boldsymbol{x}_0$ son los puntos tales que $f(\boldsymbol{x}_0) = 0$.

3.2.3 Derivadas de la delta de Dirac

La derivada de la delta de Dirac, denotada como $\delta'(x)$, cumple con la propiedad

$$\int_{-\infty}^{\infty} dx f(x)\delta'(x-x_0) = -f'(x_0), \tag{3.54}$$

como puede demostrarse por integración por partes, o simplemente reconociendo que

$$\frac{\partial}{\partial x}\delta(x-x_0) = \delta'(x-x_0) = -\frac{\partial}{\partial x_0}\delta(x-x_0), \tag{3.55}$$

con lo que podemos escribir

$$f'(x_0) = \frac{\partial}{\partial x_0}\left[\int_{-\infty}^{\infty} \mathrm{d}x f(x)\delta(x-x_0)\right] = -\int_{-\infty}^{\infty} \mathrm{d}x f(x)\delta'(x-x_0). \tag{3.56}$$

La generalización de (3.54) a derivadas de orden superior es la identidad

$$\int_{-\infty}^{\infty} \mathrm{d}x f(x)\delta^{(n)}(x-x_0) = (-1)^n f^{(n)}(x_0), \tag{3.57}$$

que podemos demostrar con la técnica de inducción matemática haciendo uso de (3.54) y la integración por partes.

► Para saber más sobre las propiedades de la delta de Dirac, se recomienda el libro de Arfken y Weber (2005) y el de Riley, Hobson y Bence (2006).

3.3 — Cambios de variable y la delta de Dirac

3.3.1 Cambios de variable en integrales simples

Consideremos la integral definida

$$Z = \int_0^{\infty} \mathrm{d}t \exp(-\lambda t) \tag{3.58}$$

para $\lambda > 0$ la cual queremos resolver introduciendo la nueva variable $u := -\lambda t$. Para realizar el cambio de variable, tradicionalmente expresamos $\mathrm{d}t$ en términos de $\mathrm{d}u$ usando

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\lambda \quad \text{y luego} \quad \mathrm{d}t = -\frac{\mathrm{d}u}{\lambda} \tag{3.59}$$

y con esto sustituimos t en función de u en la integral original, obteniendo

$$Z = -\int_{u(0)}^{u(\infty)} \frac{\mathrm{d}u}{\lambda}\exp(u) = \frac{1}{\lambda}\int_{-\infty}^{0} \mathrm{d}u \exp(u) = \frac{1}{\lambda}. \tag{3.60}$$

Veremos ahora un método sistemático y general de realizar cambios de variable, utilizando las propiedades de la delta de Dirac. En el ejemplo anterior, consideramos nuevamente el cambio de variable $t \to u = -\lambda t$ pero esta vez agregaremos un factor 1 en (3.58) como la integral de una delta de Dirac en la nueva variable u,

$$Z = \int_0^{\infty} \mathrm{d}t \underbrace{\int_{-\infty}^{\infty} \mathrm{d}u \delta(u+\lambda t)}_{=1} \exp(-\lambda t). \tag{3.61}$$

Intercambiando las integrales en u y en t luego de sustituir $-\lambda t$ por u en la exponencial como lo impone la delta de Dirac, obtenemos

$$Z = \int_{-\infty}^{\infty} \mathrm{d}u \left[\int_{0}^{\infty} \mathrm{d}t \delta(u + \lambda t) \right] \exp(u) \tag{3.62}$$

para luego desarrollar la integral en corchetes como

$$\begin{aligned} \int_{0}^{\infty} \mathrm{d}t \delta(u + \lambda t) &= \int_{0}^{\infty} \mathrm{d}t \delta\left(\lambda \left[\frac{u}{\lambda} + t\right]\right) \\ \text{usando (3.30)} \quad &= \frac{1}{\lambda} \int_{0}^{\infty} \mathrm{d}t \delta\left(t + \frac{u}{\lambda}\right) \\ &= \frac{1}{\lambda} \Theta(-u), \end{aligned} \tag{3.63}$$

la cual sólo es distinta de cero cuando $u < 0$. Reemplazando en (3.62) tenemos entonces

$$Z = \frac{1}{\lambda} \int_{-\infty}^{\infty} \mathrm{d}u \Theta(-u) \exp(u) = \frac{1}{\lambda} \int_{-\infty}^{0} \mathrm{d}u \exp(u) = \frac{1}{\lambda}. \tag{3.64}$$

Veamos cómo funciona en general este procedimiento para un cambio de variables $u = f(t)$ en una integral de la forma

$$Z = \int_{a}^{b} \mathrm{d}t G(f(t)) \tag{3.65}$$

donde $f(t)$ es invertible y $T(\bullet)$ es la función inversa de $f(\bullet)$, de forma que $T(u)$ entrega el valor de t tal que es solución de $u = f(t)$. Nuevamente agregando un 1 en (3.65) como una integral de la delta de Dirac en la variable u, tenemos

$$\begin{aligned} Z &= \int_{a}^{b} \mathrm{d}t \left[\int_{-\infty}^{\infty} \mathrm{d}u \delta(u - f(t)) \right] G(f(t)) \\ &= \int_{-\infty}^{\infty} \mathrm{d}u \left[\int_{a}^{b} \mathrm{d}t \delta(u - f(t)) \right] G(f(t)) \\ &= \int_{-\infty}^{\infty} \mathrm{d}u \left[\int_{a}^{b} \mathrm{d}t \delta(u - f(t)) \right] G(u) \end{aligned} \tag{3.66}$$

y la integral en corchetes en la última línea puede desarrollarse como

$$\int_{a}^{b} \mathrm{d}t \delta(u - f(t)) = \int_{a}^{b} \mathrm{d}t \frac{\delta(t - T(u))}{|f'(T(u))|} = \frac{\mathrm{rect}(T(u); a, b)}{|f'(T(u))|}. \tag{3.67}$$

Ahora la función rectangular $\mathrm{rect}(T(u); a, b) = 1$ impone la condición

$$a \leq T(u) \leq b,$$

que, aplicando $f(\bullet)$ a la desigualdad, es equivalente a $u_1 \leq u \leq u_2$. Esto es, $\mathrm{rect}(u; u_1, u_2) = 1$ donde

$$u_1 := \mathrm{mín}(f(a), f(b)), \tag{3.68a}$$

$$u_2 := \mathrm{máx}(f(a), f(b)), \tag{3.68b}$$

y por tanto podemos escribir

$$Z = \int_{-\infty}^{\infty} \mathrm{d}u \left[\frac{\mathrm{rect}(u; u_1, u_2)}{|f'(T(u))|}\right] G(u) = \int_{u_1}^{u_2} \frac{\mathrm{d}u G(u)}{|f'(T(u))|}. \tag{3.69}$$

Si $f(t)$ crece siempre con t dentro del intervalo $t \in [a, b]$, entonces $f'(t) \geq 0$ y podemos reemplazar $|f'(t)| = f'(t)$, además en ese caso $u_1 = f(a)$ y $u_2 = f(b)$. Por otro lado, si $f(t)$ decrece siempre con t en el intervalo $t \in [a, b]$, se tendrá $f'(t) \leq 0$ con lo que reemplazamos $|f'(t)| = -f'(t)$ y además se cumplirá $u_1 = f(b)$ y $u_2 = f(a)$. En este último caso, el signo en $-f'(t)$ permite invertir los límites de integración, y ambos casos pueden unirse como

$$\boxed{\int_a^b \mathrm{d}t G(f(t)) = \int_{f(a)}^{f(b)} \frac{\mathrm{d}u G(u)}{f'(T(u))},} \tag{3.70}$$

que coincide con la simple intuición de transformar los límites de integración y sustituir

$$\mathrm{d}t \to \frac{\mathrm{d}u}{f'(t)}.$$

En el caso donde $f'(t)$ cambia de signo en $t \in [a, b]$ siempre es posible subdividir dicho intervalo en regiones crecientes y decrecientes, aplicar (3.70) en cada región y volver a juntar los intervalos, con lo que vemos que (3.70) es siempre válida.

3.3.2 Cambios de variables en integrales múltiples

Consideremos un conjunto de variables $\boldsymbol{u} = (u_1, \ldots, u_d)$ y una representación alternativa de coordenadas $\boldsymbol{v} = (v_1, \ldots, v_d)$ tal que existe una transformación que las conecta, $\boldsymbol{u} \mapsto \boldsymbol{V}(\boldsymbol{u})$. La transformación inversa $\boldsymbol{v} \mapsto \boldsymbol{U}(\boldsymbol{v})$ es tal que

$$\boldsymbol{U}(\boldsymbol{V}(\boldsymbol{u})) = \boldsymbol{u} \qquad \text{para todo } \boldsymbol{u}. \tag{3.71}$$

Para fijar ideas, supongamos la transformación entre coordenadas cartesianas y polares esféricas en tres dimensiones. Tenemos $d = 3$ con

$$\boldsymbol{u} = (u_1, u_2, u_3) = (x, y, z), \tag{3.72a}$$
$$\boldsymbol{v} = (v_1, v_2, v_3) = (r, \theta, \phi), \tag{3.72b}$$

tal que la transformación de $\boldsymbol{u}$ a $\boldsymbol{v}$ es

$$\begin{aligned} V_1(u_1, u_2, u_3) &= \sqrt{u_1^2 + u_2^2 + u_3^2}, \\ V_2(u_1, u_2, u_3) &= \cos^{-1}(u_3 / \sqrt{u_1^2 + u_2^2 + u_3^2}), \\ V_3(u_1, u_2, u_3) &= \tan^{-1}(u_2 / u_1), \end{aligned} \tag{3.73}$$

y la transformación inversa de $\boldsymbol{v}$ a $\boldsymbol{u}$ es

$$\begin{aligned} U_1(v_1, v_2, v_3) &= v_1 \cos(v_3) \sin(v_2), \\ U_2(v_1, v_2, v_3) &= v_1 \sin(v_3) \sin(v_2), \\ U_3(v_1, v_2, v_3) &= v_1 \cos(v_2) \end{aligned} \tag{3.74}$$

La integral $Z = \int \mathrm{d}\boldsymbol{u} G(\boldsymbol{u})$ de una función $G(\boldsymbol{u})$ arbitraria puede transformarse a una integral en las coordenadas $\boldsymbol{v}$, usando la fórmula (Riley, Hobson y Bence 2006)

$$\boxed{\int \mathrm{d}\boldsymbol{u}\, G(\boldsymbol{u}) = \int \mathrm{d}\boldsymbol{v} \mathcal{J}_{\boldsymbol{uv}}\, \tilde{G}(\boldsymbol{v}),} \tag{3.75}$$

donde $\tilde{G}(\boldsymbol{v}) := G(\boldsymbol{U}(\boldsymbol{v}))$ es simplemente la función G escrita en las variables $\boldsymbol{v}$, y donde

$$\mathcal{J}_{\boldsymbol{uv}} := \left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}}\right| = \det \begin{pmatrix} \frac{\partial U_1}{\partial v_1} & \cdots & \frac{\partial U_1}{\partial v_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial U_n}{\partial v_1} & \cdots & \frac{\partial U_n}{\partial v_n} \end{pmatrix} \tag{3.76}$$

es el determinante de la *matriz jacobiana*. Al usar la notación $|\partial \boldsymbol{u}/\partial \boldsymbol{v}|$ para $\mathcal{J}_{\boldsymbol{uv}}$ tenemos la regla mnemotécnica

$$\mathrm{d}\boldsymbol{u} \to \mathrm{d}\boldsymbol{v} \left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}}\right|$$

donde podemos imaginar que $\partial \boldsymbol{u}$ toma el papel de $\mathrm{d}\boldsymbol{u}$, y se agregan $\mathrm{d}\boldsymbol{v}$ y $\partial \boldsymbol{v}$ de forma que se «cancelan».

En nuestro ejemplo de la transformación de coordenadas cartesianas a polares esféricas, el determinante es

$$\left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}}\right| = \left|\frac{\partial(u_1, u_2, u_3)}{\partial(v_1, v_2, v_3)}\right| = v_1^2 \sin v_2 = r^2 \sin\theta, \tag{3.77}$$

por lo que

$$\int \mathrm{d}x\mathrm{d}y\mathrm{d}z\; G(x, y, z) \to \int \mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi\; (r^2 \sin\theta)\; \tilde{G}(r, \theta, \phi). \tag{3.78}$$

Para ver cómo surge la ecuación (3.75) podemos hacer el mismo análisis que nos llevó a (3.70). Tomemos la integral original de $G(\boldsymbol{u})$ en las variables $\boldsymbol{u}$ y agreguemos un 1 como la integral de una delta de Dirac en las nuevas variables $\boldsymbol{v}$,

$$\int \mathrm{d}\boldsymbol{u} G(\boldsymbol{u}) = \int \mathrm{d}\boldsymbol{u} \underbrace{\left[\int \mathrm{d}\boldsymbol{v} \delta(\boldsymbol{v} - \boldsymbol{V}(\boldsymbol{u}))\right]}_{=1} G(\boldsymbol{u}). \tag{3.79}$$

Cambiando el orden de integración y reemplazando (3.71) en $G(\boldsymbol{u})$, tenemos

$$\begin{aligned}\int \mathrm{d}\boldsymbol{u} G(\boldsymbol{u}) &= \int \mathrm{d}\boldsymbol{v} \left[\int \mathrm{d}\boldsymbol{u}\delta(\boldsymbol{v}-\boldsymbol{V}(\boldsymbol{u}))\right] G(\boldsymbol{U}(\boldsymbol{V}(\boldsymbol{u})))\\ &= \int \mathrm{d}\boldsymbol{v} \left[\int \mathrm{d}\boldsymbol{u}\delta(\boldsymbol{v}-\boldsymbol{V}(\boldsymbol{u}))\right] G(\boldsymbol{U}(\boldsymbol{v}))\\ &= \int \mathrm{d}\boldsymbol{v} \left[\int \mathrm{d}\boldsymbol{u}\delta(\boldsymbol{v}-\boldsymbol{V}(\boldsymbol{u}))\right] \tilde{G}(\boldsymbol{v}).\end{aligned} \tag{3.80}$$

Transformando la delta de Dirac a una sobre las variables $\boldsymbol{u}$, y usando que la aproximación de $\boldsymbol{V}$ a primer orden en torno a un punto $\boldsymbol{u}_0$ es

$$\boldsymbol{V}(\boldsymbol{u}) \approx \boldsymbol{V}(\boldsymbol{u}_0) + \mathbb{J}_{vu}(\boldsymbol{u}_0)\cdot(\boldsymbol{u}-\boldsymbol{u}_0), \tag{3.81}$$

notamos además que el único vector $\boldsymbol{u}$ que es solución de $\boldsymbol{v} = \boldsymbol{V}(\boldsymbol{u})$ es

$$\boldsymbol{u}_0 = \boldsymbol{U}(\boldsymbol{v}),$$

por lo tanto

$$\begin{aligned}\delta(\boldsymbol{v}-\boldsymbol{V}(\boldsymbol{u})) &= \delta\left(\cancel{\boldsymbol{v}} - \left[\cancel{\boldsymbol{V}(\boldsymbol{u}_0)} + \mathbb{J}_{vu}\cdot(\boldsymbol{u}-\boldsymbol{u}_0)\right]\right)\\ &= \delta(\mathbb{J}_{vu}\cdot(\boldsymbol{u}-\boldsymbol{U}(\boldsymbol{v})))\\ &= \frac{\delta(\boldsymbol{u}-\boldsymbol{U}(\boldsymbol{v}))}{|\mathbb{J}_{vu}|}\end{aligned} \tag{3.82}$$

y se tiene

$$\int \mathrm{d}\boldsymbol{u}\,\delta(\boldsymbol{v}-\boldsymbol{V}(\boldsymbol{u})) = \left(\frac{1}{\mathcal{J}_{vu}}\right)_{\boldsymbol{u}=\boldsymbol{U}(v)} = \left|\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}}\right|. \tag{3.83}$$

Reemplazando (3.83) en (3.80) obtenemos (3.75).

3.3.3 Densidad de puntos

El siguiente caso donde usaremos la técnica de introducir una delta de Dirac para transformar una integral no es un cambio de variables como los entendemos tradicionalmente. Consideremos la integral

$$Z = \int \mathrm{d}\boldsymbol{u} G(f(\boldsymbol{u})) \tag{3.84}$$

donde el integrando depende de las variables $\boldsymbol{u}$ sólo a través de una función $f(\boldsymbol{u})$. También aquí podemos agregar un 1 como la integral de una delta de Dirac, pero ahora en la variable F que representa los posibles valores de $f(\boldsymbol{u})$,

$$Z = \int \mathrm{d}\boldsymbol{u} G(f(\boldsymbol{u})) = \int \mathrm{d}\boldsymbol{u} G(f(\boldsymbol{u})) \underbrace{\left[\int_{-\infty}^{\infty} \mathrm{d}F\,\delta(F-f(\boldsymbol{u}))\right]}_{=1}, \tag{3.85}$$

y cambiando el orden de integración, podemos escribir Z como una integral unidimensional,

$$\begin{aligned} Z &= \int \mathrm{d}\boldsymbol{u} G(f(\boldsymbol{u})) && (3.86)\\ &= \int_{-\infty}^{\infty} \mathrm{d}F \int \mathrm{d}\boldsymbol{u} \underbrace{G(f(\boldsymbol{u}))}_{=G(F)} \delta(F - f(\boldsymbol{u})) \\ &= \int_{-\infty}^{\infty} \mathrm{d}F \left[\int \mathrm{d}\boldsymbol{u}\delta(F - f(\boldsymbol{u}))\right] G(F). && (3.87) \end{aligned}$$

Es decir, tenemos[3]

$$\boxed{\int \mathrm{d}\boldsymbol{u} G(f(\boldsymbol{u})) = \int_{-\infty}^{\infty} \mathrm{d}F \mathcal{D}(F) G(F),} \tag{3.88}$$

donde la cantidad $\mathcal{D}(F)$ es la *densidad de puntos* de la función f en torno a F.

Definición 3.4 — Densidad de puntos de una función
Definimos la densidad de puntos de $f(\boldsymbol{u})$ alrededor del valor $f = F$ como

$$\mathcal{D}(F) := \int \mathrm{d}\boldsymbol{u}\delta(F - f(\boldsymbol{u})). \tag{3.89}$$

Esta densidad de puntos puede reescribirse, usando (3.53), como

$$\mathcal{D}(F) = \sum_{\boldsymbol{u}_F} \frac{1}{|\nabla f(\boldsymbol{u}_F)|} \tag{3.90}$$

donde $\boldsymbol{u}_F$ son los puntos que cumplen $f(\boldsymbol{u}_F) = F$ y donde $\nabla f = \dfrac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$.

La regla mnemotécnica queda en este caso

$$\mathrm{d}\boldsymbol{u} \to \mathrm{d}F\mathcal{D}(F)$$

Podemos ver que efectivamente $\mathcal{D}(F)$ es una densidad de puntos si calculamos su integral en un intervalo $[F_1, F_2]$,

$$\begin{aligned} \int_{F_1}^{F_2} \mathrm{d}F\mathcal{D}(F) &= \int_{F_1}^{F_2} \mathrm{d}F\mathrm{d}\boldsymbol{u}\delta(F - f(\boldsymbol{u})) \\ &= \int \mathrm{d}\boldsymbol{u} \int_{F_1}^{F_2} \mathrm{d}F\delta(F - f(\boldsymbol{u})) \\ &= \int \mathrm{d}\boldsymbol{u}\Big[\Theta(F_2 - f(\boldsymbol{u})) - \Theta(F_1 - f(\boldsymbol{u}))\Big] \\ &= \int \mathrm{d}\boldsymbol{u}\,\mathrm{rect}(f(\boldsymbol{u}); F_1, F_2) \\ &= \int\limits_{f\in[F_1,F_2]} \mathrm{d}\boldsymbol{u}, \end{aligned} \tag{3.91}$$

que es precisamente el volumen de puntos tal que $f(\boldsymbol{u}) \in [F_1, F_2]$.

[3] Aquí $\mathcal{D}(F)$ puede interpretarse como una medida de integración.

Ejemplo 3.3.1. *La integral en tres dimensiones de una función radial*

$$G = G\left(\sqrt{x^2+y^2+z^2}\right)$$

puede transformarse en una integral unidimensional sobre r,

$$\begin{aligned} Z &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \int_{-\infty}^{\infty} \mathrm{d}z\; G\left(\sqrt{x^2+y^2+z^2}\right) \\ &= \int_0^{\infty} \mathrm{d}r \int_0^{\pi} \mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\phi\; r^2 \sin\theta\; G(r) \\ &= \int_0^{\infty} \mathrm{d}r \underbrace{(4\pi r^2)}_{=\mathcal{D}(r)} G(r), \end{aligned} \tag{3.92}$$

por lo que la densidad radial de puntos $\mathcal{D}(r)$ en 3 dimensiones es $\mathcal{D}(r) = 4\pi r^2$, que precisamente corresponde al área de la esfera de radio r,

$$\mathcal{D}(r) = 4\pi r^2 = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \int_{-\infty}^{\infty} \mathrm{d}z\; \delta\left(\sqrt{x^2+y^2+z^2} - r\right). \tag{3.93}$$

Aquí también es claro ver que la interpretación de $\mathcal{D}(r)$ como una densidad tiene sentido, ya que su integral entre R_1 y R_2,

$$\int_{R_1}^{R_2} \mathrm{d}r\; 4\pi r^2 = \frac{4\pi {R_2}^3}{3} - \frac{4\pi {R_1}^3}{3} \tag{3.94}$$

es precisamente el volumen de la región que se encuentra entre los radios R_1 y R_2.

3.4 — Derivada dentro de una integral

Para calcular integrales, una herramienta útil a tener en cuenta es derivar una integral parcialmente con respecto a un parámetro[(4)]. Más precisamente, nos referimos a usar la fórmula

$$\frac{\partial}{\partial\alpha}\int_a^b \mathrm{d}x f(x;\alpha) = \int_a^b \mathrm{d}x \left[\frac{\partial f(x;\alpha)}{\partial\alpha}\right], \tag{3.95}$$

donde los límites de integración a y b no dependen de α.

Ejemplo 3.4.1. *Si conocemos la integral*

$$\int_0^{\infty} \mathrm{d}t \exp(-\lambda t) = \frac{1}{\lambda}, \tag{3.96}$$

calculemos la integral

$$R = \int_0^{\infty} \mathrm{d}t \exp(-\lambda t) t. \tag{3.97}$$

Solución. *Tenemos que*

$$\exp(-\lambda t)t = -\frac{\partial}{\partial\lambda}\exp(-\lambda t), \tag{3.98}$$

luego podemos escribir R como

$$R = -\frac{\partial}{\partial\lambda}\int_0^{\infty} \mathrm{d}t \exp(-\lambda t) = -\frac{\partial}{\partial\lambda}\left(\frac{1}{\lambda}\right) = \frac{1}{\lambda^2}. \tag{3.99}$$

[(4)] Esta herramienta fue popularizada por el físico Richard Feynman, por lo que a veces se le conoce como el *truco de Feynman*.

Lo sorprendente de este truco es que es posible usarlo en maneras insospechadas. Por ejemplo, ¿qué pasa si no hay un parámetro α respecto al cual derivar? ¡Simplemente nos podemos inventar uno!

Ejemplo 3.4.2. *Considerando la integral*

$$\int_0^\infty \mathrm{d}x \exp(-\alpha x) \sin x = \frac{1}{1+\alpha^2} \tag{3.100}$$

calculemos

$$I = \int_0^\infty \mathrm{d}x \left(\frac{\sin x}{x}\right). \tag{3.101}$$

Solución. *Podemos generalizar nuestra integral a*

$$I(\alpha) := \int_0^\infty \mathrm{d}x \exp(-\alpha x) \frac{\sin x}{x} \tag{3.102}$$

de forma que $I = I(0)$. Pero

$$\begin{aligned} -\frac{\partial I(\alpha)}{\partial \alpha} &= -\frac{\partial}{\partial \alpha} \int_0^\infty \mathrm{d}x \exp(-\alpha x) \frac{\sin x}{x} \\ &= \int_0^\infty \mathrm{d}x \exp(-\alpha x) \sin x = \frac{1}{1+\alpha^2}, \end{aligned} \tag{3.103}$$

luego integrando en α tenemos que

$$I(\alpha) = -\arctan \alpha + C, \tag{3.104}$$

y para evaluar la constante C, vemos que, de acuerdo a (3.102),

$$\lim_{\alpha \to \infty} I(\alpha) = 0 = C - \lim_{\alpha \to \infty} \arctan \alpha = C - \frac{\pi}{2} \tag{3.105}$$

luego $C = \frac{\pi}{2}$. Finalmente la integral que buscamos es

$$I = I(0) = \frac{\pi}{2} - \arctan 0 = \frac{\pi}{2}. \tag{3.106}$$

3.5 — Funciones especiales

En esta sección visitaremos algunas de las funciones especiales útiles para nuestro futuro desarrollo en inferencia, dado que aparecen frecuentemente en el estudio de modelos. En primer lugar definiremos la función gamma de Euler como sigue.

Definición 3.5 — Función gamma
La función gamma de Euler se define por la integral

$$\Gamma(z) := \int_0^\infty \mathrm{d}t \exp(-t) t^{z-1}, \tag{3.107}$$

válida para cualquier complejo *excepto cero y los enteros negativos*.

Esta función entrega una generalización del factorial de números enteros, a través de la propiedad

$$\boxed{\Gamma(n) = (n-1)!} \tag{3.108}$$

Usando la propiedad del factorial

$$n! = n(n-1)!$$

se sigue que

$$\Gamma(n+1) = n\,\Gamma(n). \tag{3.109}$$

Veamos cómo demostrar (3.108) como un ejemplo de la técnica de inducción matemática.

Demostración (3.108). Primero, notemos que

$$\Gamma(1) = \int_0^\infty \mathrm{d}t \exp(-t) = 1 = 0!, \tag{3.110}$$

luego (3.108) es cierta para $n = 1$. A continuación, debemos mostrar que (3.108) para $n = k$ se sigue de (3.108) para $n = k-1$, es decir, debemos demostrar $\Gamma(k) = (k-1)!$ **a partir de la suposición**

$$\Gamma(k-1) = (k-2)! \tag{3.111}$$

lo cual podemos hacer simplemente escribiendo $\Gamma(k)$ e integrando por partes,

$$\begin{aligned}
\Gamma(k) &= \int_0^\infty \mathrm{d}t \exp(-t)t^{k-1} \\
&= \underbrace{-\exp(-t)t^{k-1}\Big|_0^\infty}_{\to 0} + \int_0^\infty \mathrm{d}t \exp(-t)(k-1)t^{k-2} \\
&= (k-1)\int_0^\infty \mathrm{d}t \exp(-t)t^{k-2} \\
&= (k-1)\Gamma(k-1) \\
\text{usando (3.111)} \quad &= (k-1)(k-2)! \\
&= (k-1)! \qquad \checkmark
\end{aligned}$$

donde en la segunda igualdad hemos usado integración por partes escogiendo $u = t^{k-1}$ y $v = -\exp(-t)$.

Ahora introduciremos la función beta, también conocida como la integral de Euler de primer tipo, la cual es una integral en el intervalo $[0,1]$ como sigue.

Definición 3.6 — Función beta
La función beta se define como

$$B(x,y) := \int_0^1 \mathrm{d}t\; t^{x-1}(1-t)^{y-1} \tag{3.112}$$

para x, y complejos tales que $\mathrm{Re}(x) > 0, \mathrm{Re}(y) > 0$.

La conexión entre la función beta y la función gamma de Euler está dada por la relación

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{3.113}$$

Demostración (3.113). Escribimos el producto de dos funciones gamma como

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \left[\int_0^\infty \mathrm{d}t \exp(-t)t^{x-1}\right]\left[\int_0^\infty \mathrm{d}s \exp(-s)s^{y-1}\right] \\ &= \int_0^\infty \mathrm{d}t \int_0^\infty \mathrm{d}s \exp(-t-s)t^{x-1}s^{y-1},\end{aligned} \tag{3.114}$$

luego hacemos el cambio de variable $t = zu$, $s = z(1-u)$, el cual tiene determinante de la *matriz jacobiana*

$$\left|\frac{\partial(t,s)}{\partial(z,u)}\right| = z, \tag{3.115}$$

para obtener

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \int_0^\infty \mathrm{d}t \int_0^\infty \mathrm{d}s \exp(-t-s)t^{x-1}s^{y-1} \\ &= \int_0^\infty \mathrm{d}z \int_0^1 \mathrm{d}u\; z\exp(-z(u+1-u))(zu)^{x-1}(z(1-u))^{y-1} \\ &= \left[\int_0^\infty \mathrm{d}z \exp(-z)z^{x+y-1}\right]\left[\int_0^1 \mathrm{d}u u^{x-1}(1-u)^{y-1}\right] \\ &= \Gamma(x+y)B(x,y)\end{aligned}$$

Usando (3.113) podemos escribir otra cantidad importante para nosotros, el **coeficiente binomial**

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} \tag{3.116}$$

en términos de la función beta como

$$\binom{n}{k} = \frac{1}{(n-k)B(k+1,n-k)}. \tag{3.117}$$

Para ver esto, simplemente usamos (3.113) y (3.108) para desarrollar la

función beta en el denominador,

$$\begin{aligned} B(k+1, n-k) &= \frac{\Gamma(k+1)\Gamma(n-k)}{\Gamma(k+1+n-k)} \\ &= \frac{k!(n-k-1)!}{(n!)} \\ &= \frac{k!(n-k)!}{(n-k)(n!)}. \end{aligned}$$

El coeficiente binomial es el número de ordenamientos (permutaciones) posibles de k elementos en un total de n. Por ejemplo, el número de secuencias como «00101001» que contienen exactamente tres veces 1 y 5 veces 0 (total de 8 dígitos) es

$$\binom{8}{3} = \frac{8!}{3!5!} = 56.$$

La deducción de este resultado se muestra en el **Ejemplo 4.1.3**.

3.6 — Método de los multiplicadores de Lagrange

Recordaremos en esta sección cómo encontrar los puntos donde una función continua tiene su máximo o mínimo, comenzando con funciones de una variable real para luego generalizar a funciones de n variables. El punto x_0 que hace que la función continua $f(x)$ sea un máximo o mínimo debe cumplir la *condición de extremo*

$$\left(\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right)_{x=x_0} = 0, \tag{3.118}$$

que gráficamente se interpreta como el hecho que la recta tangente a f en x_0 sea horizontal. Para determinar si el extremo x_0 corresponde a un máximo o a un mínimo, es necesario mirar el signo de la segunda derivada de f en dicho punto, de forma que si

$$f''(x_0) := \left(\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right)_{x=x_0} > 0$$

el punto x_0 es un mínimo de f, y es un máximo si $f''(x_0) < 0$.

Para una función de n variables $f(x_1, \ldots, x_n)$, que denotaremos simplemente como $f(\boldsymbol{x})$, la condición de extremo se vuelve

$$\left(\frac{\partial f(\boldsymbol{x})}{\partial x_k}\right)_{\boldsymbol{x}=\boldsymbol{x}_0} = 0 \qquad \text{para todo } k = 1, \ldots, n, \tag{3.119}$$

es decir, es un sistema de n ecuaciones y n incógnitas, estas últimas las componentes del punto $\boldsymbol{x}_0$, que podemos escribir en forma vectorial como

$$\Big(\nabla f(\boldsymbol{x})\Big)_{\boldsymbol{x}=\boldsymbol{x}_0} = \boldsymbol{0}. \tag{3.120}$$

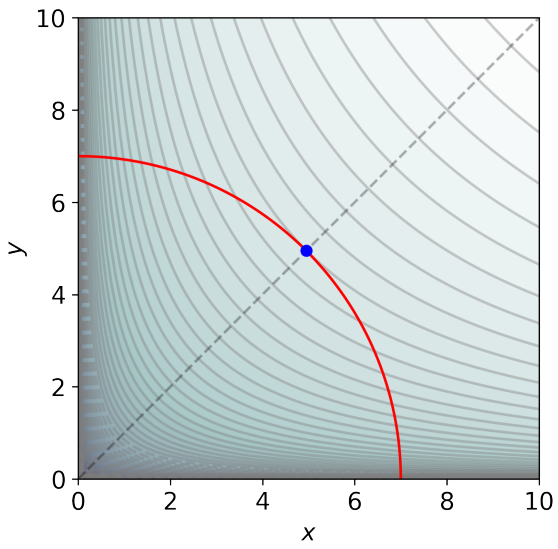


Figura 3.3: Maximización del valor de $f(x,y) = 5\ln xy$ para los puntos tales que $x^2 + y^2 = 7$, que se encuentran en la curva en rojo. El máximo ocurre en el punto marcado en azul, y es tal que los gradientes de las curvas $g(x,y) = x^2 + y^2$ y $f(x,y)$ apuntan ambos en la dirección $\hat{\boldsymbol{n}} = (\hat{\boldsymbol{x}} + \hat{\boldsymbol{y}})/\sqrt{2}$.

Definir máximos o mínimos ya no es tan directo en n dimensiones porque ahora existe la posibilidad de que un extremo $\boldsymbol{x}_0$ sea un punto de ensilladura, esto es, un punto donde el signo de la segunda derivada (parcial) depende de la dirección en que miremos. De hecho ahora debemos considerar las componentes de la *matriz hessiana* en $\boldsymbol{x}_0$,

$$H_{ij}(\boldsymbol{x}_0) := \left(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right)_{\boldsymbol{x}=\boldsymbol{x}_0}, \tag{3.121}$$

de forma que $\boldsymbol{x}_0$ sólo será un máximo o mínimo si las componentes H_{ij} tienen igual signo para todo i, j.

Supongamos ahora que queremos encontrar ya sea los máximos o mínimos de $f(\boldsymbol{x})$, pero tales que cumplan la *restricción*

$$g(\boldsymbol{x}) = G \tag{3.122}$$

donde G es un valor constante. Para esto existe el método de los multiplicadores de Lagrange, el cual se basa en la observación de que en los puntos extremos $\boldsymbol{x}_0$ de f tales que $g(\boldsymbol{x}_0) = G$ se cumple que los vectores ∇f y ∇g son colineales, esto es, paralelos o antiparalelos, como se ve en la **Figura 3.3**.

En el método de los multiplicadores de Lagrange, construimos la *función ampliada* de n+1 variables

$$\tilde{f}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda\Big(g(\boldsymbol{x}) - G\Big) \tag{3.123}$$

donde la nueva variable λ se conoce como el multiplicador de Lagrange asociado a la restricción (3.122). Los extremos de esta nueva función deben cumplir con n+1 ecuaciones. Por un lado, debe cumplirse

$$\frac{\partial \tilde{f}}{\partial x_k} = \frac{\partial f}{\partial x_k} - \lambda \frac{\partial g}{\partial x_k} = 0 \tag{3.124}$$

para $k = 1, \ldots, n$, es decir,

$$\nabla f(\boldsymbol{x}_0) = \lambda \nabla g(\boldsymbol{x}_0), \tag{3.125}$$

lo cual garantiza que los gradientes de f y g son colineales. Por otro lado, también debe cumplirse

$$\frac{\partial \tilde{f}}{\partial \lambda} = 0, \tag{3.126}$$

que es precisamente la restricción (3.122), en la forma $g(\boldsymbol{x}_0) - G = 0$.

De esta forma, sólo debemos preocuparnos de buscar los extremos de $\tilde{f}$ considerada como una función de $\boldsymbol{x}$, y luego buscar el (único) valor de λ que asegura que se cumpla la restricción.

Ejemplo 3.6.1. *Encontremos los puntos (x, y) en el cuadrante $x > 0, y > 0$ tal que el valor de la función $f(x, y) = \alpha \ln(xy)$ con $\alpha > 0$ es máximo pero el punto está a una distancia R del origen.*

Solución. *Nuestra función ampliada es*

$$\tilde{f}(x, y, \lambda) = \alpha \ln(xy) - \lambda\left(x^2 + y^2 - R^2\right), \tag{3.127}$$

la cual debemos extremar con respecto a x e y. Obtenemos las ecuaciones

$$\left(\frac{\partial \tilde{f}(x, y)}{\partial x}\right)_{x=x_0, y=y_0} = \frac{\alpha}{x_0} - 2\lambda x_0 = 0, \tag{3.128}$$

$$\left(\frac{\partial \tilde{f}(x, y)}{\partial y}\right)_{x=x_0, y=y_0} = \frac{\alpha}{y_0} - 2\lambda y_0 = 0, \tag{3.129}$$

esto es,

$$x_0^2 = y_0^2 = \frac{\alpha}{2\lambda}, \tag{3.130}$$

luego cumplir la restricción nos obliga a fijar

$$x_0^2 + y_0^2 = \frac{\alpha}{\lambda} = R^2 \tag{3.131}$$

con lo que tenemos $\lambda = \alpha / R^2$. Por lo tanto el único punto extremo en el cuadrante es

$$\boldsymbol{x}_0 = \frac{R}{\sqrt{2}}(1, 1) \tag{3.132}$$

que efectivamente es el máximo de $f(x, y) = \alpha \ln(xy)$ a la distancia R del origen.

Este problema puede resolverse también sin usar multiplicadores de Lagrange, pasando a coordenadas polares

$$x = R\cos\theta, \tag{3.133}$$

$$y = R\sin\theta \tag{3.134}$$

y buscando el ángulo $\theta \in [0, \frac{\pi}{2}]$ que maximiza

$$F(\theta) := f(R\cos\theta, R\sin\theta) = \alpha \ln(R^2 \sin\theta \cdot \cos\theta).$$

Tenemos

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\alpha \ln(R^2 \sin\theta \cdot \cos\theta)\right] = \alpha\left(\frac{\cos\theta_0}{\sin\theta_0} - \frac{\sin\theta_0}{\cos\theta_0}\right) = 0 \tag{3.135}$$

con lo que se obtiene $(\tan\theta_0)^2 = 1$, con solución $\theta_0 = \frac{\pi}{4}$, que precisamente corresponde al punto en (3.132).

El método de los multiplicadores de Lagrange se extiende naturalmente para m restricciones de la forma

$$g_j(\boldsymbol{x}) = G_j \tag{3.136}$$

para $j = 1, \ldots, m$. En este caso la función ampliada

$$\tilde{f}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{j=1}^{m} \lambda_j \Big(g_j(\boldsymbol{x}) - G_j \Big) \tag{3.137}$$

posee m multiplicadores de Lagrange $\lambda_1, \ldots, \lambda_m$, los cuales se fijan con las ecuaciones

$$\frac{\partial \tilde{f}(\boldsymbol{x}, \boldsymbol{\lambda})}{\partial \lambda_j} = 0 \tag{3.138}$$

para $j = 1, \ldots, m$.

3.7 — Funcionales y cálculo de variaciones

Ahora introduciremos el concepto de *funcional*, que podemos entender como una «función de orden superior» cuyo argumento es una función. Esto es, un funcional $\mathcal{F}$ asigna a cada función $x(\bullet)$ en su dominio un número real $\mathcal{F}[x]$. Por ejemplo, el funcional

$$\mathcal{F}_1[x] := \int_0^1 \mathrm{d}t \, x(t)^2 \tag{3.139}$$

asigna a la función $x_1(t) = 2t$ el valor

$$\mathcal{F}_1[x_1] = \int_0^1 \mathrm{d}t \, (2t)^2 = \frac{4}{3},$$

mientras que a la función $x_2(t) = \exp(-t)$ le asigna el valor

$$\mathcal{F}_1[x_2] = \int_0^1 \mathrm{d}t \, \exp(-2t) = \frac{1}{2}\Big(1 - \frac{1}{e^2}\Big) \approx 0.43233.$$

Así como para una función $f(x)$ podemos pensar en el cambio en f cuando x cambia ligeramente, que por supuesto es el concepto de derivada, el equivalente para funcionales es la *derivada funcional*

$$\frac{\delta \mathcal{F}[x]}{\delta x(t)},$$

que representa el cambio en el valor que entrega el funcional $\mathcal{F}[x]$ cuando la función $x(\bullet)$ cambia infinitesimalmente en el punto t. La derivada funcional se define formalmente a través de la *variación*[5]

$$\delta \mathcal{F} := \int \mathrm{d}t \left(\frac{\delta \mathcal{F}[x]}{\delta x(t)} \right) \eta(t) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \Big(\mathcal{F}[x + \epsilon \cdot \eta] - \mathcal{F}[x] \Big) \tag{3.140}$$

donde $\eta(\bullet)$ es una *función de prueba* arbitraria.

[5] A veces denominada *primera variación*.

Cálculo vectorial		Cálculo de variaciones	
Vector	$\boldsymbol{x}$	Función	$x(\bullet)$
Componente del vector	x_i	Función evaluada	$x(t)$
Función vectorial	$f(\boldsymbol{x})$	Funcional	$\mathcal{F}[x]$
Diferencial total	$\mathrm{d}f$	Variación	$\delta\mathcal{F}$
Componente del gradiente	$\dfrac{\partial f}{\partial x_i}$	Derivada funcional	$\dfrac{\delta\mathcal{F}}{\delta x(t)}$

Tabla 3.1: Correspondencia entre conceptos aplicables a funciones vectoriales y a funcionales.

Si pensamos en un vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ de n componentes como una función de un argumento discreto $i = 1, \ldots, n$, podemos ahora interpretar una función continua $x(\bullet)$ como si fuera un vector de infinitas componentes, esto es, si promovemos el índice entero i a un parámetro continuo $t \in [a, b]$, tomando el límite $n \to \infty$ de forma tal que

$$x(a) = x_1, \tag{3.141a}$$
$$x(b) = x_n. \tag{3.141b}$$

Entonces, un funcional $\mathcal{F}$ corresponde a la generalización de una función vectorial $f(\boldsymbol{x})$. En esta interpretación, la variación $\delta\mathcal{F}$ no es más que la generalización del *diferencial total* de $f(\boldsymbol{x})$,

$$\mathrm{d}f = \sum_{i=1}^{n} \mathrm{d}x_i \frac{\partial f(\boldsymbol{x})}{\partial x_i} = \mathrm{d}\boldsymbol{x} \cdot \nabla f(\boldsymbol{x}) \tag{3.142}$$

mientras que la derivada funcional no es más que la generalización de la componente i-ésima del gradiente,

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} \longrightarrow \frac{\delta\mathcal{F}[x]}{\delta x(t)},$$

como puede verse en la **Tabla 3.1**.

De hecho, podemos obtener una definición alternativa de la derivada funcional si hacemos la identificación concreta

$$t := a + (i-1)\Delta_n, \tag{3.143a}$$
$$\Delta_n := \frac{b-a}{n-1}, \tag{3.143b}$$

de forma que $\Delta_n \to 0$ cuando $n \to \infty$. En tal caso tenemos que (**3.140**) se reduce a

$$\frac{\delta\mathcal{F}[x]}{\delta x(t)} = \lim_{n\to\infty} \left[\frac{1}{\Delta_n} \frac{\partial f(\boldsymbol{x})}{\partial x_i} \right]_t \tag{3.144}$$

donde el índice i en el lado derecho es el que corresponde a t de acuerdo a (**3.143a**).

Demostración. Comenzamos estableciendo la correspondencia entre un funcional y una función de infinitas variables, con lo que

$$\mathcal{F}[x+\epsilon\eta] = \lim_{n\to\infty} f(\boldsymbol{x}+\epsilon\boldsymbol{\eta}). \tag{3.145}$$

Ahora usamos el hecho que ϵ es suficientemente pequeño para justificar aproximar $f(\boldsymbol{x}+\epsilon\boldsymbol{\eta})$ en Taylor a primer orden,

$$f(\boldsymbol{x}+\epsilon\boldsymbol{\eta}) \approx f(\boldsymbol{x}) + \epsilon\boldsymbol{\eta}\cdot\nabla f(\boldsymbol{x}) \tag{3.146}$$

con lo que podemos formar

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(f(\boldsymbol{x}+\epsilon\boldsymbol{\eta}) - f(\boldsymbol{x})\Big) = \boldsymbol{\eta}\cdot\nabla f(\boldsymbol{x}) \tag{3.147}$$

y entonces se tiene

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(\mathcal{F}[x+\epsilon\cdot\eta] - \mathcal{F}[x]\Big) = \lim_{n\to\infty}\boldsymbol{\eta}\cdot\nabla f(\boldsymbol{x}). \tag{3.148}$$

Expandiendo el lado derecho tenemos

$$\lim_{n\to\infty}\boldsymbol{\eta}\cdot\nabla f(\boldsymbol{x}) = \lim_{n\to\infty}\sum_{i=1}^{n}\eta_i\frac{\partial f(\boldsymbol{x})}{\partial x_i} = \lim_{n\to\infty}\sum_{i=1}^{n}\varepsilon_n\eta_i\left[\frac{1}{\varepsilon_n}\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right], \tag{3.149}$$

y si ahora agregamos un 1 como la integral de una delta de Dirac en t, podemos reescribir

$$\begin{aligned}
\lim_{n\to\infty}\boldsymbol{\eta}\cdot\nabla f(\boldsymbol{x}) &= \lim_{n\to\infty}\sum_{i=1}^{n}\underbrace{\int_a^b \mathrm{d}t\,\delta(t-a-(i-1)\varepsilon_n)}_{=1}\,\varepsilon_n\eta_i\left[\frac{1}{\varepsilon_n}\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right] \\
&= \int_a^b \mathrm{d}t\,\eta(t)\left\{\lim_{n\to\infty}\sum_{i=1}^{n}\varepsilon_n\delta(t-a-(i-1)\varepsilon_n)\left[\frac{1}{\varepsilon_n}\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right]\right\} \\
&= \int_a^b \mathrm{d}t\,\eta(t)\left\{\lim_{n\to\infty}\left[\frac{1}{\varepsilon_n}\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right]_t\right\}\underbrace{\lim_{n\to\infty}\sum_{i=1}^{n}\varepsilon_n\delta(t-a-(i-1)\varepsilon_n)}_{=\int_a^b \mathrm{d}t'\,\delta(t-t')=1},
\end{aligned} \tag{3.150}$$

donde en la última línea hemos usado la suma de Riemann para evaluar la suma de la delta de Dirac, por medio del cambio de variable $t' := (i-1)\varepsilon_n + a$. Finalmente llegamos a

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(\mathcal{F}[x+\epsilon\cdot\eta] - \mathcal{F}[x]\Big) = \int_a^b \mathrm{d}t\,\eta(t)\left\{\lim_{n\to\infty}\left[\frac{1}{\varepsilon_n}\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right]_t\right\} \tag{3.151}$$

y comparando (3.151) con (3.140) demostramos (3.144) ✔

Veamos ahora cómo calcular derivadas funcionales de manera práctica. Primero calculemos la derivada funcional de $\mathcal{F}_1[x]$ en (3.139) usando la definición formal. Para esto, aproximamos $\mathcal{F}[x+\epsilon\cdot\eta]$ a primer orden en ϵ, obteniendo

$$\begin{aligned}\mathcal{F}_1[x+\epsilon\cdot\eta] &= \int_0^1 \mathrm{d}t\,(x(t)+\epsilon\eta(t))^2 \\ &= \int_0^1 \mathrm{d}t\Big(x(t)^2+2\epsilon x(t)\eta(t)+\underbrace{\epsilon^2}_{\approx 0}\eta(t)^2\Big) \\ &\approx \underbrace{\int_0^1 \mathrm{d}t x(t)^2}_{=\mathcal{F}_1[x]} + 2\epsilon\int_0^1 \mathrm{d}t\,x(t)\eta(t)\end{aligned} \tag{3.152}$$

con lo que

$$\mathcal{F}_1[x+\epsilon\cdot\eta]-\mathcal{F}_1[x] = 2\epsilon\int_0^1 \mathrm{d}t\,x(t)\eta(t) \tag{3.153}$$

y entonces

$$\int \mathrm{d}t\left(\frac{\delta\mathcal{F}[x]}{\delta x(t)}\right)\eta(t) = 2\int_0^1 \mathrm{d}t\,x(t)\eta(t) = 2\int \mathrm{d}t\,x(t)\eta(t)\,\mathrm{rect}(t;0,1). \tag{3.154}$$

Es decir, tenemos que

$$\int \mathrm{d}t\,\eta(t)\left[\frac{\delta\mathcal{F}[x]}{\delta x(t)} - 2x(t)\,\mathrm{rect}(t;0,1)\right] = 0 \tag{3.155}$$

para todo $\eta(\bullet)$, y por lo tanto el paréntesis cuadrado debe ser idénticamente cero. Se sigue entonces que

$$\frac{\delta\mathcal{F}[x]}{\delta x(t)} = 2x(t)\,\mathrm{rect}(t;0,1) \quad \text{para todo } t. \tag{3.156}$$

Veamos un atajo interesante para el cálculo de derivadas funcionales. Para funcionales de la forma

$$\mathcal{F}[x] = \int \mathrm{d}t L(x(t)), \tag{3.157}$$

la derivada funcional es simplemente

$$\frac{\delta\mathcal{F}[x]}{\delta x(t)} = \left(\frac{\mathrm{d}L}{\mathrm{d}x}\right)_{x=x(t)}, \tag{3.158}$$

que corresponde al uso de la regla de la cadena

$$\frac{\delta}{\delta x(t)}\left(\int \mathrm{d}t' L(x(t'))\right) = \int \mathrm{d}t'\left(\frac{\mathrm{d}L}{\mathrm{d}x}\right)_{x=x(t')}\frac{\delta x(t')}{\delta x(t)} = \left(\frac{\mathrm{d}L}{\mathrm{d}x}\right)_{x=x(t)} \tag{3.159}$$

donde hemos incorporado una nueva regla,

$$\boxed{\frac{\delta x(t')}{\delta x(t)} = \delta(t'-t).} \tag{3.160}$$

Demostración. Nuevamente aproximamos $\mathcal{F}[x+\epsilon\cdot\eta]$ a primer orden en ϵ,

$$\begin{aligned}\mathcal{F}[x+\epsilon\cdot\eta] &= \int \mathrm{d}t L(x(t)+\epsilon\eta(t)) \\ &\approx \int \mathrm{d}t \left(L(x)+\epsilon\eta(t)\frac{\mathrm{d}L}{\mathrm{d}x}\right)_{x=x(t)}\end{aligned} \tag{3.161}$$

luego

$$\frac{1}{\epsilon}\Big(\mathcal{F}[x+\epsilon\cdot\eta]-\mathcal{F}[x]\Big) = \int \mathrm{d}t\,\eta(t)\left(\frac{\mathrm{d}L}{\mathrm{d}x}\right)_{x=x(t)}, \tag{3.162}$$

y comparando con (3.140),

$$\frac{\delta\mathcal{F}[x]}{\delta x(t)} = \left(\frac{\mathrm{d}L}{\mathrm{d}x}\right)_{x=x(t)}$$

Adicionalmente, podemos verificar (3.160) evaluando (3.140) directamente para $\mathcal{F}[x]=x(t')$,

$$\int \mathrm{d}t\,\eta(t)\frac{\delta x(t')}{\delta x(t)} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(\cancel{x(t')}+\epsilon\eta(t')-\cancel{x(t')}\Big) = \eta(t'). \tag{3.163}$$

En un caso más general, para funcionales de la forma

$$\mathcal{F}[x] = \int \mathrm{d}t L(x(t),\dot{x}(t)). \tag{3.164}$$

se cumple

$$\frac{\delta\mathcal{F}[x]}{\delta x(t)} = \partial_1 L(x(t),\dot{x}(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_2 L(x(t),\dot{x}(t))\Big) \tag{3.165}$$

con

$$\partial_1 L(x,v) := \frac{\partial L(x,v)}{\partial x}, \qquad \partial_2 L(x,v) := \frac{\partial L(x,v)}{\partial v}, \tag{3.166}$$

y la condición de extremo del funcional $\mathcal{F}[x]$ es la renombrada *ecuación de Euler-Lagrange*,

$$\boxed{\partial_1 L(x(t),\dot{x}(t)) = \frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_2 L(x(t),\dot{x}(t))\Big).} \tag{3.167}$$

Esto puede interpretarse como el uso de la regla de la cadena con

$$\frac{\delta\dot{x}(t')}{\delta x(t)} = \delta'(t'-t), \tag{3.168}$$

que también podemos verificar evaluando (3.140) directamente para el funcional $\mathcal{F}[x]=\dot{x}(t')$,

$$\int \mathrm{d}t\,\eta(t)\frac{\delta\dot{x}(t')}{\delta x(t)} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(\cancel{\dot{x}(t')}+\epsilon\dot{\eta}(t')-\cancel{\dot{x}(t')}\Big) = \dot{\eta}(t'). \tag{3.169}$$

Demostración. Aproximamos $\mathcal{F}[x+\epsilon\cdot\eta]$ a primer orden en ϵ, esta vez en ambos argumentos de L,

$$\begin{aligned}\mathcal{F}[x+\epsilon\cdot\eta] &= \int \mathrm{d}t\, L\big(x(t)+\epsilon\eta(t), \dot{x}(t)+\epsilon\dot{\eta}(t)\big)\\ &\approx \int \mathrm{d}t\, \Big(L(x(t)) + \epsilon\eta(t)\partial_1 L(x(t),\dot{x}(t))\\ &\qquad + \epsilon\dot{\eta}(t)\partial_2 L(x(t),\dot{x}(t))\Big)\end{aligned} \tag{3.170}$$

luego se obtiene

$$\begin{aligned}\frac{1}{\epsilon}\Big(\mathcal{F}[x+\epsilon\cdot\eta]-\mathcal{F}[x]\Big) = \int \mathrm{d}t\, \Big(&\eta(t)\partial_1 L(x(t),\dot{x}(t))\\ &+\dot{\eta}(t)\partial_2 L(x(t),\dot{x}(t))\Big).\end{aligned} \tag{3.171}$$

Como en el caso anterior, el primer término aporta $\partial_1 L$ a la derivada funcional, mientras que el segundo término puede integrarse por partes, obteniendo

$$\int \mathrm{d}t\, \dot{\eta}(t)\partial_2 L(x(t),\dot{x}(t)) = -\int \mathrm{d}t\, \eta(t)\frac{\mathrm{d}}{\mathrm{d}t}\Big(\partial_2 L(x(t),\dot{x}(t))\Big) \tag{3.172}$$

con lo que finalmente se tiene

$$\frac{1}{\epsilon}\Big(\mathcal{F}[x+\epsilon\cdot\eta]-\mathcal{F}[x]\Big) = \int \mathrm{d}t\, \eta(t)\Big(\partial_1 L - \frac{\mathrm{d}}{\mathrm{d}t}\big[\partial_2 L\big]\Big) \tag{3.173}$$

y luego (3.165).

Ejemplo 3.7.1. *Calculemos la derivada del funcional*

$$\mathcal{F}[x] = \int_{-\infty}^{\infty} \mathrm{d}t\, x(t)\sqrt{1+\dot{x}(t)^2}. \tag{3.174}$$

Tenemos, de acuerdo a (3.165), que

$$\begin{aligned}\frac{\delta\mathcal{F}[x]}{\delta x(t)} &= \frac{\partial}{\partial x}\Big(x\sqrt{1+v^2}\Big)_{x=x(t),v=\dot{x}(t)} - \frac{d}{dt}\Big(\frac{\partial}{\partial v}x\sqrt{1+v^2}\Big)_{x=x(t),v=\dot{x}(t)}\\ &= \sqrt{1+\dot{x}(t)^2} - \frac{d}{dt}\Big(x\frac{\partial}{\partial v}\sqrt{1+v^2}\Big)_{x=x(t),v=\dot{x}(t)}\\ &= \sqrt{1+\dot{x}(t)^2} - \frac{\dot{x}(t)^2}{\sqrt{1+\dot{x}(t)^2}} - x(t)\frac{d}{dt}\left(\frac{\dot{x}(t)}{\sqrt{1+\dot{x}(t)^2}}\right)\\ &= \frac{1+\dot{x}(t)^2 - x(t)\ddot{x}(t)}{\big(\sqrt{1+\dot{x}(t)^2}\big)^3}\end{aligned} \tag{3.175}$$

luego los extremos de $\mathcal{F}[x]$ son soluciones de la ecuación diferencial

$$\ddot{x}(t) = \frac{1+\dot{x}(t)^2}{x(t)}. \tag{3.176}$$

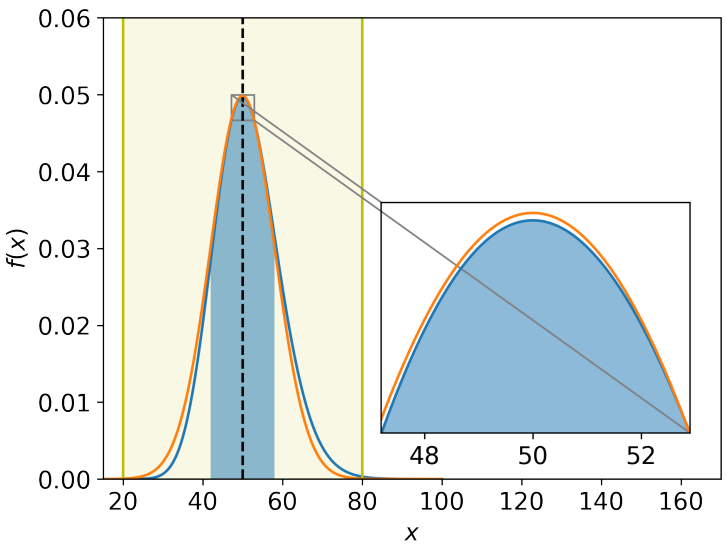


Figura 3.4: La curva en naranjo es la aproximación de Laplace de la función en azul, la cual se desea integrar en el intervalo $[a, b]$, denotado por las líneas roja y verde. La banda en azul corresponde al intervalo $[x_0 - s, x_0 + s]$, con s definido en (3.183).

3.8 — La aproximación de Laplace

La idea detrás de la aproximación de Laplace es sencilla: vista desde cerca (como si miráramos con una lupa) cualquier función en torno a su máximo se ve como una parábola, usando una expansión de Taylor a segundo orden. Si fuera de ese máximo la función decae rápidamente podemos aproximar la integral de dicha función como una integral gaussiana. Más precisamente, consideremos la integral unidimensional

$$Z = \int_a^b \mathrm{d}x f(x), \tag{3.177}$$

donde el integrando $f(x) \geq 0$ está fuertemente concentrado en el punto $x = x_0$, el cual es un máximo de f, como se ve en la **Figura 3.4**. Esto es,

$$\left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)_{x=x_0} = 0, \tag{3.178a}$$

$$\left(\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right)_{x=x_0} < 0. \tag{3.178b}$$

Como $f(x)$ es no negativa, podemos reescribirla como

$$f(x) = \exp(\ln f(x)) \tag{3.179}$$

y, en lugar de expandir la propia función f podemos expandir $\ln f$ a segundo orden en torno a $x = x_0$. De esta forma obtenemos

$$\ln f(x) \approx \ln f(x_0) + (x - x_0)\left(\frac{\mathrm{d}}{\mathrm{d}x}\ln f\right)_{x=x_0} + \frac{1}{2}(x - x_0)^2\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2}\ln f\right)_{x=x_0}. \tag{3.180}$$

Como x_0 es un máximo de f, también lo es de $\ln f$ y se tiene

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}\ln f\right)_{x=x_0} = 0, \tag{3.181}$$

por la condición de extremo, luego

$$\ln f(x) \approx \ln f(x_0) + \frac{1}{2}(x - x_0)^2 \left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} \ln f \right)_{x=x_0}. \tag{3.182}$$

Como la segunda derivada de $\ln f$ es negativa en x_0, definimos

$$\frac{1}{s^2} := -\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} \ln f(x) \right)_{x=x_0}, \tag{3.183}$$

con lo que nuestra aproximación a $\ln f$ puede escribirse como

$$\ln f(x) \approx \ln f(x_0) - \frac{1}{2s^2}(x - x_0)^2, \tag{3.184}$$

y finalmente reemplazando en la integral Z, tenemos

$$\begin{aligned} Z = \int_a^b \mathrm{d}x \exp(\ln f(x)) &\approx f(x_0) \int_a^b \mathrm{d}x \exp\left(-\frac{1}{2s^2}(x - x_0)^2 \right) \\ &\approx f(x_0) \int_{-\infty}^{\infty} \mathrm{d}x \exp\left(-\frac{1}{2s^2}(x - x_0)^2 \right) \\ \text{usando (12)} \quad &= \sqrt{2\pi} s \, f(x_0), \end{aligned} \tag{3.185}$$

donde hemos hecho uso de la condición de que f está suficientemente concentrada en torno a $x = x_0$ tal que se cumple que $|f(a)| \approx 0$ y $|f(b)| \approx 0$, lo cual a su vez nos permite extender los límites de integración de a hasta $-\infty$, y de b hasta ∞. Claramente esto sólo se cumplirá si $|a - x_0| \gg s$ y $|b - x_0| \gg s$.

Recuadro 3.1 — Aproximación de Laplace
Si la función $f(x)$ posee un máximo suficientemente agudo en $x = x_0$, podemos aproximar la integral $Z = \int_a^b \mathrm{d}x f(x)$ como

$$Z \approx f(x_0) \frac{\sqrt{2\pi}}{\sqrt{\left(-\frac{\mathrm{d}^2}{\mathrm{d}x^2} \ln f \right)_{x=x_0}}}. \tag{3.186}$$

En d dimensiones, la aproximación de Laplace se generaliza a

$$Z = \int_U \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) \int_U \mathrm{d}\boldsymbol{x} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\intercal \mathbb{H}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \right) \tag{3.187}$$

donde $\mathbb{H}$ es la matriz de componentes

$$H_{ij}(\boldsymbol{x}) = -\frac{\partial^2}{\partial x_i \partial x_j} \ln f(\boldsymbol{x}), \tag{3.188}$$

con lo que usando la integral gaussiana multidimensional tenemos

$$Z \approx f(\boldsymbol{x}_0) \frac{\sqrt{(2\pi)^d}}{\sqrt{\det \mathbb{H}(\boldsymbol{x}_0)}}. \tag{3.189}$$

► Para más detalles sobre la aproximación de Laplace, ver el libro de MacKay (2003).

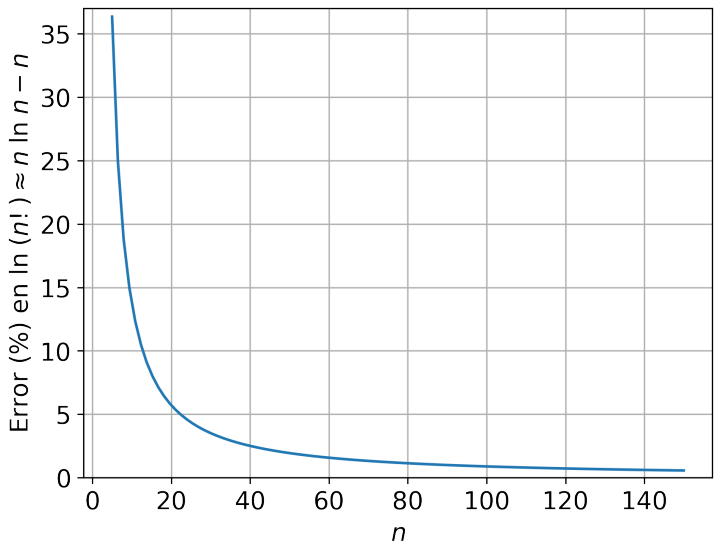


Figura 3.5: Error porcentual cometido al aproximar $\ln(n!)$ por $n\ln n - n$.

3.9 — La aproximación de Stirling

El objetivo es obtener una buena aproximación del factorial $n!$ para $n \to \infty$. Comenzamos expresando el factorial en términos de la función gamma de Euler según (3.108),

$$n! = \Gamma(n+1) = \int_0^\infty \mathrm{d}t\; t^n \exp(-t) = \int_0^\infty \mathrm{d}t \exp(n\ln t - t), \tag{3.190}$$

y usando el cambio de variable $t = nu$, tenemos que

$$n! = \int_0^\infty \mathrm{d}t \exp(n\ln t - t) = \exp(n\ln n)n \int_0^\infty \mathrm{d}u \exp(n[\ln u - u]). \tag{3.191}$$

Vemos que la función $f(u) := \exp(n[\ln u - u])$ para $n \to \infty$ está cada vez más concentrada en $u_0 = 1$, ya que

$$\left(\frac{\partial}{\partial u}\ln f(u)\right)_{u=u_0} = n\left(\frac{1}{u_0} - 1\right) = 0, \tag{3.192}$$

luego $u_0 = 1$, y

$$\left(\frac{\partial^2}{\partial u^2}\ln f(u)\right)_{u=u_0} = -\frac{n}{u_0^2} = -n \tag{3.193}$$

con lo que se obtiene que $s^2 = 1/n$. Entonces, usando la aproximación de Laplace, encontramos que

$$\int_0^\infty \mathrm{d}u \exp(n[\ln u - u]) \approx \sqrt{2\pi}sf(u_0) = \sqrt{\frac{2\pi}{n}}\exp(-n), \tag{3.194}$$

y reemplazando en (3.191) obtenemos

$$n! = \exp(n\ln n)n\int_0^\infty \mathrm{d}u \exp(n[\ln u - u]) \approx \sqrt{2\pi n}\exp(n\ln n - n). \tag{3.195}$$

Si ahora aplicamos logaritmo a ambos lados tenemos

$$\ln(n!) \approx \ln(\sqrt{2\pi}) + \left(n + \tfrac{1}{2}\right)\ln n - n, \tag{3.196}$$

pero para $n \to \infty$ se tiene que $n + \frac{1}{2} \approx n$ y el término constante $\ln \sqrt{2\pi}$ es despreciable, luego obtenemos la forma más conocida de la aproximación de Stirling.

Recuadro 3.2 — Aproximación de Stirling
Para $n \to \infty$ es posible aproximar $\ln(n!)$ como

$$\ln(n!) \approx n \ln n - n. \tag{3.197}$$

3.10 — Funciones generadoras

Para una secuencia $f_0, f_1, f_2, \ldots$ definimos la función generadora $F(t)$ como

$$F(t) := \sum_{n=0}^{\infty} f_n \, t^n, \tag{3.198}$$

donde t es un parámetro continuo. Esta función generadora es una manera conveniente de manipular el conjunto de valores de la secuencia y permite, entre otros usos, transformar relaciones de recurrencia en ecuaciones diferenciales. En particular, la derivada de (3.198) nos entrega

$$\begin{aligned} \frac{\mathrm{d}F}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t} \sum_{n=0}^{\infty} f_n t^n \\ &= \sum_{n=1}^{\infty} n f_n t^{n-1} \\ \text{(usando } m = n-1\text{)} \quad &= \sum_{m=0}^{\infty} (m+1) f_{m+1} t^m, \end{aligned} \tag{3.199}$$

mientras que la segunda derivada es

$$\begin{aligned} \frac{\mathrm{d}^2F}{\mathrm{d}t^2} &= \sum_{n=2}^{\infty} n(n-1) f_n t^{n-2} \\ \text{(usando } m = n-2\text{)} \quad &= \sum_{m=0}^{\infty} (m+2)(m+1) f_{m+2} t^m. \end{aligned} \tag{3.200}$$

Otras relaciones convenientes son

$$t \frac{\mathrm{d}F}{\mathrm{d}t} = \sum_{n=0}^{\infty} n f_n t^n, \tag{3.201}$$

$$t^2 \frac{\mathrm{d}^2F}{\mathrm{d}t^2} = \sum_{n=0}^{\infty} n(n-1) f_n t^n, \tag{3.202}$$

y en general

$$t^k \frac{\mathrm{d}^k F}{\mathrm{d}t^k} = \sum_{n=0}^{\infty} \left(\frac{n!}{(n-k)!} \right) f_n t^n. \tag{3.203}$$

Ejemplo 3.10.1. *la función generadora de la secuencia geométrica* $1, a, a^2, a^3, \ldots$ *es*

$$F(t) = \sum_{n=0}^{\infty} a^n t^n = \frac{1}{1-at}. \tag{3.204}$$

Ejemplo 3.10.2. *Queremos encontrar una forma cerrada para los* c_k *que cumplen la relación de recurrencia*

$$c_{k+1} = b\left(\frac{c_k}{k+1}\right), \qquad c_0 = 2. \tag{3.205}$$

Escribimos la relación de recurrencia usando como índice la variable n*, de la forma*

$$c_{n+1}(n+1) = bc_n, \tag{3.206}$$

y luego multiplicamos por t^n *y sumamos sobre* $n = 0, 1, 2, \ldots$*,*

$$\sum_{n=0}^{\infty} bc_n t^n = \sum_{n=0}^{\infty} c_{n+1}(n+1)t^n. \tag{3.207}$$

Definiendo la función generadora de la secuencia $\{c_i\}$ *como*

$$F(t) := \sum_{n=0}^{\infty} c_n t^n, \tag{3.208}$$

tenemos

$$\begin{aligned} bF(t) &= \sum_{n=0}^{\infty} c_{n+1}(n+1)t^n = \sum_{n=1}^{\infty} c_n n t^{n-1} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{n=0}^{\infty} c_n t^n = \frac{\mathrm{d}}{\mathrm{d}t}F(t). \end{aligned} \tag{3.209}$$

Resolviendo esta ecuación diferencial de primer orden para $F(t)$ *tenemos*

$$F(t) = F(0)\exp(bt), \tag{3.210}$$

que a continuación podemos expandir para encontrar los coeficientes c_n*,*

$$\sum_{n=0}^{\infty} c_n t^n = F(0)\sum_{n=0}^{\infty} \frac{(bt)^n}{n!} \tag{3.211}$$

con lo que se tiene, comparando potencias iguales de t*, que*

$$\sum_{n=0}^{\infty} \Big\{ \underbrace{c_n - F(0)\frac{b^n}{n!}}_{=0 \text{ para todo } n} \Big\} t^n = 0. \tag{3.212}$$

Finalmente tenemos

$$c_n = F(0)\left(\frac{b^n}{n!}\right), \qquad n = 0, 1, 2, \ldots \tag{3.213}$$

donde usamos la condición inicial $c_0 = F(0) = 2$ *para escribir*

$$c_n = 2\left(\frac{b^n}{n!}\right). \tag{3.214}$$

3.11 — Transformadas integrales y convolución

Recordaremos aquí los conceptos de transformada integral lineal de una función, y de convolución entre dos funciones. Diremos que $T[f]$ es una transformada integral si es un funcional de f de la forma

$$T[f](k) = \int_{-\infty}^{\infty} \mathrm{d}x K(x;k) f(x) \tag{3.215}$$

donde $K(x;k)$ se denomina el *kernel* de la transformada. Esta transformada entonces toma una función de la variable x y produce una nueva función de la variable k. Por otro lado, la convolución $f * g$ entre dos funciones f y g es una nueva función que se define como

$$(f * g)(x) = \int_{-\infty}^{\infty} \mathrm{d}x' f(x') g(x - x'). \tag{3.216}$$

Para una transformada $T[f]$ bajo ciertas condiciones se cumple el *teorema de convolución*.

Teorema 3.1 — Teorema de convolución
Si el *kernel* K de una transformada $T[f]$ definida según (3.215) cumple

$$K(x + y; k) = K(x;k) K(y;k) \tag{3.217}$$

entonces se tiene

$$T[f * g] = T[f] \cdot T[g], \tag{3.218}$$

esto es, la transformada de la convolución es el producto de las transformadas.

Demostración.

$$\begin{aligned}
T[f * g] &= \int_{-\infty}^{\infty} \mathrm{d}z K(z)(f * g)(z) \\
&= \int_{-\infty}^{\infty} \mathrm{d}z K(z) \left[\int_{-\infty}^{\infty} \mathrm{d}x f(x) g(z - x) \right] \\
&= \int_{-\infty}^{\infty} \mathrm{d}z \int_{-\infty}^{\infty} \mathrm{d}x K(z + \underbrace{x - x}_{=0}) f(x) g(z - x) \\
\text{(usando } y = z - x\text{)} \quad &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y K(x + y) f(x) g(y) \\
\text{usando (3.217)} \quad &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y K(x) K(y) f(x) g(y) \\
&= \left[\int_{-\infty}^{\infty} \mathrm{d}x K(x) f(x) \right] \cdot \left[\int_{-\infty}^{\infty} \mathrm{d}y K(y) g(y) \right] \\
&= T[f] \cdot T[g] \qquad \checkmark
\end{aligned} \tag{3.219}$$

La utilidad de este teorema es que, si la transformada T tiene una inversa T^{-1}, podemos calcular la convolución $f * g$ como

$$(f * g) = T^{-1}\Big[T[f] \cdot T[g]\Big]. \tag{3.220}$$

Podemos definir una transformada inversa integral para el *kernel* K como

$$T^{-1}[F](x) = \int_{-\infty}^{\infty} \mathrm{d}k K^{\dagger}(k;x) F(k) \tag{3.221}$$

tal que

$$\int_{-\infty}^{\infty} \mathrm{d}k K^{\dagger}(k;x) K(x';k) = \delta(x - x'), \tag{3.222}$$

de forma que se cumpla que la transformada inversa de la transformada de cualquier función f sea la misma función f,

$$\begin{aligned} T^{-1}\Big[T[f]\Big](x) &= \int_{-\infty}^{\infty} \mathrm{d}k K^{\dagger}(k;x) \int_{-\infty}^{\infty} \mathrm{d}z K(z;k) f(z) \\ &= \int_{-\infty}^{\infty} \mathrm{d}z f(z) \underbrace{\left[\int_{-\infty}^{\infty} \mathrm{d}k K^{\dagger}(k;x) K(z;k)\right]}_{=\delta(z-x)} = f(x). \end{aligned} \tag{3.223}$$

La representación (3.46) de la delta de Dirac sugiere buscar pares K y $K^{\dagger}$ tal que

$$K^{\dagger}(k;x) K(x';k) = \frac{1}{2\pi} \exp(\mathrm{i}k(x - x')), \tag{3.224}$$

por ejemplo es conveniente la siguiente definición

$$K(x;k) := \exp(\mathrm{i}kx) \tag{3.225a}$$

$$K^{\dagger}(k;x') := \frac{1}{2\pi} \exp(-\mathrm{i}kx'), \tag{3.225b}$$

que además cumple

$$\begin{aligned} K(x + y;k) &= \exp(\mathrm{i}k(x + y)) \\ &= \exp(\mathrm{i}kx) \exp(\mathrm{i}ky) \\ &= K(x;k) K(y;k), \end{aligned} \tag{3.226}$$

con lo cual la transformadas T y su inversa T^{-1} son

$$T[f] := \int_{-\infty}^{\infty} \mathrm{d}x \exp(\mathrm{i}kx) f(x), \tag{3.227a}$$

$$T^{-1}[\tilde{f}] := \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx) \tilde{f}(k). \tag{3.227b}$$

Este par de transformadas son esencialmente la transformada de Fourier y su inversa, pero siguiendo una convención particular. En física es más común definirlas de manera simétrica (Riley, Hobson y Bence 2006, pág. 435) y con una convención de signo distinta, como

$$\mathcal{F}[f] := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}x \exp(-\mathrm{i}kx) f(x), \tag{3.228a}$$

$$\mathcal{F}^{-1}[\tilde{f}] := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}k \exp(\mathrm{i}kx) \tilde{f}(k). \tag{3.228b}$$

PROBLEMAS

Problema 3.1. *Exprese la función signo,*

$$\operatorname{sgn}(x) = \begin{cases} 1 \; si \; x > 0, \\ \\ -1 \; si \; x < 0, \end{cases} \tag{3.229}$$

en términos de la función escalón de Heaviside $\Theta(x)$. Utilice este resultado para escribir el valor absoluto de x únicamente en términos de x y $\Theta(x)$.

Problema 3.2. *Demuestre que la delta de Dirac es la derivada de la función escalón de Heaviside, haciendo uso de (3.5) y del teorema fundamental del cálculo.*

Problema 3.3. *Calcule explícitamente el determinante en la ecuación (3.77) usando la transformación en (3.74).*

Problema 3.4. *Demuestre que*

$$\frac{\partial}{\partial z}\Big[\int_0^z \mathrm{d}x f(x;z)\Big] = f(z;z) + \int_0^z \mathrm{d}x \frac{\partial f(x;z)}{\partial z}, \tag{3.230}$$

sin usar el teorema fundamental del cálculo. Cambie los límites con un escalón de Heaviside.

Problema 3.5. *Resuelva las siguientes integrales,*

$$I_1 = \int_{-\infty}^{\infty} \mathrm{d}x \delta(x^2 - 2x - 3), \tag{3.231a}$$

$$I_2 = \int_{0}^{\infty} \mathrm{d}x \delta(x^2 - 2x - 3), \tag{3.231b}$$

$$I_3 = \int_{0}^{4\pi} \mathrm{d}\theta \delta(cos(\theta)). \tag{3.231c}$$

Problema 3.6. *Construya la aproximación de Laplace para la integral*

$$Z = \int_{-\pi}^{\pi} \mathrm{d}x \exp(n \cos x) \tag{3.232}$$

con $n \to \infty$. Compare numéricamente su resultado para distintos valores de n con el valor exacto,

$$Z = 2\pi I_0(n), \tag{3.233}$$

donde I_0 es la función de Bessel modificada de primera especie.

Problema 3.7. *Demuestre que*

$$\operatorname{rect}(\alpha x; -\alpha L, \alpha L) = \operatorname{rect}(x; -L, L), \tag{3.234}$$

y usando este resultado, verifique que el límite en (3.17) es un ejemplo de la familia de límites en (3.36) que dan origen a la delta de Dirac. Escriba la función $\eta(u)$ en este caso y verifique que ésta cumple (3.37).

Problema 3.8. *Muestre que la función beta (3.112) tiende a concentrarse en torno a un valor único cuando $\alpha \to \infty$ y $\beta \to \infty$. Determine dicho valor y obtenga la aproximación de Laplace para $B(\alpha, \beta)$ en esos límites.*

Problema 3.9. *Calcule la integral gaussiana*

$$\int_0^\infty \mathrm{d}u \exp(-u^2) = \frac{\sqrt{\pi}}{2} \tag{3.235}$$

llevándola a una integral en coordenadas esféricas en 3 dimensiones.

Problema 3.10. *Demuestre usando la aproximación de Stirling (3.197) que el coeficiente binomial definido en (3.116) puede aproximarse como*

$$\binom{n}{k} \approx \exp\Big(-n\big(p\ln p + (1-p)\ln(1-p)\big)\Big) \tag{3.236}$$

para $n \to \infty$ y $k \to \infty$ con $p := \dfrac{k}{n} \in [0,1]$.

Problema 3.11. *Demuestre que $x\delta(x) = 0$.*

Problema 3.12. *Demuestre que $\Gamma\left(1/2\right) = \sqrt{\pi}$.*

Problema 3.13. *Demuestre usando la relación de recurrencia*

$$a_{n+2} = a_{n+1} + a_n$$

que cumple la secuencia de Fibonacci, que la función generadora asociada a dicha secuencia es

$$F(t) = \frac{1}{1-(t+t^2)}.$$

Problema 3.14. *Calcule la convolución entre dos funciones rectangular $\mathrm{rect}(\bullet; a, b)$.*

Problema 3.15. *Demuestre (3.57) usando (3.54) y la técnica de inducción matemática.*

Problema 3.16. *Demuestre la relación (3.203) para una función generadora $F(t)$ definida según (3.198).*

Problema 3.17. *Demuestre por inducción matemática que la densidad de puntos de la función $f(x_1, \ldots, x_N) = \sum_{i=1}^N x_i$ con $x_i \geq 0$ es*

$$\mathcal{D}_N(S) = \int_0^\infty \mathrm{d}x_1 \int_0^\infty \mathrm{d}x_2 \ldots \int_0^\infty \mathrm{d}x_N\, \delta\left(\sum_{i=1}^N x_i - S\right) = \frac{S^{N-1}}{(N-1)!}. \tag{3.237}$$

Capítulo 4

Variables discretas y continuas

Come, Watson, come! The game is afoot!

Sherlock Holmes, The Return of Sherlock Holmes

Ya hemos conquistado nuestras herramientas matemáticas, y ahora haremos uso de ellas para conectar la idea de proposiciones lógicas con las variables discretas y continuas a las que estamos acostumbrados. El concepto clave aquí será el de *funciones indicador*.

4.1 — Funciones indicador

Introduciremos ahora un concepto fundamental para el desarrollo de los siguientes capítulos, la función indicador, mediante la cual podemos transformar identidades de la lógica en identidades aritméticas.

Definición 4.1 — Función indicador
Para una proposición lógica A definiremos la función indicador $\mathrm{Q}(A)$ como

$$\mathrm{Q}(A) = \begin{cases} 1 \text{ si } A = \mathbb{T} \\ \\ 0 \text{ si } A = \mathbb{F}. \end{cases} \tag{4.1}$$

En otras palabras, la función indicador simplemente define una transformación (mapa) que lleva $\mathbb{T}$ hacia 1 y $\mathbb{F}$ hacia 0. Por ejemplo, la negación en la Tabla 2.1 puede transformarse en la siguiente identidad,

$$\mathrm{Q}(\neg A) = 1 - \mathrm{Q}(A), \tag{4.2}$$

mientras que la conjunción en la Tabla 2.1 se reduce a

$$\mathrm{Q}(A \wedge B) = \mathrm{Q}(A)\mathrm{Q}(B). \tag{4.3}$$

Condición	Función indicador	Notación algebraica
$n = m$	$\mathrm{Q}(n = m)$	$\delta(n, m)$
$x \geq a$	$\mathrm{Q}(x \geq a)$	$\Theta(x - a)$
$x \leq b$	$\mathrm{Q}(x \leq b)$	$\Theta(b - x)$
$x \in [a, b]$	$\mathrm{Q}(x \in [a, b])$	$\mathrm{rect}(x; a, b)$

Tabla 4.1: Algunas funciones indicador útiles.

Para determinar la función indicador de la disyunción, usamos (4.2), (4.3) y una de las leyes de De Morgan, (2.9b),

$$\begin{aligned}
\mathrm{Q}(A \vee B) &= 1 - \mathrm{Q}(\neg(A \vee B)) \\
&= 1 - \mathrm{Q}((\neg A) \wedge (\neg B)) \\
&= 1 - \mathrm{Q}(\neg A)\mathrm{Q}(\neg B) \\
&= 1 - (1 - \mathrm{Q}(A))(1 - \mathrm{Q}(B)) \\
&= \cancel{1} - \Big\{\cancel{1} - \mathrm{Q}(A) - \mathrm{Q}(B) + \mathrm{Q}(A)\mathrm{Q}(B)\Big\},
\end{aligned} \tag{4.4}$$

por lo tanto,

$$\boxed{\mathrm{Q}(A \vee B) = \mathrm{Q}(A) + \mathrm{Q}(B) - \mathrm{Q}(A \wedge B),} \tag{4.5}$$

que podemos verificar que está en acuerdo con la Tabla 2.1. La función indicador de la implicación es

$$\mathrm{Q}(A \Rightarrow B) = \mathrm{Q}((\neg A) \vee B) = \Big(1 - \mathrm{Q}(A)\Big)\Big(1 - \mathrm{Q}(B)\Big) + \mathrm{Q}(B), \tag{4.6}$$

con lo que automáticamente vemos que $\mathrm{Q}(A \Rightarrow B) = 1$ si $\mathrm{Q}(A) = 0$, que es el *principio de explosión*.

La Tabla 4.1 muestra algunas de las funciones indicador comunes en matemática. Entre ellas, la función indicador de la igualdad entre variables discretas es probablemente la más conocida, y se le llama la *delta de Kronecker*,

$$\mathrm{Q}(n = m) = \begin{cases} 1 \text{ si } n = m, \\ \\ 0 \text{ si } n \neq m. \end{cases} = \delta(n, m) \tag{4.7}$$

Usando las propiedades (4.2), (4.3) y (4.5) podemos crear nuevas funciones indicador. Por ejemplo, a partir de la función indicador (4.7) fácilmente construimos

$$\begin{aligned}
\mathrm{Q}(n \neq m) &= \mathrm{Q}(\neg(n = m)) \\
&= 1 - \mathrm{Q}(n = m) \\
&= 1 - \delta(n, m).
\end{aligned} \tag{4.8}$$

Ejemplo 4.1.1. *La función indicador asociada a la proposición*

$$A = \text{«}Y \text{ no es negativo y es menor o igual que } \exp(X)\text{»}$$

es

$$\mathrm{Q}(A) = \mathrm{Q}(Y \geq 0)\mathrm{Q}(Y \leq \exp(X)) = \Theta(Y)\Theta(\exp(X) - Y). \tag{4.9}$$

Ejemplo 4.1.2. *La función indicador de pertenencia a un intervalo $[a, b]$ es la función rectangular*

$$\mathrm{Q}(x \in [a, b]) = \mathrm{rect}(x; a, b). \tag{4.10}$$

Sin embargo, sabemos que un intervalo puede escribirse como la intersección o la unión de otros intervalos, y esto nos permite deducir otras propiedades de la función rectangular. Si $[b, c]$ es la intersección de dos intervalos $[a, c]$ y $[b, d]$ con $a < b < c < d$, entonces

$$\begin{aligned} \mathrm{Q}(x \in [b, c]) &= \mathrm{Q}((x \in [a, c]) \wedge (x \in [b, d])) \\ &= \mathrm{Q}(x \in [a, c])\mathrm{Q}(x \in [b, d]) \\ &= \mathrm{rect}(x; a, c)\,\mathrm{rect}(x; b, d) \end{aligned} \tag{4.11}$$

pero por otro lado $\mathrm{Q}(x \in [b, c]) = \mathrm{rect}(x; b, c)$ por lo que obtenemos

$$\mathrm{rect}(x; a, c)\,\mathrm{rect}(x; b, d) = \mathrm{rect}(x; b, c).$$

Ejemplo 4.1.3 (Coeficiente binomial)**.** *Queremos determinar el número de secuencias con n dígitos, k de los cuales son 1 y el resto son 0, que sabemos es el coeficiente binomial (3.116). Para una secuencia dada $e_1, e_2, \ldots, e_n$ con $e_i \in \{0, 1\}$ contaremos las ocurrencias de 1 usando la suma*

$$k = \sum_{i=1}^{n} e_i, \tag{4.12}$$

por medio de la cual la condición de que existan k ocurrencias de verdadero en n repeticiones estará dada por

$$\mathrm{Q}(e_1 + \ldots + e_n = k).$$

Ahora podemos definir el número $C(k, n)$ de secuencias con k verdaderos entre n realizaciones,

$$C(k, n) = \sum_{e_1=0}^{1} \sum_{e_2=0}^{1} \ldots \sum_{e_n=0}^{1} \mathrm{Q}(e_1 + e_2 + \ldots + e_{n-1} + e_n = k). \tag{4.13}$$

Desarrollando explícitamente la suma sobre e_n, tenemos

$$\begin{aligned} C(k,n) &= \sum_{e_1=0}^{1}\sum_{e_2=0}^{1}\cdots\sum_{e_{n-1}=0}^{1}\Big\{Q(e_1+\ldots+e_{n-1}+0=k) \\ &\qquad + Q(e_1+\ldots+e_{n-1}\underbrace{+1}=k)\Big\} \\ &= \sum_{e_1=0}^{1}\sum_{e_2=0}^{1}\cdots\sum_{e_{n-1}=0}^{1}\Big\{Q(e_1+\ldots+e_{n-1}=k) \\ &\qquad + Q(e_1+\ldots+e_{n-1}=k\underbrace{-1})\Big\}, \end{aligned} \tag{4.14}$$

esto es, se tiene la relación de recurrencia

$$C(k,n) = C(k,n-1) + C(k-1,n-1), \tag{4.15}$$

conocida como la fórmula de Pascal, *la cual resolveremos utilizando la técnica de la función generadora. Definiendo*

$$F_n(t) := \sum_{k=0}^{\infty} C(k,n)t^k \tag{4.16}$$

podemos multiplicar (4.15) por t^k y sumar, obteniendo

$$\begin{aligned} \sum_{k=0}^{\infty} C(k,n)t^k &= \sum_{k=0}^{\infty} C(k,n-1)t^k + \sum_{k=0}^{\infty} C(k-1,n-1)t^k \\ &\hookrightarrow F_n(t) = F_{n-1}(t) + \sum_{k'=0}^{\infty} C(k',n-1)t^{k'+1} \\ &\hookrightarrow F_n(t) = (1+t)F_{n-1}(t). \end{aligned} \tag{4.17}$$

Notando que

$$F_1(t) = \sum_{k=0}^{\infty} C(k,1)t^k = C(0,1) + C(1,1)t = 1+t \tag{4.18}$$

vemos que

$$F_n(t) = \sum_{k=0}^{\infty} C(k,n)t^k = (1+t)^n \tag{4.19}$$

y al usar el teorema del binomio para expandir $F_n(t)$ tenemos

$$F_n(t) = (1+t)^n = \sum_{k=0}^{n} \binom{n}{k} t^k \cancel{1^{n-k}}^{1} \tag{4.20}$$

luego, comparando con (4.16), se tiene

$$C(k,n) = \binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{4.21}$$

4.2 — Cambios de dominio de integración

Una de las propiedades más útiles de las funciones indicador es que permiten seleccionar un subconjunto U de una región o dominio de integración Ω. Sea una región Ω que contiene a una región $U \subset \Omega$. Usando la función indicador $\mathrm{Q}(\boldsymbol{x} \in U)$ para la pertenencia a U podemos escribir

$$\begin{aligned}\int_{\Omega} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \mathrm{Q}(\boldsymbol{x} \in U) &= \int_{U} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \cancelto{1}{\mathrm{Q}(\boldsymbol{x} \in U)} + \int_{\Omega - U} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \cancelto{0}{\mathrm{Q}(\boldsymbol{x} \in U)} \\ &= \int_{U} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}), \end{aligned} \tag{4.22}$$

donde $\Omega - U$ es el complemento de la región U que forma parte de Ω. Tenemos entonces que la regla universal para «proyectar» hacia un subdominio de integración $U \subset \Omega$ es

$$\text{Si } U \subset \Omega, \text{ entonces } \int_{U} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) = \int_{\Omega} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \mathrm{Q}(\boldsymbol{x} \in U). \tag{4.23}$$

Además de poder unificar (3.5), (3.6), (3.10) y (3.12), esta propiedad nos permite escribir (3.25) como

$$\int_{U} \mathrm{d}\boldsymbol{x} \delta(\boldsymbol{x} - \boldsymbol{x}_0) = \int_{\Omega} \mathrm{d}\boldsymbol{x} \delta(\boldsymbol{x} - \boldsymbol{x}_0) \mathrm{Q}(\boldsymbol{x} \in U) = \mathrm{Q}(\boldsymbol{x}_0 \in U), \tag{4.24}$$

lo cual a su vez nos entrega una interpretación de la delta de Dirac multidimensional como el límite

$$\delta(\boldsymbol{x} - \boldsymbol{x}_0) = \lim_{V \to 0} \frac{1}{V} \mathrm{Q}(\boldsymbol{x} \in \mathcal{V}), \tag{4.25}$$

donde $\mathcal{V}$ es una vecindad de volumen pequeño V en torno al punto $\boldsymbol{x}_0$.

4.3 — De proposiciones a variables

Procederemos ahora a expresar la idea de variables, tanto discretas como continuas, como proposiciones lógicas y usando la función indicador obtendremos algunas identidades que serán claves para el desarrollo de nuestra teoría de la inferencia. Antes de desarrollar esta idea, presentaremos ahora una propiedad realmente sencilla, pero que sin embargo nos será de utilidad en lo que sigue.

Lema 4.1. *Si $a \in \{0, 1\}$ entonces*

$$a\, f(a) = a\, f(1). \tag{4.26}$$

Claramente esto es así, ya que para $a = 0$ se tiene $0 = 0$, y para $a = 1$ se tiene $f(1) = f(1)$. Un corolario importante de este lema es que

$$\mathrm{Q}(A) f(\boldsymbol{u}; A) = \mathrm{Q}(A) f(\boldsymbol{u}; \mathbb{T}) \tag{4.27}$$

donde f es una función arbitraria cuyo valor depende de la proposición A. Vemos que la función indicador «proyecta» a la función f al caso donde $A = \mathbb{T}$. Usando este resultado podemos deducir

$$\mathrm{Q}(x = x_0) f(x) = \mathrm{Q}(x = x_0) f(x_0), \tag{4.28}$$

para una variable discreta x, que es el análogo discreto de (3.28).

Volvamos ahora al problema de la descripción de variables numéricas en términos de proposiciones. Consideremos el caso de una variable discreta que toma uno de n valores permitidos, $X \in \{x_1, \ldots, x_n\}$. Definiendo la proposición

$$X_i := \text{«la variable } X \text{ toma el valor } x_i\text{»}$$

podemos ver que las proposiciones X_i y X_j con $i \neq j$ son mutuamente excluyentes,

$$X_i \wedge X_j = \mathbb{F} \quad \text{si } i \neq j, \tag{4.29}$$

ya que X no puede tomar el valor x_i y el valor x_j a la vez si $i \neq j$. Por supuesto, si $i = j$, se tiene $X_i \wedge X_i = X_i$. Además, el conjunto completo de proposiciones $X_1, X_2, \ldots, X_n$ son exhaustivas, esto es,

$$X_1 \vee \ldots \vee X_n = \mathbb{T}, \tag{4.30}$$

dado que el valor de X debe estar en el conjunto $\{x_1, \ldots, x_n\}$. Aplicando función indicador a ambos lados se tiene

$$\mathrm{Q}(X_1 \vee \ldots \vee X_n) = \mathrm{Q}(\mathbb{T}) = 1, \tag{4.31}$$

pero

$$\begin{aligned} \mathrm{Q}(X_1 \vee \ldots \vee X_n) &= \mathrm{Q}(X_1) + \mathrm{Q}(X_2 \vee \ldots \vee X_n) - \mathrm{Q}(\underbrace{X_1 \wedge [X_2 \vee \ldots \vee X_n]}_{=\mathbb{F}}) \\ &= \mathrm{Q}(X_1) + \mathrm{Q}(X_2 \vee \ldots \vee X_n), \end{aligned} \tag{4.32}$$

ya que X_1 es mutuamente excluyente con $X_2 \vee \ldots \vee X_n$. Aplicando (4.32) iterativamente vemos que

$$\mathrm{Q}(X_1 \vee X_2 \vee \ldots \vee X_n) = \mathrm{Q}(X_1) + \mathrm{Q}(X_2) + \ldots + \mathrm{Q}(X_n) = \sum_{i=1}^{n} \mathrm{Q}(X_i), \tag{4.33}$$

y por lo tanto

$$\sum_{i=1}^{n} \mathrm{Q}(X_i) = 1. \tag{4.34}$$

Esto podemos escribirlo de una forma más descriptiva, que utilizaremos en el **Capítulo 5**, como

$$\boxed{\sum_{i=1}^{n} \mathrm{Q}(X = x_i) = 1.} \tag{4.35}$$

Función indicador	**delta de Dirac**
$\sum_{i=1}^{n} \mathrm{Q}(X = x_i) = 1$	$\int_{-\infty}^{\infty} \mathrm{d}x\delta(X - x) = 1$
$\sum_{i=1}^{n} \mathrm{Q}(X = x_i)x_i = X$	$\int_{-\infty}^{\infty} \mathrm{d}x\delta(X - x)x = X$
$\sum_{i=1}^{n} \mathrm{Q}(X = x_i)f(x_i) = f(X)$	$\int_{-\infty}^{\infty} \mathrm{d}x\delta(X - x)f(x) = f(X)$

Tabla 4.2: Comparación entre propiedades de funciones indicador y la delta de Dirac.

La interpretación de este resultado es la siguiente: dado un valor de X, digamos $X = x_k$, sólo el término k-ésimo de la sumatoria de la izquierda será igual a 1, mientras que todos los demás términos serán iguales a cero, resultando la suma siempre igual a 1 (sin importar cuál sea k). Más aún, es posible deducir otra propiedad interesante. Multiplicando (4.35) por X a ambos lados, tenemos que

$$X \sum_{i=1}^{n} \mathrm{Q}(X = x_i) = X, \tag{4.36}$$

y si ahora distribuimos X al interior de la suma y usamos (4.28), obtenemos

$$X \sum_{i=1}^{n} \mathrm{Q}(X = x_i) = \sum_{i=1}^{n} \Big(\mathrm{Q}(X = x_i)X\Big) = \sum_{i=1}^{n} \Big(\mathrm{Q}(X = x_i)x_i\Big), \tag{4.37}$$

con lo que finalmente llegamos a la sugerente propiedad

$$\sum_{i=1}^{n} \mathrm{Q}(X = x_i)x_i = X \tag{4.38}$$

que nos dice que cualquier variable discreta puede escribirse como la suma de sus valores posibles, ponderados por la función indicador. Es inmediato ver que también se cumple, para cualquier función $f(X)$,

$$\sum_{i=1}^{n} \mathrm{Q}(X = x_i)f(x_i) = f(X). \tag{4.39}$$

La analogía entre estas propiedades y las de la delta de Dirac se muestran en la **Tabla 4.2**. Para ver en mayor profundidad la conexión entre la función indicador y la delta de Dirac, veamos cómo pasar de una variable discreta a una continua a través de un proceso de límite. Sea $X \in \{x_1, \ldots, x_n\}$ tal que $x_1 = a$ y $x_n = b$, y los valores intermedios son equiespaciados. El conjunto de valores[(1)] puede ser escrito como

$$x_i := a + \frac{i-1}{n-1}(b - a) = a + (i - 1)\Delta_n \tag{4.40}$$

con

$$\Delta_n := \frac{b - a}{n - 1}, \tag{4.41}$$

[(1)] A este conjunto de valores se le denomina una *partición* del intervalo $[a, b]$.

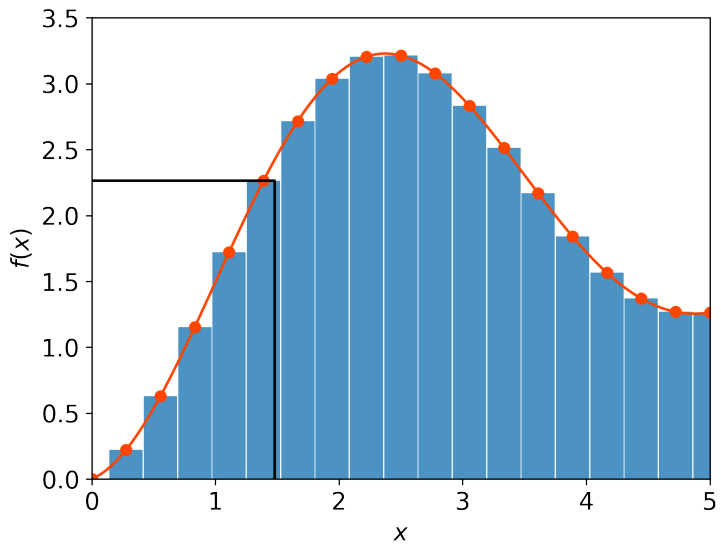


Figura 4.1: Discretización $\tilde{f}_n(x)$ según (4.43) de una función continua $f(x)$. La discretización retorna el mismo valor $f(x_k)$ marcado con círculo para cualquier punto x dentro de la k-ésima barra.

y con estas condiciones podemos escribir la suma de Riemann para cualquier función de los x_i. Inicialmente, supongamos $X = x_k$. Por (4.39) tenemos

$$\sum_{i=1}^{n} \mathrm{Q}(X = x_i) f(x_i) = f(x_k), \tag{4.42}$$

pero ahora queremos extender X para que se mueva continuamente en $[a, b]$. Luego reemplazaremos la proposición «X es igual a x_i» por

> «x_i es el elemento más cercano a X».

Es decir, para una variable continua $X \in [a, b]$ podemos ahora escribir la *discretización* de la función $f(X)$ como

$$\tilde{f}_n(X) := \sum_{i=1}^{n} \mathrm{Q}\Big(|X - x_i| \leq \tfrac{\Delta_n}{2}\Big) f(x_i) = f(x_k) \tag{4.43}$$

donde x_k será el valor más cercano a X en el conjunto $\{x_1, \ldots, x_n\}$, esto es, el de menor $|X - x_i|$. Un ejemplo de esta discretización se muestra en la **Figura 4.1**. Ahora nos interesa tomar el límite $n \to \infty$, que también implica el límite $\Delta \to 0$ por lo que podemos simplemente reemplazar Δ_n por Δ por simplicidad de notación. En este caso, como $|X - x_k| \leq \frac{\Delta}{2}$ y estamos tomando $\Delta \to 0$, se tiene que podemos reemplazar x_k por X y la discretización se vuelve exacta,

$$\lim_{n\to\infty} \tilde{f}_n(X) = f(X). \tag{4.44}$$

Multiplicando y dividiendo por Δ, y tomando los límites anteriores, tenemos que

$$f(X) = \lim_{n\to\infty} \sum_{i=1}^{n} \Delta_n \left[\lim_{\Delta\to 0} \frac{1}{\Delta} \mathrm{Q}\Big(|X - x_i| \leq \tfrac{\Delta}{2}\Big) \right] f(x_i), \tag{4.45}$$

pero usando la definición de la suma de Riemann se tiene

$$f(X) = \int_a^b \mathrm{d}x \left[\lim_{\Delta\to 0} \frac{1}{\Delta} \mathrm{Q}\Big(|X - x| \leq \tfrac{\Delta}{2}\Big) \right] f(x), \tag{4.46}$$

es decir,

$$\delta(X-x) = \lim_{\Delta\to 0} \frac{1}{\Delta} \mathrm{Q}\Big(|X-x| \leq \tfrac{\Delta}{2}\Big), \tag{4.47}$$

en pleno acuerdo con la representación (3.17) en términos de la función rectangular. Vemos que la delta de Dirac no es una función indicador, sino el límite de una.

PROBLEMAS

Problema 4.1. *Escriba la función indicador* $\mathrm{Q}(C)$ *para el punto* $\boldsymbol{x} = (x, y, z)$ *asociada a la proposición*

$C =$ *"El punto* $\boldsymbol{x}$ *está dentro de una esfera de radio* R *y centro* $(0, 0, 0)$*".*

Su resultado sólo debe depender de las variables x, y, z, R*.*

Problema 4.2. *Escriba la tabla de verdad de* $(F \vee G) \wedge \neg(F \wedge G)$*, y represente la función indicador de esta expresión en términos de* $\mathrm{Q}(F)$ *y* $\mathrm{Q}(G)$*.*

Problema 4.3. *Usando funciones indicador, traduzca la afirmación lógica "dos enteros* a *y* b *iguales a un tercero* c *son iguales entre sí" a una propiedad de la función delta de Kronecker.*

Problema 4.4. *Verifique que el coeficiente binomial en (3.116) es solución de la ecuación de recurrencia (4.15).*

Capítulo 5

Estimación

> Before we start to investigate, let us try to realize what we do know, so as to make the most of it, and to separate the essential from the accidental.
>
> Sherlock Holmes, The Adventure of the Priory School

Como describimos previamente en el **Capítulo 1**, es necesario llegar a la idea de inferencia en los casos en que no se tiene acceso a los detalles que permitirían el cálculo o la medición exacta de las cantidades, y es por esto que se necesita una operación distinta a la simple evaluación de expresiones matemáticas. De esta forma es que, en este capítulo, introduciremos la idea de estimación de una cantidad en un estado de conocimiento.

Pensaremos en la estimación como la operación intuitiva con la cual obtenemos una *idea aproximada* del valor de una cantidad dada cierta información. Por ejemplo, podríamos pensar en estimar el número de granos de trigo en un costal, la concentración de una sustancia química en una muestra desconocida, el número de pasajeros en un bus antes de tomarlo, la superficie en metros cuadrados de un departamento, la masa de la Luna, etc.

> **Importante:** En física, podríamos pensar en la llamada *estimación de Fermi*, pero en aquella se trata de obtener una estimación dentro del orden de magnitud, mientras que nosotros buscaremos la mejor estimación con la información que poseemos. En ese sentido, nuestra estimación no es una mera *adivinanza*, sino que debe estar restringida fuertemente por las leyes de la matemática y la lógica, además de por la información conocida acerca de la cantidad.

Por esto mismo, la operación que buscamos deberá ser *objetiva*, en el sentido de que dos personas en el mismo estado de conocimiento I deberán asignar la misma estimación a una cantidad.

Lo que haremos a continuación será *diseñar* nuestra operación estimación a la medida, de tal forma que cumpla cierto número de requerimientos que la harían consistente con nuestras intuiciones. Lograremos esto gracias a la potencia de la lógica y las técnicas exploradas en el **Capítulo 2** y el **Capítulo 3**.

5.1 — Los postulados de la estimación

Podemos comparar la idea de estimación con la de *evaluación de una expresión matemática* sustituyendo los valores de sus incógnitas, por ejemplo, en la expresión

$$\left(X^2 + 3Y\right)_{X=3,Y=2} = 3^2 + 3 \cdot 2 = 15. \tag{5.1}$$

Queremos generalizar esta idea al caso donde no tenemos todos los valores necesarios para producir un número como respuesta. Por ejemplo, ¿qué sucedería si sólo conocemos $X = 3$ y no el valor de Y? Tendríamos algo como

$$\left(X^2 + 3Y\right)_{X=3} = 9 + 3Y, \tag{5.2}$$

cuyo resultado al lado derecho no es un número bien definido sino que es la función

$$y \mapsto 9 + 3y,$$

cuyo valor sigue siendo desconocido hasta que podamos evaluarla en un punto $y = y_0$ particular. Necesariamente la estimación de una cantidad X dependerá únicamente de lo que sabemos respecto a X, y para dejar explícito este hecho introduciremos la siguiente notación que generaliza la evaluación de funciones. Si denotamos la evaluación de una función $f(\boldsymbol{u})$ de variables $\boldsymbol{u}$ en un punto $\boldsymbol{u}_0$ como

$$\left(f(\boldsymbol{u})\right)_{\boldsymbol{u}=\boldsymbol{u}_0} = f(\boldsymbol{u}_0),$$

entonces la **estimación** de la función $f(\boldsymbol{u})$ en el estado de conocimiento I será escrita como

$$\left\langle f \right\rangle_I$$

donde los *brackets* $\langle\rangle$ encierran la cantidad a estimar, y el subíndice I indica la información con la cual se realiza la estimación. El estado de conocimiento I será siempre una conjunción de proposiciones lógicas, esto es,

$$I = A_1 \wedge A_2 \wedge \ldots A_n \tag{5.3}$$

donde los A_i representan premisas. Por simplicidad de notación, escribiremos también $I = A_1, A_2, \ldots, A_n$, donde la coma reemplaza al operador $\wedge$.

Importante: La estimación es una operación que *actúa siempre sobre funciones*, asignando a cada par {función, estado de conocimiento} un valor numérico. Lo que llamamos cantidades desconocidas o aleatorias son en realidad funciones complicadas cuya evaluación requeriría muchos detalles inaccesibles en la práctica.

Ahora postularemos ciertos requerimientos que la operación estimación debe cumplir, tales que todos constituyen propiedades que ya posee la evaluación convencional de expresiones. De esta forma, nos aseguramos que la evaluación convencional sea un caso particular de la estimación. En primer lugar, dado que

$$\Big(f(\boldsymbol{u})+g(\boldsymbol{u})\Big)_{\boldsymbol{u}=\boldsymbol{u}_0} = \Big(f(\boldsymbol{u})\Big)_{\boldsymbol{u}=\boldsymbol{u}_0} + \Big(g(\boldsymbol{u})\Big)_{\boldsymbol{u}=\boldsymbol{u}_0}, \tag{5.4}$$

requeriremos que la estimación de una suma sea la suma de las estimaciones.

Postulado 5.1 — Aditividad de la estimación

$$\big\langle X+Y\big\rangle_I = \big\langle X\big\rangle_I + \big\langle Y\big\rangle_I \ \text{ para todo } I. \tag{5.5}$$

Es decir, si separamos una cantidad Z en dos contribuciones, $Z = X + Y$, entonces el que Z sea estimada como un todo, o que sean estimados X e Y por separado y luego sumados al final no debe hacer diferencia en el resultado, siempre que la misma información I sea usada en todos los casos.
A continuación, notando que si a y b son constantes, la evaluación cumple

$$\Big(a\,f(\boldsymbol{u})+b\Big)_{\boldsymbol{u}=\boldsymbol{u}_0} = a\Big(f(\boldsymbol{u})\Big)_{\boldsymbol{u}=\boldsymbol{u}_0} + b = a\,f(\boldsymbol{u}_0)+b, \tag{5.6}$$

requeriremos que la estimación cumpla el análogo a esta propiedad ante transformaciones lineales constantes.

Postulado 5.2 — Transformación lineal constante

$$\text{Si } \ a,b \text{ son constantes bajo } I, \text{ entonces } \big\langle aX+b\big\rangle_I = a\big\langle X\big\rangle_I + b. \tag{5.7}$$

Notemos que este postulado de inmediato nos dice dos cosas. Primero, si $a = 0$ entonces

$$\big\langle b\big\rangle_I = b,$$

es decir, la estimación de una función constante es la propia constante. Por otro lado, si $b = 0$, entonces se tiene que

$$\big\langle aX\big\rangle_I = a\big\langle X\big\rangle_I,$$

es decir, una constante puede ser «extraída» fuera de una estimación.

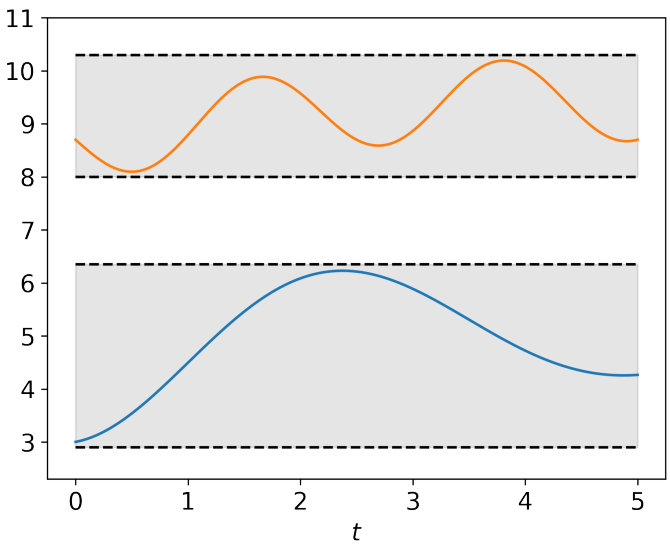


Figura 5.1: Dos cantidades variables tal que una (en naranjo) siempre es mayor que la otra (en azul). El postulado de conservación del orden asegura que la estimación de la primera siempre será mayor que la estimación de la segunda.

Una notación más explícita, pero por lo mismo más difícil de leer a primera vista, permite indicar en el estado de conocimiento que las cantidades A y B son conocidas y tienen valores a y b respectivamente. Con ella el postulado de **Transformación lineal constante** queda como

$$\langle AX + B \rangle_{A=a,B=b,I} = a\langle X \rangle_{A=a,B=b,I} + b, \tag{5.8}$$

y como caso particular muy útil podemos escribir

$$\langle AX \rangle_{A=a,I} = a\langle X \rangle_{A=a,I}. \tag{5.9}$$

Nuestro siguiente postulado asegura que si dos cantidades pueden ser claramente ordenadas de mayor a menor, aunque no se conozcan sus valores, entonces la estimación de dichas cantidades también puede ser ordenada, y debe respetar el orden original.

Postulado 5.3 — Conservación del orden

$$\text{Si } \; X > Y \text{ bajo } I, \text{ entonces } \langle X \rangle_I > \langle Y \rangle_I. \tag{5.10}$$

Esto claramente lo cumple la evaluación convencional de expresiones, ya que si $f(\boldsymbol{u}) > g(\boldsymbol{u})$ para todo $\boldsymbol{u}$, podemos decir que $f(\boldsymbol{u}) - g(\boldsymbol{u}) > 0$ y

$$\Big(f(\boldsymbol{u}) - g(\boldsymbol{u})\Big)_{\boldsymbol{u}=\boldsymbol{u}_0} = f(\boldsymbol{u}_0) - g(\boldsymbol{u}_0) > 0 \tag{5.11}$$

y por lo tanto $f(\boldsymbol{u}_0) > g(\boldsymbol{u}_0)$.

Una descripción gráfica del requerimiento en el Postulado de **Conservación del orden** puede verse en la **Figura 5.1**. Finalmente, nuestro último postulado tiene relación con la evaluación parcial de funciones, como se mostró

en (5.2). Consideremos una función de dos variables, $f(x,y)$, la cual evaluaremos convencionalmente en el punto (x_0, y_0),

$$\Big(f(x,y)\Big)_{x=x_0,y=y_0} = f(x_0,y_0). \tag{5.12}$$

Alternativamente, podemos decidir evaluar parcialmente sólo en $x = x_0$,

$$\Big(f(x,y)\Big)_{x=x_0} = f(x_0,y) \tag{5.13}$$

con lo que, como vimos anteriormente, el resultado es la función $y \mapsto f(x_0,y)$ y no un valor. Si evaluamos esta función en $y = y_0$,

$$\Big(f(x_0,y)\Big)_{y=y_0} = f(x_0,y_0) \tag{5.14}$$

claramente obtenemos el valor de la función f evaluada completamente en (x_0, y_0), lo cual podemos escribir de forma más explícita como la evaluación de una evaluación,

$$\Big(\Big[f(x,y)\Big]_{x=x_0}\Big)_{y=y_0} = f(x_0,y_0). \tag{5.15}$$

Combinando con (5.12) vemos que la evaluación en $x = x_0$ primero, seguida de la evaluación en $y = y_0$, es igual a la evaluación simultánea en $x = x_0, y = y_0$,

$$\Big(\Big[f(x,y)\Big]_{x=x_0}\Big)_{y=y_0} = \Big(f(x,y)\Big)_{x=x_0,y=y_0}. \tag{5.16}$$

Entonces, como una extensión de la propiedad (5.16) de la evaluación convencional, introducimos el último postulado que define la estimación.

Postulado 5.4 — Doble estimación

$$\Big\langle \langle Y \rangle_{X=\bullet,I} \Big\rangle_I = \langle Y \rangle_I \text{ para todo } I. \tag{5.17}$$

Usando una notación más explícita, este postulado significa lo siguiente. Si realizamos una estimación parcial de Y manteniendo fijo el valor de X, llamémoslo x, obtendremos la función

$$Y_p(x) := \langle Y \rangle_{X=x,I} \tag{5.18}$$

que nos lleva de x a la estimación de Y sabiendo que $X = x$ y el conocimiento I. Como Y_p es una función (de un argumento), podemos estimarla usando el conocimiento I, y debemos obtener

$$\langle Y_p \rangle_I = \langle Y \rangle_I. \tag{5.19}$$

Estos cuatro postulados, 5.1 a 5.4, son suficientes para determinar la forma que tiene la operación estimación[(1)] , como veremos en las siguientes secciones.

[(1)] Y por tanto constituyen lo que se podría denominar una *axiomatización* del concepto de estimación.

5.2 — Estimación de variables discretas

Ahora haremos uso de los tres primeros postulados, **Aditividad de la estimación**, **Transformación lineal constante** y **Conservación del orden**, para determinar la forma que toma la estimación de una cantidad discreta. Supongamos para esto una variable X tal que toma uno de n valores, $X \in \{x_1, x_2, \ldots, x_n\}$. Considerando que el estado de conocimiento I contiene los valores conocidos de $x_1, \ldots, x_n$ y usando la propiedad (4.38), podemos escribir

$$\begin{aligned} \langle X \rangle_I &= \left\langle \sum_{i=1}^{n} \mathrm{Q}(X = x_i) x_i \right\rangle_I \\ &= \sum_{i=1}^{n} \langle \mathrm{Q}(X = x_i) \rangle_I \, x_i \\ &= \sum_{i=1}^{n} p_i x_i, \end{aligned} \tag{5.20}$$

donde hemos definido n coeficientes $p_1, p_2, \ldots, p_n$ de acuerdo a

$$p_i := \langle \mathrm{Q}(X = x_i) \rangle_I \,. \tag{5.21}$$

Dado que $0 \leq \mathrm{Q}(X = x_i) \leq 1$, usamos el postulado de **Conservación del orden** para los límites inferior y superior de p_i,

$$\langle \mathrm{Q}(X = x_i) \rangle_I \geq 0, \tag{5.22a}$$

$$\langle \mathrm{Q}(X = x_i) \rangle_I \leq 1, \tag{5.22b}$$

por lo tanto

$$0 \leq p_i \leq 1, \qquad i = 1, 2, \ldots, n. \tag{5.23}$$

Además, usando (4.35), vemos que los coeficientes p_i suman 1, ya que

$$\begin{aligned} \sum_{i=1}^{n} p_i &= \sum_{i=1}^{n} \langle \mathrm{Q}(X = x_i) \rangle_I \\ &= \left\langle \sum_{i=1}^{n} \mathrm{Q}(X = x_i) \right\rangle_I \\ &= \langle 1 \rangle_I = 1, \end{aligned} \tag{5.24}$$

y vemos que, por medio de (4.39), la estimación de una función arbitraria $f(X)$ de la variable X es

$$\begin{aligned} \langle f \rangle_I &= \left\langle \sum_{i=1}^{n} \mathrm{Q}(X = x_i) f(x_i) \right\rangle_I \\ &= \sum_{i=1}^{n} \langle \mathrm{Q}(X = x_i) \rangle_I \, f(x_i) \\ &= \sum_{i=1}^{n} p_i f(x_i), \end{aligned} \tag{5.25}$$

por lo que teniendo el conjunto $p_1, p_2, \ldots, p_n$ podemos calcular $\langle f \rangle_I$ para cualquier $f(X)$. Podemos resumir estos resultados en

$$p_i := \langle \mathrm{Q}(X = x_i) \rangle_I, \quad i = 1, 2, \ldots, n, \tag{5.26a}$$

$$0 \leq p_i \leq 1, \quad i = 1, 2, \ldots, n, \tag{5.26b}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{5.26c}$$

$$\langle f \rangle_I = \sum_{i=1}^{n} p_i f(x_i), \tag{5.26d}$$

con lo que los coeficientes p_i tienen la apariencia de ser *probabilidades* de variables discretas, como se usan en la teoría de la probabilidad y estadística tradicional. Sin embargo, aunque sigan la misma matemática, para nosotros tendrán una semántica algo distinta, dado que son estimaciones de los valores de verdad de las proposiciones $X = x_i$ dado un cierto conocimiento I. Veremos las implicaciones de esta nueva semántica en toda su generalidad en el **Capítulo 6**, pero por ahora nos bastará introducir la notación

$$P(X = x_i | I)$$

para referirnos al coeficiente p_i, notación que leeremos como **la probabilidad de que X sea igual a x_i en el estado de conocimiento I**.

Recuadro 5.1 — Notación para las probabilidades
Escribiremos $P(A|B)$ para denotar la probabilidad de que A sea cierto, dado que suponemos B como cierto. Llamaremos a $|$ la *barra condicional*, notando que siempre lo que se encuentra a la izquierda de dicha barra es algo que no conocemos, mientras que lo que se encuentra a la derecha de la barra es siempre algo considerado como conocido, para efectos del cálculo de la probabilidad.

Con esta nueva notación podemos escribir la estimación de una función $f(X)$ de una variable discreta $X \in \{x_1, x_2, \ldots, x_n\}$ como

$$\langle f \rangle_I = \sum_{i=1}^{n} P(X = x_i | I) f(x_i), \tag{5.27}$$

o incluso como

$$\langle f \rangle_I = \sum_{i=1}^{n} P(x_i | I) f(x_i), \tag{5.28}$$

donde seguiremos la convención de reemplazar la proposición $X = x_i$ por el valor x_i cuando la variable X se entienda en el contexto. Notemos además que $P(\bullet|I)$, al ser una estimación, siempre lleva un estado de conocimiento a la derecha de la barra, en este caso I.

5.3 — Estimación de variables continuas

De manera análoga a (5.20), usaremos para una variable continua $X \in [a, b]$ la identidad

$$X = \int_a^b \mathrm{d}x' \delta(X - x') x',$$

para poder desarrollar la estimación de X como

$$\begin{aligned} \langle X \rangle_I &= \left\langle \int_a^b \mathrm{d}x' \delta(X - x') x' \right\rangle_I \\ &= \int_a^b \mathrm{d}x' \langle \delta(X - x') \rangle_I \, x' \\ &= \int_a^b \mathrm{d}x' p(x') \, x', \end{aligned} \tag{5.29}$$

donde ahora hemos definido la función continua $p(x)$ según

$$p(x) := \langle \delta(X - x) \rangle_I. \tag{5.30}$$

La función $p(x)$ cumple el mismo rol de la probabilidad p_i para variables discretas, aunque en sí misma no es una probabilidad, ya que puede ser mayor que 1. De hecho, como

$$\delta(X - x) \geq 0,$$

se sigue del postulado de **Conservación del orden** que

$$p(x) \geq 0, \qquad x \in [a, b] \tag{5.31}$$

sin embargo $p(x)$ no tiene un límite superior, debido a que $\delta(0) \to +\infty$. A pesar de esto, se cumple el análogo continuo de (5.24),

$$\begin{aligned} \int_a^b \mathrm{d}x' p(x') &= \int_a^b \mathrm{d}x' \langle \delta(X - x') \rangle_I \\ &= \left\langle \int_a^b \mathrm{d}x' \delta(X - x') \right\rangle_I \\ &= \langle 1 \rangle_I = 1. \end{aligned} \tag{5.32}$$

Notemos que según (4.24), si integramos la delta de Dirac en un intervalo $[r, s]$ el resultado es la función indicador de pertenencia al intervalo,

$$\int_r^s \mathrm{d}x' \delta(X - x') = \int_{-\infty}^{\infty} \mathrm{d}x' \delta(X - x') \mathrm{Q}(x' \in [r, s]) = \mathrm{Q}(X \in [r, s]), \tag{5.33}$$

por lo que tenemos

$$\begin{aligned} \int_r^s \mathrm{d}x' p(x') &= \int_r^s \mathrm{d}x' \langle \delta(X - x') \rangle_I \\ &= \left\langle \int_r^s \mathrm{d}x' \delta(X - x') \right\rangle_I \\ &= \langle \mathrm{Q}(X \in [r, s]) \rangle_I \\ &= P(X \in [r, s] | I). \end{aligned} \tag{5.34}$$

En otras palabras, aunque $p(x)$ no es una probabilidad, la integral de $p(x)$ en un intervalo sí lo es, y corresponde precisamente a la probabilidad de que el valor de la variable X se encuentre en ese intervalo. De esta forma, llamaremos a $p(x)$ la **densidad de probabilidad** de X. En concordancia con el caso discreto, cambiaremos la notación $p(x)$ por

$$P(X = x|I)$$

o incluso $P(x|I)$ cuando la variable X se entienda del contexto. Usando la representación (4.47) de la delta de Dirac, podemos escribir la densidad de probabilidad como

$$\begin{aligned} P(X = x|I) = \langle \delta(X - x) \rangle_I &= \left\langle \lim_{\Delta \to 0} \frac{1}{\Delta} \mathrm{Q}(|X - x| \leq \tfrac{\Delta}{2}) \right\rangle_I \\ &= \lim_{\Delta \to 0} \frac{1}{\Delta} \left\langle \mathrm{Q}(|X - x| \leq \tfrac{\Delta}{2}) \right\rangle_I \\ &= \lim_{\Delta \to 0} \frac{1}{\Delta} P(|X - x| \leq \tfrac{\Delta}{2}|I). \end{aligned} \tag{5.35}$$

En un espacio de más dimensiones, llamando $\boldsymbol{X} = (X_1, \ldots, X_n)$ y usando (3.25) tenemos que la generalización de (5.34) es

$$\begin{aligned} \int_{\mathcal{V}} \mathrm{d}\boldsymbol{x}' p(\boldsymbol{x}') &= \int_{\mathcal{V}} \mathrm{d}\boldsymbol{x}' \left\langle \delta(\boldsymbol{X} - \boldsymbol{x}') \right\rangle_I \\ &= \left\langle \int_{\mathcal{V}} \mathrm{d}\boldsymbol{x}' \delta(\boldsymbol{X} - \boldsymbol{x}') \right\rangle_I \\ &= \left\langle \mathrm{Q}(\boldsymbol{X} \in \mathcal{V}) \right\rangle_I \\ &= P(\boldsymbol{X} \in \mathcal{V}|I), \end{aligned} \tag{5.36}$$

esto es, la integral de la densidad de probabilidad sobre una región es la probabilidad de estar en dicha región. Suponiendo ahora que la región $\mathcal{V}$ es una vecindad $\mathcal{V}(\boldsymbol{x})$ de volumen pequeño en torno a un punto $\boldsymbol{x}$, podemos decir

$$P(\boldsymbol{X} \in \mathcal{V}(\boldsymbol{x})|I) = \int_{\mathcal{V}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}' p(\boldsymbol{x}') = p_0 V \tag{5.37}$$

donde

$$p_0 := \frac{1}{V} \int_{\mathcal{V}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}' p(\boldsymbol{x}'), \tag{5.38a}$$

$$V := \int_{\mathcal{V}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}' \tag{5.38b}$$

con V el volumen de la vecindad $\mathcal{V}$. En el límite $V \to 0$ se tiene $p_0 \to p(\boldsymbol{x})$ y entonces

$$p(\boldsymbol{x}) = \lim_{V \to 0} \frac{1}{V} P(\boldsymbol{X} \in \mathcal{V}(\boldsymbol{x})|I), \tag{5.39}$$

que es precisamente la estimación de (4.25).

Con todos estos elementos podemos definir más formalmente la densidad de probabilidad para una variable o múltiples variables.

Definición 5.1 — Densidad de probabilidad
La densidad de probabilidad para la variable continua X en el punto x en el estado de conocimiento I es

$$P(X = x|I) := \lim_{\Delta \to 0} \frac{1}{\Delta} P(|X - x| \leq \tfrac{\Delta}{2}|I). \tag{5.40}$$

Para un conjunto de variables $\boldsymbol{X}$ definiremos la densidad de probabilidad en el punto $\boldsymbol{x}$ en el estado de conocimiento I como

$$P(\boldsymbol{X} = \boldsymbol{x}|I) := \lim_{V \to 0} \frac{1}{V} P(\boldsymbol{X} \in \mathcal{V}(\boldsymbol{x})|I), \tag{5.41}$$

donde $\mathcal{V}(\boldsymbol{x})$ es una vecindad en torno a $\boldsymbol{x}$ y V es el volumen de dicha vecindad.

Podemos ver aquí que la densidad de probabilidad en variables continuas es a la probabilidad en variables discretas como la delta de Dirac es a la función indicador.

5.4 — Invarianza de la estimación

La estimación de una función $f(\boldsymbol{u})$ no depende de la parametrización utilizada para representar dicha función. Por ejemplo, una función en dos dimensiones en coordenadas cartesianas $f(x, y)$ tiene la misma estimación en un estado I que la función transformada a coordenadas polares,

$$\tilde{f}(r, \phi) := f(r \cos \phi, r \sin \phi). \tag{5.42}$$

Por un lado, podemos pensar en (x, y) como las variables aleatorias fundamentales y en f como una función de ellas, o alternativamente en (r, ϕ) como las variables aleatorias fundamentales siendo $\tilde{f}$ una función de ellas. Más aún, podemos olvidarnos de que f es una función y tratarla como una variable aleatoria sin «estructura interna», por lo que podemos referirnos a la estimación de f en el estado I simplemente como $\langle f \rangle_I$ en lugar de usar la notación $\langle f(x, y) \rangle_I$ o $\langle \tilde{f}(r, \phi) \rangle_I$.

Este hecho, de suma importancia, asegura algo que damos por sentado en Física, y es que la elección del sistema de coordenadas es completamente arbitraria y por tanto no puede afectar nuestros resultados. La invarianza de la estimación no es un postulado adicional, sino que su validez la asegura el siguiente teorema.

Teorema 5.1 — Invarianza de la estimación
Sea una función arbitraria $f(\boldsymbol{u})$ y un cambio de coordenadas

$$\boldsymbol{v} \mapsto \boldsymbol{U}(\boldsymbol{v}),$$

de forma que podemos definir una función alternativa $\tilde{f}(\boldsymbol{v})$ de las nuevas variables según

$$\tilde{f}(\boldsymbol{v}) := f(\boldsymbol{U}(\boldsymbol{v})).$$

Entonces, se cumple que

$$\left\langle \tilde{f} \right\rangle_I = \left\langle f \right\rangle_I \tag{5.43}$$

para todo estado de conocimiento I.

La demostración es la siguiente.

Demostración. Llamemos por simplicidad

$$p_u(\boldsymbol{u}) := P(\boldsymbol{U} = \boldsymbol{u}|I), \tag{5.44a}$$
$$p_v(\boldsymbol{v}) := P(\boldsymbol{V} = \boldsymbol{v}|I), \tag{5.44b}$$

a las densidades de probabilidad asociadas a los puntos en las coordenadas $\boldsymbol{u}$ y $\boldsymbol{v}$, respectivamente, con lo que se cumple que

$$p_v(\boldsymbol{v}) = \left\langle \delta(\boldsymbol{V} - \boldsymbol{v}) \right\rangle_I. \tag{5.45}$$

Si ahora usamos el hecho que $\boldsymbol{V}(\boldsymbol{u}) = \boldsymbol{v}$ tiene como única solución $\boldsymbol{u} = \boldsymbol{U}(\boldsymbol{v})$ podemos usar (3.82) para reescribir

$$\begin{aligned} p_v(\boldsymbol{v}) &= \left\langle \frac{\delta(\boldsymbol{U} - \boldsymbol{U}(\boldsymbol{v}))}{|\mathbb{J}_{vu}|} \right\rangle_I \\ &= \left(\frac{p_u}{\mathcal{J}_{vu}} \right)_{\boldsymbol{u}=\boldsymbol{U}(\boldsymbol{v})} \\ &= \tilde{p}_u(\boldsymbol{v}) \left| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}} \right|, \end{aligned} \tag{5.46}$$

donde $\tilde{p}_u(\boldsymbol{v}) = p_u(\boldsymbol{U}(\boldsymbol{v}))$ es la función p_u pero ahora escrita en términos de las variables $\boldsymbol{v}$. Ahora desarrollamos $\left\langle \tilde{f} \right\rangle_I$ en términos de las integrales asociadas a las estimaciones,

$$\begin{aligned} \left\langle \tilde{f} \right\rangle_I &= \int d\boldsymbol{v} p_v(\boldsymbol{v}) \tilde{f}(\boldsymbol{v}) \\ \text{usando (5.46)} \quad &= \int d\boldsymbol{v} \left| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}} \right| \tilde{p}_u(\boldsymbol{v}) \tilde{f}(\boldsymbol{v}) \\ \text{usando (3.75)} \quad &= \int d\boldsymbol{u} \cancel{\left| \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{u}} \right|} \cancel{\left| \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{v}} \right|} p_u(\boldsymbol{u}) f(\boldsymbol{u}) \\ &= \int d\boldsymbol{u} p_u(\boldsymbol{u}) f(\boldsymbol{u}) = \left\langle f \right\rangle_I \end{aligned} \tag{5.47}$$

En el siguiente capítulo veremos las consecuencias del concepto de probabilidad, entendida como una estimación, y las reglas que permiten operar con probabilidades.

PROBLEMAS

Problema 5.1. *Demuestre usando sólo el postulado de* ***Aditividad de la estimación*** *que*

(a) $\langle 0 \rangle_I = 0$ *para todo* I*,*

(b) $\langle nX \rangle_I = n\langle X \rangle_I$ *para* n *entero y todo* I*.*

Problema 5.2. *Calcule la estimación de la función* $f(x,y,z) = (xy)^2$ *bajo la densidad de probabilidad*

$$P(x,y,z|\lambda) = \frac{1}{Z(\lambda)} \exp\big(-\lambda(x^2+y^2+z^2)\big) \tag{5.48}$$

y determine explícitamente la densidad de probabilidad para las coordenadas polares esféricas, $P(r,\phi,\theta|\lambda)$*. Verifique que la estimación de* $\tilde{f}(r,\phi,\theta) = r^4(\sin\theta)^4(\sin\phi\cos\phi)^2$ *es la misma que la de* f*.*

CAPÍTULO 6

Probabilidad

> Each fact is suggestive in itself. Together they have a cumulative force.
>
> Sherlock Holmes, The Bruce-Partington Plans

En el **Capítulo 5**, en la búsqueda de la manera correcta de calcular estimaciones de variables discretas, hemos descubierto el concepto de *probabilidad* de una afirmación en un cierto estado de conocimiento. Esta probabilidad fue definida sólo para cierto tipo de proposiciones, pero ahora la definiremos de forma más general.

Definición 6.1 — Probabilidad
Para una proposición lógica A cualquiera y un estado de conocimiento I definiremos

$$P(A|I) := \langle \mathrm{Q}(A) \rangle_I \tag{6.1}$$

como la **probabilidad** de A dado I.

Esta definición implica que la probabilidad $P(A|I)$ es un número real en el intervalo $[0, 1]$ que representa la *estimación del valor de verdad de A bajo la información I*. En otras palabras, la probabilidad de A dado I nos indica cuán plausible debe considerarse la proposición A a la luz de lo conocido en I. En particular, si

$$I \Rightarrow (\neg A)$$

se tiene que $P(A|I) = 0$, mientras que si

$$I \Rightarrow A$$

se cumple que $P(A|I) = 1$. Esta es exactamente la descripción *bayesiana* de la probabilidad, donde las probabilidades son grados de plausibilidad (o grados de creencia plausible) que nosotros asignamos a las distintas afirmaciones sobre la realidad, en base a nuestro conocimiento.

6.1 — Probabilidad bayesiana y frecuentista

Aunque fue la interpretación original de la probabilidad desde los tiempos de Jacob Bernoulli (1713), Thomas Bayes (1763) y Pierre-Simon de Laplace (1820), la probabilidad bayesiana perdió su posición dominante ante la probabilidad frecuentista, impulsada en gran medida por Karl Pearson (1902) y Ronald A. Fisher (1956). Esta visión frecuentista sólo trata con *fenómenos aleatorios repetibles*, y define la probabilidad de dichos eventos como una razón o frecuencia,

$$p(A) := \frac{\text{número de casos donde } A \text{ ocurre}}{\text{número de casos totales}}, \tag{6.2}$$

en el límite de un número muy grande de casos totales.

Hoy en día, luego del resurgimiento de la teoría bayesiana en gran medida gracias a las nuevas capacidades de cálculo que no estaban disponibles al principio del siglo XX, entendemos que todos los resultados de la probabilidad frecuentista están contenidos en la probabilidad bayesiana como casos particulares, como fue mostrado extensivamente por Jeffreys (1998) y Jaynes (2003). No sólo eso, sino que hay muchos problemas intratables —o incluso imposibles de formular— desde el punto de vista frecuentista que pueden ser tratados de forma bayesiana.

Por ejemplo,

(1) ¿Cuál es la probabilidad de que en este momento exista vida en el planeta Marte?

(2) ¿Cuál es la probabilidad de que el dígito decimal[(1)] de π en la posición 999 997 sea 3?

(3) ¿Cuál es la probabilidad de que la teoría de la relatividad de Einstein sea correcta?

En ninguno de estos casos podemos contar un número de casos donde la pregunta sea cierta y usar (6.2), ya que ¿qué serían estos casos? ¿distintas historias alternativas del universo donde todo se mantiene igual excepto aquello que se pregunta?

6.2 — La regla de la suma y el producto

Para verificar que nuestras probabilidades $P(A|I)$ como las definimos en (6.1) son efectivamente equivalentes a las probabilidades de la interpretación bayesiana, debemos ser capaces de demostrar que la estimación de una función indicador cumple con las propiedades usuales de una probabilidad. Estas propiedades se resumen en las llamadas regla de la suma y regla del producto, que se muestran en el recuadro siguiente.

(1) Si quiere eliminar su incerteza respecto a esta pregunta, mire al final de la página https://www.piday.org/million/

Recuadro 6.1 — Álgebra de probabilidades
La regla de la suma,

$$P(A|I) + P(\neg A|I) = 1, \tag{6.3}$$

y la regla del producto,

$$P(A \wedge B|I) = P(A|I)P(B|A \wedge I), \tag{6.4}$$

son suficientes para definir el *álgebra de probabilidades*. De ellas es posible derivar la regla extendida de la suma,

$$P(A \vee B|I) = P(A|I) + P(B|I) - P(A \wedge B|I). \tag{6.5}$$

Nuestra tarea es entonces demostrar (6.3) y (6.4) a partir de los postulados de la estimación. En primer lugar veamos cómo eliminar la negación en $P(\neg A|I)$, es decir, cómo expresar $P(\neg A|I)$ en términos de $P(A|I)$. Simplemente tomamos la identidad (4.2) para la función indicador de la negación,

$$\mathrm{Q}(\neg A) = 1 - \mathrm{Q}(A),$$

y aplicamos estimación a ambos lados en el estado de conocimiento I, obteniendo

$$\big\langle \mathrm{Q}(\neg A) \big\rangle_I = \big\langle 1 - \mathrm{Q}(A) \big\rangle_I = 1 - \big\langle \mathrm{Q}(A) \big\rangle_I, \tag{6.6}$$

es decir,

$$\boxed{P(\neg A|I) = 1 - P(A|I),} \tag{6.7}$$

que es la regla de la suma. Esto es, mientras más probable es A, menos probable es su negación $\neg A$, y viceversa.

Para ver cómo eliminar la disyunción en $P(A \vee B|I)$, también podemos hacer uso directamente de la propiedades de la función indicador, en este caso, de la identidad (4.5),

$$\mathrm{Q}(A \vee B) = \mathrm{Q}(A) + \mathrm{Q}(B) - \mathrm{Q}(A \wedge B).$$

Aplicando estimación a ambos lados en el estado I, tenemos

$$\big\langle \mathrm{Q}(A \vee B) \big\rangle_I = \big\langle \mathrm{Q}(A) \big\rangle_I + \big\langle \mathrm{Q}(B) \big\rangle_I - \big\langle \mathrm{Q}(A \wedge B) \big\rangle, \tag{6.8}$$

esto es,

$$\boxed{P(A \vee B|I) = P(A|I) + P(B|I) - P(A \wedge B|I).} \tag{6.9}$$

que es la *regla extendida de la suma*.

El determinar cómo eliminar la conjunción en $P(A \wedge B|I)$ es más complicado, y de hecho requiere cambiar de estado de conocimiento, cosa que sólo es posible a través del postulado de **Doble estimación**.
Primero, definamos por simplicidad de notación $a := \mathrm{Q}(A)$, $b := \mathrm{Q}(B)$ y usemos el postulado de **Doble estimación** con $X = a$, $Y = ab$,

$$\langle ab \rangle_I = \left\langle \langle ab \rangle_{a=\bullet,I} \right\rangle_I . \tag{6.10}$$

Ahora concentrémonos en la cantidad $\langle ab \rangle_{a=\bullet,I}$ dentro de la estimación del lado derecho, que de hecho es la función

$$\alpha \mapsto \langle ab \rangle_{a=\alpha,I} \qquad \text{con } \alpha \in \{0,1\}.$$

Podemos sacar la constante a con el valor α y sustituir la función por

$$\alpha \mapsto \alpha \langle b \rangle_{a=\alpha,I},$$

pero por (4.26) sabemos que ésta es equivalente a

$$\alpha \mapsto \alpha \langle b \rangle_{a=1,I}.$$

Por lo tanto, definiendo $\beta := \langle b \rangle_{a=1,I}$ tenemos que (6.10) se reduce a

$$\langle ab \rangle_I = \langle a\beta \rangle_I = \langle a \rangle_I \beta = \langle a \rangle_I \langle b \rangle_{a=1,I}. \tag{6.11}$$

Esto en la notación original nos dice que

$$\langle \mathrm{Q}(A)\mathrm{Q}(B) \rangle_I = \langle \mathrm{Q}(A) \rangle_I \langle \mathrm{Q}(B) \rangle_{A,I}, \tag{6.12}$$

donde hemos cambiado el estado de conocimiento $(a = 1, I)$ por el equivalente (A, I). Finalmente, usando (4.3) tenemos

$$\langle \mathrm{Q}(A \wedge B) \rangle_I = \langle \mathrm{Q}(A)\mathrm{Q}(B) \rangle_I = \langle \mathrm{Q}(A) \rangle_I \langle \mathrm{Q}(B) \rangle_{A,I}, \tag{6.13}$$

y por tanto hemos demostrado la regla del producto,

$$\boxed{P(A \wedge B|I) = P(A|I)P(B|A \wedge I).} \tag{6.14}$$

Las propiedades (6.3) y (6.4) son de hecho las que definen la probabilidad en términos **bayesianos**, y de las cuales es posible deducir toda la estructura de la probabilidad, con lo que hemos validado entonces nuestra definición en (6.1). Más aún, ahora que podemos afirmar que los coeficientes

$$p_i = P(X = x_i|I)$$

son legítimamente las probabilidades (bayesianas) asociadas a una variable discreta, y que

$$p(x) = P(X = x|I)$$

es la densidad de probabilidad (bayesiana) de una variable continua, también podemos reconocer nuestra operación estimación como la **expectación** o valor esperado de la teoría de probabilidad pero en el sentido bayesiano. Por lo tanto, nos referiremos de ahora en adelante a la estimación como expectación, y leeremos $\langle X \rangle_I$ como la expectación de X en el estado de conocimiento I.

En las siguientes secciones revisaremos las consecuencias de estas reglas, en particular la **regla de marginalización** y el **teorema de Bayes**, y cómo éstas describen un proceso de razonamiento y aprendizaje que nos permitirá crear modelos y actualizarlos al recibir nueva información, como lo esbozamos en el **Capítulo 1**. Pero primero veamos cómo asignar números concretos a las probabilidades en un caso de nula información.

6.3 — El principio de indiferencia

Supongamos n proposiciones $A_1, \ldots, A_n$ mutuamente excluyentes y exhaustivas. En este caso la identidad (**4.34**) nos dice que

$$\sum_{i=1}^{n} \mathrm{Q}(A_i) = 1 \tag{6.15}$$

y por tanto, tomando expectación en un estado de conocimiento I arbitrario, se tendrá

$$\sum_{i=1}^{n} P(A_i|I) = 1. \tag{6.16}$$

Si nos encontramos en un estado de conocimiento I_0 tal que el hecho que existen estas n alternativas es nuestra única información, ¿qué probabilidades

$$p_i := P(A_i|I_0)$$

deberíamos asignar?

Claramente, si sólo conocemos el número de alternativas, no tenemos preferencia por ninguna de ellas sobre las otras, y por tanto, no se justifica que entre dos probabilidades p_i y p_j con $i \neq j$ una de ellas sea mayor que la otra. Dicho de otra manera, si por ejemplo hemos asignado $p_2 > p_3$, ¿qué nos hizo preferir A_2 por sobre A_3 si todas las proposiciones A_i son equivalentes en I_0? En términos de la idea de simetría, las etiquetas 2 y 3 fueron asignadas de forma arbitraria, por lo que podríamos haber llamado A_3 a la proposición A_2 y viceversa, y obtendríamos ahora $p_3 > p_2$.

Por supuesto, la única alternativa que refleja nuestro honesto desconocimiento de las proposiciones más allá de su número es que todas las probabilidades p_i sean iguales, es decir, debemos asignar

$$P(A_i|I_0) = \frac{1}{n}$$

para que además se cumpla la *condición de normalización* (**6.16**). A esta regla se le denomina el *principio de indiferencia*.

Recuadro 6.2 — El principio de indiferencia
Para un conjunto de n proposiciones mutuamente excluyentes y exhaustivas $A_1, \ldots, A_n$ en un estado de conocimiento I_0 donde únicamente conocemos el número total de proposiciones y no somos capaces de distinguir entre una y otra, las únicas probabilidades consistentes con I_0 son

$$P(A_i|I_0) = \frac{1}{n}, \tag{6.17}$$

de forma que

$$\sum_{i=1}^{n} P(A_i|I_0) = 1.$$

Este principio justifica el asignar probabilidad $1/2$ a obtener «cara» en el lanzamiento de una moneda, o asignar probabilidad $1/6$ a obtener «5» en el lanzamiento de un dado de 6 caras.

6.4 — La regla de marginalización

A continuación introduciremos la **regla de marginalización**, que nos permite eliminar proposiciones o variables irrelevantes en un problema, o agregar nuevas proposiciones o variables relevantes. Consideremos una proposición A y un estado de conocimiento I. Escribiendo la probabilidad $P(A|I)$ como una expectación, podemos agregar una nueva proposición B cualquiera de la siguiente manera,

$$\begin{aligned} P(A|I) = \langle \mathrm{Q}(A) \rangle_I &= \langle \mathrm{Q}(A) \underbrace{\left[\mathrm{Q}(B) + \mathrm{Q}(\neg B) \right]}_{=1} \rangle_I \\ &= \langle \mathrm{Q}(A)\mathrm{Q}(B) \rangle_I + \langle \mathrm{Q}(A)\mathrm{Q}(\neg B) \rangle_I \\ &= P(A \wedge B|I) + P(A \wedge [\neg B]|I), \end{aligned} \tag{6.18}$$

es decir, tenemos

$$\boxed{P(A|I) = P(A \wedge B|I) + P(A \wedge \neg B|I),} \tag{6.19}$$

que es la regla de marginalización para proposiciones.

Otra manera de entender esta regla es que podemos eliminar una proposición (B en este caso) que se encuentre **a la izquierda de la barra condicional** sumando sobre todos sus valores, B y $\neg B$. Si además usamos la regla del producto, podemos escribir

$$P(A|I) = P(A|B \wedge I)P(B|I) + P(A|\neg B \wedge I)P(\neg B|I), \tag{6.20}$$

lo cual nos indica que para eliminar una proposición **a la derecha de la barra condicional** hay que sumar sobre todos sus valores pero multiplicando por las probabilidades respectivas.

Ejemplo 6.4.1. *Supongamos que voy de la casa al trabajo usando una de dos opciones: voy en bus (B) o en metro (M). Como ir en bus significa no ir en metro y viceversa, se tiene $M = (\neg B)$ y la probabilidad de llegar atrasado (A) al trabajo es entonces*

$$P(A|I) = P(A|B,I)P(B|I) + P(A|M,I)P(M|I),$$

es decir, es la probabilidad de llegar atrasado si voy en bus por la probabilidad de ir en bus más la probabilidad de llegar atrasado si voy en metro por la probabilidad de ir en metro.

Una demostración alternativa de (6.19) pasa por usar la regla de la suma para el estado de conocimiento $(A \wedge I)$, esto es,

$$P(B|A,I) + P(\neg B|A,I) = 1 \tag{6.21}$$

y multiplicar por $P(A|I)$ a ambos lados, obteniendo

$$\underbrace{P(A|I)P(B|A,I)}_{=P(A,B|I)} + \underbrace{P(A|I)P(\neg B|A,I)}_{=P(A,\neg B|I)} = P(A|I), \tag{6.22}$$

Es posible llevar a cabo el mismo procedimiento con variables discretas. Supongamos una variable discreta $X \in \{x_1, \ldots, x_n\}$ con probabilidades $P(X = x_i|I)$, y agreguemos una nueva variable $Y \in \{y_1, \ldots, y_m\}$.
Podemos escribir

$$\begin{aligned} P(x_i|I) = \langle \mathrm{Q}(X = x_i) \rangle_I &= \Big\langle \mathrm{Q}(X = x_i) \underbrace{\sum_{j=1}^{m} \mathrm{Q}(Y = y_j)}_{=1} \Big\rangle_I \\ &= \sum_{j=1}^{m} \langle \mathrm{Q}(X = x_i)\mathrm{Q}(Y = y_j) \rangle_I \\ &= \sum_{j=1}^{m} P(x_i, y_j|I), \end{aligned} \tag{6.23}$$

es decir, tenemos en la notación completa,

$$\boxed{P(X = x_i|I) = \sum_{j=1}^{m} P(X = x_i, Y = y_j|I).} \tag{6.24}$$

Nuevamente usando la regla del producto podemos separar el sumando y obtener

$$P(X = x_i|I) = \sum_{j=1}^{m} P(X = x_i|Y = y_j, I)P(Y = y_j|I). \tag{6.25}$$

En el caso de variables continuas el procedimiento es el mismo, primero escribimos

$$\begin{aligned} P(\boldsymbol{x}|I) = \langle \delta(\boldsymbol{X}-\boldsymbol{x})\rangle_I &= \Big\langle \delta(\boldsymbol{X}-\boldsymbol{x}) \underbrace{\int d\boldsymbol{y}\ \delta(\boldsymbol{Y}-\boldsymbol{y})}_{=1} \Big\rangle_I \\ &= \int d\boldsymbol{y} \langle \delta(\boldsymbol{X}-\boldsymbol{x})\delta(\boldsymbol{Y}-\boldsymbol{y})\rangle_I \\ &= \int d\boldsymbol{y} P(\boldsymbol{x},\boldsymbol{y}|I), \end{aligned} \tag{6.26}$$

con lo que llegamos al equivalente continuo de (6.24), dado por

$$P(\boldsymbol{X}=\boldsymbol{x}|I) = \int d\boldsymbol{y} P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y}|I). \tag{6.27}$$

Finalmente usando la regla del producto obtenemos

$$P(\boldsymbol{X}=\boldsymbol{x}|I) = \int d\boldsymbol{y} P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y},I) P(\boldsymbol{Y}=\boldsymbol{y}|I). \tag{6.28}$$

6.5 — El teorema de Bayes

A continuación veremos el que será el teorema más importante dentro de la inferencia, el teorema de Bayes, originalmente introducido por el reverendo Thomas Bayes (1763) pero en realidad popularizado por Laplace (1820). Aunque su derivación es inmediata, sus consecuencias son claves en cómo entenderemos el proceso de aprendizaje racional. La formulación que nos interesa del teorema de Bayes dice relación con el proceso de actualización de modelos descrito en la **Figura 1.3** en el **Capítulo 1**, e involucra una hipótesis o teoría T que se pone a prueba, una evidencia E que necesitamos considerar, y el estado de conocimiento base I_0 que no incluye la evidencia E.

Teorema 6.1 — Teorema de Bayes
Consideremos las siguientes definiciones.

T : Hipótesis o teoría que se está poniendo a prueba.

E : Nueva evidencia a considerar.

$\mathbf{I_0}$: Estado de conocimiento que no incluye la evidencia E.

La probabilidad $P(T|E,I_0)$ de la hipótesis T incorporando la evidencia E es proporcional a la probabilidad $P(T|I_0)$ de T sin incorporar E, y está dada por

$$P(T|E,I_0) = P(T|I_0)\frac{P(E|T,I_0)}{P(E|I_0)}. \tag{6.29}$$

El teorema de Bayes captura elegantemente la idea de **racionalidad**: somos racionales cuando modificamos nuestras creencias al aparecer nueva evidencia que las desafía, de acuerdo a (6.29).

Por el contrario, ante la pregunta «¿qué evidencia sería capaz de hacerle cambiar de opinión?», si la respuesta es *ninguna*, entonces tendría sentido llamar a esa creencia **irracional**.

La demostración del teorema es sencilla, y pasa por recordar el hecho que la conjunción $A \wedge B$ es una operación conmutativa, es decir,

$$A \wedge B = B \wedge A.$$

Esto significa que podemos escribir $P(E, T|I_0) = P(T, E|I_0)$, y por tanto, usando la regla del producto a ambos lados, tenemos

$$P(E|I_0)P(T|E, I_0) = P(T|I_0)P(E|T, I_0). \tag{6.30}$$

Si ahora despejamos $P(T|E, I_0)$ en función de los otros 3 factores, obtenemos el teorema de Bayes en (**6.29**)

$$P(T|E, I_0) = \frac{P(T|I_0)P(E|T, I_0)}{P(E|I_0)}$$

Volviendo a la interpretación en términos de hipótesis y evidencia, de inmediato vemos que la razón

$$R(T; E) := \frac{P(E|T, I_0)}{P(E|I_0)} \tag{6.31}$$

es la que determina si la evidencia E favorece o perjudica a T. Si $R > 1$, entonces $P(T|E, I_0) > P(T|I_0)$ y el incluir la evidencia E aumenta la probabilidad de T, es decir, la evidencia E favorece a T. Por el contrario, si $R < 1$, entonces $P(T|E, I_0) < P(T|I_0)$ y la evidencia E disminuye la probabilidad de T al ser incluida. Diremos que la evidencia E perjudica a T en este caso. La magnitud de R también es crucial, ya que si ésta es del orden de 1, el efecto de la evidencia no será tan dramático como si $R \gg 1$ o si $R \ll 1$. Como $P(E|T, I_0)$ tiene como máximo 1, la única manera en que R puede crecer indefinidamente es que $P(E|I_0) \to 0$, es decir, cuando la evidencia es algo que considerábamos prácticamente imposible. **Las evidencias ridículamente improbables son las que tendrán mayor valor para una hipótesis**.

Ejemplo 6.5.1. *Consideremos el caso en que una persona, pretendiendo que tiene poderes sobrenaturales, hace la predicción de que lloverá mañana. Si es invierno, no sería para nada raro que lloviera por lo que la predicción tiene una importancia mínima. Si fuera verano en un clima seco, entonces la predicción acertada de lluvia adquiere mucho más peso, porque algo que difícilmente hubiéramos esperado (antes de oír la predicción) resulta cierto. Por otro lado, si la predicción es que mañana lloverán langostas, y efectivamente así ocurre, entonces no nos queda otra opción que maravillarnos ante las capacidades de dicha persona.*

De cierta manera podemos relacionar lo informativa que resulta una evidencia para nosotros con lo inesperada o sorpresiva que ésta es bajo la información actual que tenemos. Más adelante, en el **Capítulo 11**, revisaremos el concepto de *sorpresa* y su relación con la información medida en *bits*.

Analicemos ahora el que es posiblemente el caso más habitual: cuando la hipótesis de interés es una entre n alternativas mutuamente excluyentes y exhaustivas $T_1, \ldots, T_n$. En ese caso podemos escribir la probabilidad de la evidencia E dado I_0 usando la regla de marginalización como

$$\begin{aligned} P(E|I_0) &= \sum_{i=1}^{n} P(E, T_i|I_0) \\ &= \sum_{i=1}^{n} P(T_i|I_0)P(E|T_i, I_0), \end{aligned} \tag{6.32}$$

y luego el teorema de Bayes para la k-ésima hipótesis toma la forma

$$P(T_k|E, I_0) = \frac{P(T_k|I_0)P(E|T_k, I_0)}{\sum_{i=1}^{n} P(T_i|I_0)P(E|T_i, I_0)}. \tag{6.33}$$

El caso particular cuando existen sólo dos alternativas, T y $\neg T$, es de especial importancia.

Recuadro 6.3 — Teorema de Bayes para dos hipótesis alternativas

Cuando analizamos la influencia de una evidencia E para escoger entre dos hipótesis alternativas, T y $\neg T$, el teorema de Bayes toma la forma

$$P(T|E, I_0) = \frac{P(T|I_0)P(E|T, I_0)}{P(T|I_0)P(E|T, I_0) + \big(1 - P(T|I_0)\big)P(E|\neg T, I_0)}. \tag{6.34}$$

Esto implica que, además de la probabilidad previa $P(T|I_0)$ sólo necesitamos conocer $P(E|T, I_0)$ y $P(E|\neg T, I_0)$. Esto es, no sólo se necesita la probabilidad de observar la evidencia en caso que T sea cierta, sino también en caso de ser falsa.

Existen dos situaciones interesantes que podemos analizar.

(A) La evidencia E es imposible de acuerdo a T.

En este caso, $P(E|T, I_0) = 0$, por tanto $P(T|E, I_0) = 0$ para cualquier valor de $P(T|I_0)$, esto es, la hipótesis es refutada inmediatamente. En términos simples, si T predice que E es imposible, y luego E es observado, entonces T es automáticamente falsa.

Este es el típico escenario de lo que se denomina criterio de *falsabilidad* de Popper (1959), de acuerdo al cual una teoría científica no puede ser demostrada cierta, sino que sólo puede ser refutada, cuando una de sus predicciones falla al contacto con la realidad. Es más, de acuerdo al teorema de Bayes, existe un único caso en que es posible que una evidencia

demuestre con certeza absoluta una hipótesis T, y es cuando dicha evidencia refuta todas las alternativas. Este es literalmente un ejemplo de la regla de Sherlock Holmes, «cuando se ha eliminado lo imposible, lo que queda debe ser la verdad».

Para ver cómo ocurre esto, consideremos nuestra teoría T_1 junto a otras $(n-1)$ alternativas $T_2, \ldots, T_n$. Escribiendo $P(E|I_0)$ por medio de la regla de marginalización, tenemos

$$P(E|I_0) = P(T_1|I_0)P(E|T_1, I_0) + \sum_{i=2}^{n} P(T_i|I_0)P(E|T_i, I_0). \tag{6.35}$$

Ahora imponemos que T_1 se vuelva cierta bajo la evidencia E, es decir $P(T_1|E, I_0) = 1$, luego

$$\frac{P(T_1|I_0)\ P(E|T_1, I_0)}{P(T_1|I_0)P(E|T_1, I_0) + \sum_{i=2}^{n} P(T_i|I_0)P(E|T_i, I_0)} = 1, \tag{6.36}$$

pero la igualdad de numerador y denominador implica que

$$\sum_{i=2}^{n} \underbrace{P(T_i|I_0)P(E|T_i, I_0)}_{\geq 0} = 0$$

$$\hookrightarrow \underbrace{P(T_i|I_0)}_{>0} P(E|T_i, I_0) = 0$$

luego $P(E|T_i, I_0) = 0$ para $i = 2, \ldots, n$ y por el teorema de Bayes en su forma (6.33) se sigue que

$$P(T_i|E, I_0) = 0 \text{ para } i = 2, \ldots, n.$$

En resumen, una evidencia E sólo puede demostrar definitivamente una hipótesis T cuando dicha evidencia refuta a todas las hipótesis rivales a T. Pero esto sólo puede ocurrir en la deducción lógica, la cual es recuperada aquí como un caso particular de la inferencia.

(B) La evidencia E es muy improbable *a priori*, pero de acuerdo a T es cierta.
En este caso $P(E|I_0) \ll 1$ y $P(E|T, I_0) = 1$, luego se tiene que

$$P(T|E, I_0) \gg P(T|I_0),$$

esto es, la hipótesis T se hace enormemente más plausible al incorporar la evidencia E.

Esto nos alerta de tener en cuenta el cuán improbable es previamente una evidencia con el fin de evaluar su potencial para soportar o refutar cualquier hipótesis. La moraleja aquí es que no sólo basta que $T \Rightarrow E$ para declarar que T es soportada por E, sino que es necesario saber si otras hipótesis también implican E de forma similar. Por ejemplo, si miro por mi ventana a la casa vecina y encuentro que su patio está

mojado, podría concluir que esto es evidencia de que ha llovido —dado que si lloviera, el patio se mojaría— pero esta no es la única hipótesis que explica los hechos. Si los vecinos regaron el patio, también éste se mojaría, por lo tanto el patio mojado es también evidencia a favor de esta última explicación. Lo que manda en este caso es la probabilidad previa de que haya llovido o que hayan regado.

► Entre los autores que proponen una visión bayesiana de la filosofía de la Ciencia, además del propio Jaynes (2003) se encuentran Rosenkrantz (1977) y Howson y Urbach (2006).

Ejemplo 6.5.2. *La reciente enfermedad COVID-19 tiene, para cierto test en una región, una tasa de positividad de 5.1 %. Además, sabemos que la tasa de incidencia en la población mundial es de 1.3 %. La tasa de falsos positivos del test en cuestión es de 4 %. Si me realizo el test y obtengo un resultado positivo, ¿cuál es la probabilidad de que realmente tenga la enfermedad?*

***Solución**. De acuerdo al teorema de Bayes, tenemos*

$$P(C|t, I_0) = \frac{P(C|I_0)P(t|C, I_0)}{P(t|I_0)}. \tag{6.37}$$

Del enunciado tenemos los valores

$$P(C|I_0) = 0.013, \tag{6.38a}$$
$$P(t|I_0) = 0.051, \tag{6.38b}$$
$$P(t|\neg C, I_0) = 0.04, \tag{6.38c}$$

y usando la regla de marginalización, podemos escribir $P(t|I_0)$ como

$$\underbrace{P(t|I_0)}_{=0.051} = \underbrace{P(C|I_0)}_{=0.013} P(t|C, I_0) + \underbrace{P(\neg C|I_0)}_{=1-0.013} \underbrace{P(t|\neg C, I_0)}_{=0.04}. \tag{6.39}$$

De aquí despejamos $P(t|C, I_0) = 0.886$ y entonces

$$P(C|t, I_0) = \frac{P(C|I_0)P(t|C, I_0)}{P(t|I_0)} = \frac{0.013 \times 0.886}{0.051} = 0.226. \tag{6.40}$$

Vemos que un único test positivo en este caso no es por sí mismo evidencia suficiente para inferir que se tiene la enfermedad, principalmente debido a la alta tasa de falsos positivos.

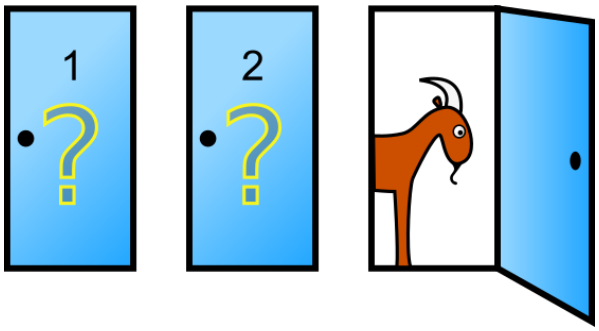


Figura 6.1: En el problema de Monty Hall, un concursante participa por un premio que se encuentra detrás de una de tres puertas. Luego de elegir una de ellas, el animador abre una de las puertas restantes, revelando una cabra. (Imagen de Wikimedia Commons)

6.5.1 El problema de Monty Hall

En el famoso *problema de Monty Hall* (Selvin 1975), un concursante se enfrenta a tres puertas cerradas. Tras una de estas puertas se encuentra un premio (digamos, un automóvil cero kilómetro), mientras que tras las otras dos puertas se encuentran cabras. Una vez que el concursante ha elegido una puerta, la cual se mantiene cerrada, el animador del programa abre una de las dos puertas restantes, mostrando por supuesto que hay una cabra, como se ve en la **Figura 6.1**. Con esto queda claro que el premio se encuentra, ya sea en la puerta que eligió el concursante, o en la otra que permanece cerrada.

En lugar de abrir de inmediato la puerta del concursante, con lo que se acabaría el concurso, el animador le pregunta al concursante si desea cambiar su puerta por la otra que permanece cerrada. La pregunta es, **¿Le conviene al concursante cambiarse de puerta?** ¿Y si es así, por qué?

Solución. Pongamos números 1, 2 y 3 a las puertas, y supongamos que el concursante escoge la puerta 2. Definamos las proposiciones mutuamente excluyentes y exhaustivas G_1, G_2, G_3 como

$$G_n = \text{«La puerta que contiene el premio es la puerta } n\text{»},$$

con n=1,2,3, luego sólo ganaremos si $G_2 = \mathbb{T}$. En nuestro estado inicial de conocimiento I con todas las puertas cerradas, tendremos de acuerdo al principio de indiferencia que

$$P(G_1|I) = P(G_2|I) = P(G_3|I) = \frac{1}{3}. \tag{6.41}$$

Una vez que el animador abre una de las puertas distinta a la 2, digamos la puerta 3, para revelar que ahí no está el premio, tenemos información adicional R_3. Usando el teorema de Bayes, tenemos

$$P(G_i|R_3, I) = \frac{P(G_i|I)P(R_3|G_i, I)}{P(R_3|I)}, \tag{6.42}$$

donde nuestra constante de normalización $P(R_3|I)$ puede ser calculada como

$$\begin{aligned} P(R_3|I) &= \underbrace{P(G_1|I)}_{=1/3} P(R_3|G_1, I) + \underbrace{P(G_2|I)}_{=1/3} P(R_3|G_2, I) + \underbrace{P(G_3|I)}_{=1/3} P(R_3|G_3, I) \\ &= \frac{1}{3}\Big(P(R_3|G_1, I) + P(R_3|G_2, I) + P(R_3|G_3, I)\Big). \end{aligned} \tag{6.43}$$

Hasta aquí la situación parece simétrica a simple vista. Sin embargo, notemos que *es imposible para el animador abrir la puerta que contiene el premio*, por lo tanto se tiene que

$$P(R_3|G_3, I) = 0. \tag{6.44}$$

Por otro lado, si la puerta ganadora es la 1, y el concursante escogió la 2, ninguna de ellas puede ser revelada y el animador está obligado a abrir la 3, luego se tiene que

$$P(R_3|G_1, I) = 1. \tag{6.45}$$

Sólo en el caso en que la puerta ganadora sea la escogida por el concursante, en este caso la 2, es que el animador tiene alguna elección sobre qué puerta abrir. De hecho puede abrir ya sea la 1 o la 3, indistintamente. La probabilidad de ambas es por tanto $1/2$,

$$P(R_1|G_2, I) = P(R_3|G_2, I) = \frac{1}{2}. \tag{6.46}$$

Reemplazando (6.41), (6.43), (6.44), (6.45) y (6.46) en el teorema de Bayes (6.42) tenemos

$$P(G_i|R_3, I) = \frac{\cancel{(\frac{1}{3})} P(R_3|G_i, I)}{\cancel{(\frac{1}{3})}(\frac{1}{2} + 1)} = \frac{2}{3} P(R_3|G_i, I), \tag{6.47}$$

es decir, las probabilidades luego de ver la puerta 3 abierta son

$$P(G_1|R_3, I) = \frac{2}{3}, \tag{6.48a}$$

$$P(G_2|R_3, I) = \frac{1}{3}, \tag{6.48b}$$

$$P(G_3|R_3, I) = 0. \tag{6.48c}$$

Sorprendentemente, ¡al concursante le conviene cambiarse de puerta, ya que tiene el doble de probabilidad de ganar! Este resultado puede verificarse mediante simulaciones computacionales en un par de líneas de código, y así es posible convencerse de que en realidad la información ganada al abrirse la puerta es relevante y no debe ser desechada.

Recuadro 6.4 — Marilyn vos Savant
Aunque el problema fue presentado y resuelto en 1975, sólo se hizo famoso luego de la controversia en la columna «Pregúntale a Marilyn» de la brillante[(2)] escritora Marilyn vos Savant. Cuando se le plantea el problema en 1990, Vos Savant explica la solución correcta: conviene cambiarse de puerta ya que hay una probabilidad $2/3$ de ganar el premio, versus $1/3$ si uno no se cambia, debido al hecho clave de que el animador siempre debe revelar una cabra. Esta solución fue severamente criticada por sus lectores, recibiendo alrededor de 10 mil cartas, incluidas unas mil de distinguidos doctores en matemática y ciencias exactas, quienes tardaron bastante tiempo en reconocer que Vos Savant siempre tuvo razón. Por ejemplo, un doctor de la Universidad de Florida termina su argumento así:

> «There is enough mathematical illiteracy in this country, and we don't need the world's highest IQ propagating more. Shame!»

6.5.2 Teorías conspirativas y la navaja de Ockham

Los alunizajes del programa Apolo en la década de los 70 nunca ocurrieron, sino que fueron falsificados por la NASA bajo órdenes del gobierno de Estados Unidos. Las tecnologías basadas en el movimiento perpetuo y la fusión fría son efectivas, pero han sido suprimidas por agencias de gobierno. Algunos virus como el HIV o el SARS-CoV-2 no son naturales sino que fueron creados en laboratorios. Paul McCartney murió en un accidente automovilístico a finales de los 60 y fue reemplazado por un doble, y Los Beatles dejaron pistas de aquello en las letras de algunas canciones.

Todas estas afirmaciones son falsas, por supuesto, pero son ejemplos de *teorías conspirativas* que han atraído en su momento bastante atención. Un elemento transversal a todas estas teorías, de la cual toman su nombre, es que existe una gran conspiración que ha logrado ocultar la verdad al mundo, salvo a unos pocos elegidos.

Para entender qué nos dice el teorema de Bayes acerca de este fenómeno, diremos que una teoría conspirativa T tiene como característica que es una hipótesis compuesta de m proposiciones individuales $T_1, T_2, \ldots, T_m$ (con m inusualmente grande) tal que

$$T = T_1 \wedge T_2 \wedge T_3 \wedge \ldots \wedge T_m, \tag{6.49}$$

y además T es tal que explica perfectamente la evidencia observada E, es decir, $T \Rightarrow E$, por lo tanto

$$P(E|T, I_0) = 1.$$

[(2)] Marilyn vos Savant fue poseedora del récord Guiness al mayor coeficiente intelectual (CI) en 1988.

De acuerdo a (6.49), la probabilidad de T en un estado de conocimiento I_0 siempre puede descomponerse por medio de la regla del producto como

$$\begin{aligned} P(T|I_0) &= P(T_1, \ldots, T_m|I_0) \\ &= P(T_1|I_0)P(T_2, \ldots, T_m|T_1, I_0) \\ &= P(T_1|I_0)P(T_2|T_1, I_0)P(T_3, \ldots, T_m|T_1, T_2, I_0) \\ &\vdots \end{aligned} \tag{6.50}$$

y así podemos continuar iterativamente, con lo que finalmente obtenemos

$$P(T|I_0) = \underbrace{P(T_1|I_0)}_{=\rho_1} \prod_{i=2}^{m} \underbrace{P(T_i|T_1, \ldots, T_{i-1}, I_0)}_{=\rho_i \text{ para } i=2,\ldots,m} = \prod_{i=1}^{m} \rho_i, \tag{6.51}$$

es decir, $P(T|I)$ es el producto de m factores ρ_i con $i = 1, \ldots, m$, siendo cada uno de ellos la probabilidad de una proposición y por tanto

$$\rho_i \leq 1.$$

Ahora volviendo al teorema de Bayes, tenemos que

$$P(T|E, I_0) = \frac{P(T|I_0)\cancelto{1}{P(E|T, I_0)}}{P(E|I_0)} = \frac{P(T|I_0)}{P(E|I_0)}, \tag{6.52}$$

pero como la probabilidad previa de la evidencia $P(E|I_0)$ no depende de los ρ_i ni de m, podemos reemplazarla por una constante η, y usar (6.51) para escribir

$$P(T|E, I_0) = \frac{1}{\eta} \prod_{i=1}^{m} \rho_i = \frac{1}{\eta} \exp\Big(\sum_{i=1}^{m} \ln \rho_i \Big). \tag{6.53}$$

Como $\rho_i \leq 1$ vemos que $\ln \rho_i \leq 0$ y por lo tanto

$$\sum_{i=1}^{m} \ln \rho_i \leq 0.$$

Definiendo por conveniencia

$$L := \frac{1}{m} \sum_{i=1}^{m} \ln \rho_i \leq 0 \tag{6.54}$$

podemos escribir

$$P(T|E, I_0) = \frac{1}{\eta} \exp(-m|L|) = \frac{q^m}{\eta} \tag{6.55}$$

con

$$q := \exp(-|L|) \leq 1.$$

De esta forma, a medida que crece m la probabilidad de T disminuye, esto es, la teoría conspirativa se hace cada vez menos probable a medida que aumenta el número de suposiciones auxiliares necesarias para soportarla, en una clara manifestación del principio llamado de la *navaja de Ockham*.

Este principio originalmente se refiere a la frase [3]

Entia non-sunt multiplicanda praeter necessitatem

esto, es, «las entidades no deben multiplicarse sin necesidad». El nombre «navaja» es justamente porque pretende cortar los elementos innecesarios de una explicación. Nosotros enunciaremos el principio de la siguiente forma.

Recuadro 6.5 — La navaja de Ockham

Entre dos explicaciones posibles a un hecho, deberíamos escoger la que requiere un menor número de suposiciones.

A veces se suele parafrasear el principio de la navaja de Ockham como «La explicación más simple es la correcta» pero esto comete dos errores. Primero, una explicación más simple no siempre es la que posee menos suposiciones, y segundo, el principio no declara con certeza que la explicación elegida será la correcta. Puesto que es un principio de inferencia, será la mejor explicación con la información disponible, provisto que en efecto sea capaz de explicar los hechos.

En *Dr. House*, precisamente en el capítulo llamado «La navaja de Occam», aparece el siguiente diálogo.

Dr. House: Ninguna condición explica todos estos síntomas. Pero el naranja y el verde lo abarcan todo.

Dr. Chase: ¿El naranja *y el verde*? ¿Dos condiciones contraídas simultáneamente?

Dr. Foreman: La navaja de Occam. La explicación más simple es siempre la mejor.

Dr. House: ¿Y crees que uno es más simple que dos?

Dr. Cameron: Seguro que lo es, sí.

Dr. House: Se aparece un bebé. Chase te dice que dos personas intercambiaron fluidos para hacer a la criatura. Yo les digo que una cigüeña dejó caer al pequeño en un pañal. ¿Van a ir por dos o por uno?

Dr. Foreman: Creo que su argumento es engañoso.

Dr. House: Creo que tu corbata es fea. ¿Por qué es uno más simple que dos? ¿Es más bajo? ¿Más solitario? ¿Es más simple? Cada una de esas condiciones tiene una posibilidad en mil, lo que significa que ambas ocurriendo a la vez es de una en un millón. Chase dice que la infección cardíaca es de una en 10 millones, entonces mi idea es diez veces mejor que la suya.

Efectivamente, el ejemplo de House usando a la cigüeña es engañoso, porque estamos comparando dos hipótesis que no explican de igual manera los hechos, luego en esta etapa (donde una falla en explicar y la otra no) no tiene

[3] Frase atribuida a William of Ockham (1287 - 1347), fraile franciscano inglés, cuyo apellido es a veces escrito como Occam.

sentido comparar cuál involucra más suposiciones. Por otro lado, el cálculo respecto a las dos enfermedades ocurriendo simultáneamente es precisamente el tipo de regla que hemos obtenido en (6.55).

6.6 — Modelos probabilísticos

Como un adelanto de lo que veremos en el Capítulo 7, describiremos lo que comúnmente se conoce como distribución de probabilidad como un *modelo probabilístico*, esto es, un conjunto de números que capturan la información que poseemos sobre una situación o sistema. Fijaremos ahora la notación que usaremos para describir estos modelos probabilísticos.

Denotaremos el hecho que una variable X es descrita por un modelo M como $X \sim M$. Si este es el caso, la distribución de probabilidad de X será escrita como $P(X|M)$, y en el caso en que el modelo M reciba parámetros $\boldsymbol{\theta}$ escribiremos $X \sim M(\boldsymbol{\theta})$. Sin embargo, simplificaremos la notación para la distribución de probabilidad de X de forma que

$$P(X|M(\boldsymbol{\theta})) \to P(X|\boldsymbol{\theta})$$

cuando el modelo M se entienda a partir del contexto y de los nombres de los parámetros.

Como ejemplo, tomemos la famosa distribución normal, la cual analizaremos en detalle en el Capítulo 8. Diremos que una variable real X es descrita por un modelo normal si su densidad de probabilidad es

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{\sigma^2}(x-\mu)^2\Big) \tag{6.56}$$

donde μ y σ son los parámetros θ_1, θ_2 del modelo. Diremos que $X \sim \mathcal{N}(\mu, \sigma)$, entendiendo que la parte de nuestro estado de conocimiento que es relevante a la variable X se encuentra condensada en los parámetros μ, σ. La misma notación será usada para variables discretas, por ejemplo diremos que una variable entera $k \in \{0, 1, 2, \ldots\}$ es descrita por un modelo de Poisson si su probabilidad es

$$P(k|\lambda) = \frac{\exp(-\lambda)\lambda^k}{k!}, \tag{6.57}$$

donde toda la información relevante a k queda capturada en el parámetro λ, y en este caso escribiremos $k \sim \text{Pois}(\lambda)$.

6.7 — Probabilidad y frecuencia de eventos

Veremos a continuación la aplicación del teorema de Bayes a un problema del tipo que exploraremos con más detalle en el Capítulo 9, la determinación de los mejores parámetros de un modelo cuando tenemos observaciones o datos.

Imaginemos un plebiscito entre dos opciones, digamos azul (■) y roja (■), respecto a la cual quisiéramos hacer una encuesta. Para esto tomamos una muestra de n personas de un total N mucho mayor y les consultamos acerca de su preferencia en el plebiscito. Supongamos más aún que la preferencia de cada persona ya está decidida aunque nosotros no la conocemos, y es tal que una persona cualquiera votará azul con probabilidad $p := P(\blacksquare|I)$, y por lo tanto rojo con probabilidad

$$P(\blacksquare|I) = P(\neg\,\blacksquare|I) = 1 - p.$$

Nuestro objetivo es determinar p cuando sabemos que k personas de la muestra de n contestaron azul en nuestra encuesta.

Si llamamos A_i a la proposición «El voto i-ésimo es azul» y designamos «El voto j-ésimo es rojo» como R_j, de forma que A_i y R_i son mutuamente excluyentes y exhaustivas, entonces la proposición

$$E = \text{«}k \text{ votantes son del grupo azul en una muestra de } n\text{»},$$

que es nuestra evidencia, es la disyunción

$$E = D_1 \vee D_2 \vee \ldots \vee D_m$$

de m permutaciones (ordenamientos) de k proposiciones de tipo A y $(n-k)$ de tipo R. Cada ordenamiento posible tendrá exactamente la misma probabilidad, dada por

$$P(D_i|k,p,n) = P(\blacksquare|I)^k P(\blacksquare|I)^{n-k} = p^k\,(1-p)^{n-k}, \tag{6.58}$$

y existen $m = C(k,n)$ posibles ordenamientos (permutaciones) de k elementos de un total de n, donde $C(k,n)$ es el coeficiente binomial

$$C(k,n) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

que introdujimos en (**3.116**). Luego, la probabilidad de obtener k azules en la muestra de n cuando no nos interesa cuál de los ordenamientos posibles obtuvimos está dada por

$$\begin{aligned} P(k|p,n) &= P(D_1,\ldots,D_m|k,p,n) \\ &= \sum_{i=1}^{m} P(D_i|k,p,n) \\ &= \sum_{i=1}^{m} p^k(1-p)^{n-k} = \binom{n}{k} p^k(1-p)^{n-k}, \end{aligned} \tag{6.59}$$

conocida como la distribución binomial.

Definición 6.2 — Distribución binomial
La probabilidad de k resultados verdaderos independientes de un total de n, cuando p es la probabilidad de un resultado verdadero, está dada por

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{6.60}$$

conocida como la distribución binomial, donde

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{6.61}$$

es el coeficiente binomial.

Queremos determinar el mejor valor de p, o en realidad, la distribución de probabilidad de los valores de p dado el conocimiento de k y n. Para esto necesitamos aplicar el teorema de Bayes, sin embargo antes debemos asignar una distribución inicial para p, que supondremos uniforme entre 0 y 1. Entonces diremos

$$P(p|I_0) = \text{rect}(p;0,1), \tag{6.62}$$

con lo que la aplicación del teorema de Bayes nos entrega

$$\begin{aligned} P(p|k,n) &= \frac{P(p|I_0)\ P(k|p,n)}{P(k|n)} \\ &= \frac{\text{rect}(p;0,1)}{P(k|n)} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \text{rect}(p;0,1) \frac{p^k (1-p)^{n-k}}{\eta(k,n)}, \end{aligned} \tag{6.63}$$

donde hemos absorbido $\binom{n}{k}$ y $P(k|n)$ en la constante de normalización $\eta(k,n)$, que a su vez está dada por una función Beta,

$$\eta(k,n) = \int_0^1 \mathrm{d}p\ p^k (1-p)^{n-k} = B(k+1, n-k+1). \tag{6.64}$$

Finalmente, escribimos nuestra distribución posterior como

$$P(p|k,n) = \frac{p^k (1-p)^{n-k}}{B(k+1,n-k+1)} \quad \text{con} \ \ p \in [0,1] \tag{6.65}$$

la cual se conoce como la distribución beta.

Definición 6.3 — Distribución beta
Diremos que una variable $X \in [0,1]$ sigue una distribución beta si

$$P(X|\alpha,\beta) = \frac{X^{\alpha-1}(1-X)^{\beta-1}}{B(\alpha,\beta)}, \tag{6.66}$$

donde $B(\alpha,\beta)$ es la función beta.

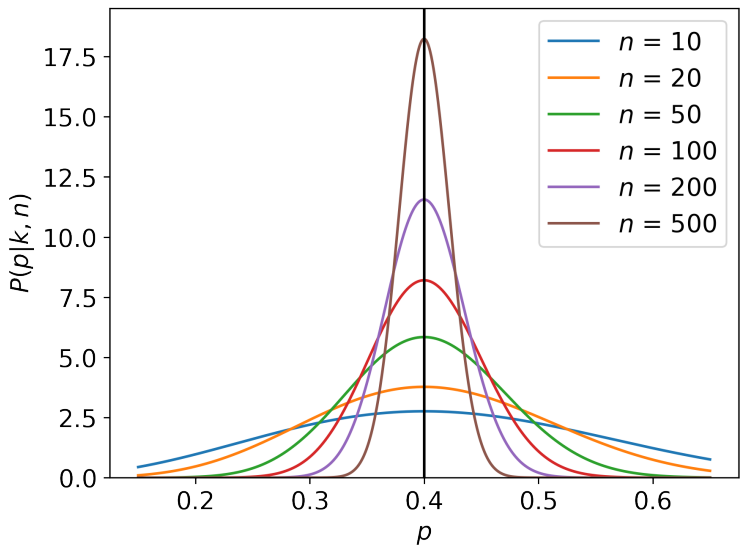


Figura 6.2: Distribución posterior para el parámetro p de la distribución binomial, de acuerdo a (6.65), para distintos valores de n con $f = k/n = 0.4$.

Ahora definamos la frecuencia f de votos azules como

$$f := \frac{k}{n} = \frac{\text{número de votos azules}}{\text{número total de votos}} \tag{6.67}$$

y consideremos el límite con $n \to \infty$ y $k \to \infty$ pero manteniendo f fijo. Vemos que

$$\begin{aligned} p^k(1-p)^{n-k} &= \exp(k \ln p + (n-k)\ln(1-p)) \\ &= \exp\left(n[f \ln p + (1-f)\ln(1-p)]\right) \end{aligned} \tag{6.68}$$

por lo tanto el valor más probable de p según $P(p|k,n)$, denotado por p^*, es tal que

$$\left(\frac{\partial}{\partial p} \ln P(p|k,n)\right)_{p=p^*} = 0, \tag{6.69}$$

es decir,

$$0 = \frac{\partial}{\partial p}\Big(f \ln p + (1-f)\ln(1-p)\Big)_{p=p^*} = \frac{f}{p^*} - \frac{1-f}{1-p^*}, \tag{6.70}$$

por lo que se obtiene el notable resultado

$$p^* = f \tag{6.71}$$

para cualquier valor de n. Esto significa que un valor representativo de la probabilidad $P(\blacksquare|I)$ es la frecuencia f de votos azules. Es más, un poco de cálculo nos permite ver que, de acuerdo a la distribución beta en (6.65), el valor esperado de p es

$$\langle p \rangle_{k,n} = \frac{B(k+2, n-k+1)}{B(k+1, n-k+1)} = \frac{k+1}{n+2} = \frac{nf+1}{n+2} \tag{6.72}$$

que también tiende a f para $n \to \infty$. Usando la distribución posterior (6.65) podemos construir el cuociente

$$\frac{P(p|k,n)}{P(p=f|k,n)} = \exp\left(-n\left|f \ln \frac{p}{f} + (1-f)\ln\frac{(1-p)}{(1-f)}\right|\right) \leq 1 \tag{6.73}$$

de forma que va a cero rápidamente con $n \to \infty$ para $p \neq f$, y va a 1 para $p = f$. La probabilidad posterior $P(p|k,n)$ entonces estará infinitamente concentrada en $p = f$, como puede verse en la **Figura 6.2**. Dado que la distribución está normalizada para todo n, debe cumplirse que

$$\lim_{n\to\infty} P(p|k,n) = \delta(p-f), \tag{6.74}$$

y por tanto

$$\lim_{n\to\infty} P(\blacksquare|k,n) = \lim_{n\to\infty} \left[\int_0^1 \mathrm{d}p \underbrace{P(\blacksquare|p)}_{=p} \underbrace{P(p|k,n)}_{=\delta(p-f)} \right] = f = \frac{k}{n}. \tag{6.75}$$

Hemos llegado a un resultado notable de la teoría de la probabilidad en la interpretación bayesiana, y es que, a pesar de que la probabilidad p es un grado de plausibilidad «personal», suficiente evidencia lo lleva a coincidir con una cantidad objetiva, la frecuencia de casos observados, bajo ciertas suposiciones. Veremos en mayor profundidad este hecho en la **Sección 9.3**.

▶ Para una introducción accesible y didáctica acerca del teorema de Bayes y su uso en inferencia se recomienda el libro de Stone (2013).

6.8 — Probabilidades en ausencia de incerteza

Una de las «condiciones de contorno» de la inferencia es que debe reducirse a la deducción cuando tenemos toda la información. Esto significa que debe haber un paso «suave» a medida que acumulamos información hacia una *distribución de conocimiento completo*, que asigna cero incerteza a una variable o conjunto de variables. ¿Cuáles son estas distribuciones de conocimiento completo? Recordemos que la expectación de un valor conocido es el propio valor, esto es,

$$\langle X \rangle_{X=x_0,I} = x_0, \tag{6.76}$$

para cualquier estado de conocimiento I. Si usando este resultado construimos, para una función arbitraria $\omega(f)$ donde $f = f(\boldsymbol{u})$, la expectación

$$\langle \omega(f) \rangle_{\boldsymbol{u}=\boldsymbol{u}_0,I} = \omega(f(\boldsymbol{u}_0)) \tag{6.77}$$

y luego escogemos $\omega(f) = \delta(f-F)$, obtenemos

$$\langle \delta(f-F) \rangle_{\boldsymbol{u}=\boldsymbol{u}_0,I} = \delta(F - f(\boldsymbol{u}_0)) \tag{6.78}$$

pero por la definición de la densidad de probabilidad,

$$\boxed{P(f = F|\boldsymbol{u} = \boldsymbol{u}_0, I) = \delta(F - f(\boldsymbol{u}_0)),} \tag{6.79}$$

es decir, la densidad de probabilidad de una función f en cualquier estado de conocimiento donde dicha función puede ser evaluada exactamente, es la delta de Dirac. A este tipo de distribuciones las denominaremos distribuciones de conocimiento completo, ya que representan un estado de cero incerteza respecto a la cantidad. De la misma manera, para variables continuas $\boldsymbol{X}$ la densidad de probabilidad que representa conocimiento completo es la delta de Dirac,

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{X} = \boldsymbol{x}_0, I) = \delta(\boldsymbol{x} - \boldsymbol{x}_0) \tag{6.80}$$

para todo I. En el caso de una variable discreta $X \in \{x_1, \ldots, x_n\}$, la probabilidad que representa conocimiento completo es la función indicador,

$$P(X = x|X = x_k, I) = \mathrm{Q}(x = x_k) \tag{6.81}$$

para todo I.

Ejemplo 6.8.1. *La altura $h(t)$ de un cuerpo en caída libre desde una altura inicial $h(0) = H$ es*

$$h(t; g, H) = H - \tfrac{1}{2}gt^2, \qquad \text{con } t \leq \sqrt{2H/g}. \tag{6.82}$$

La distribución de h dado H, g y t es de conocimiento completo, y representa el modelo determinista

$$P(h|t, g, H) = \delta\left(h - H + \tfrac{1}{2}gt^2\right) \quad \text{con } H \geq \tfrac{1}{2}gt^2. \tag{6.83}$$

Supongamos ahora que no conocemos H exactamente, sino que le asignamos un modelo $P(H|I) = p(H)$. Podemos eliminar el parámetro H de nuestro modelo determinista y el costo de ello es obtener el modelo probabilístico

$$\begin{aligned} P(h|t, g, I) &= \int_0^\infty \mathrm{d}H P(h, H|t, g, I) \\ &= \int_0^\infty \mathrm{d}H P(h|t, g, H, I) P(H|t, g, I) \\ &= \int_0^\infty \mathrm{d}H P(h|t, g, H) \frac{P(H|I)P(g, t|H, I)}{P(g, t|I)}, \end{aligned} \tag{6.84}$$

donde hemos usado el teorema de Bayes en la última línea. Llamando

$$Z(g, t) := P(g, t|I) = \int_0^\infty \mathrm{d}H P(H|I) P(g, t|H) \tag{6.85}$$

y tomando $P(g, t|H) = \Theta(H - \frac{1}{2}gt^2)$ podemos escribir

$$\begin{aligned} P(h|t, g, I) &= \frac{1}{Z(g, t)} \int_0^\infty \mathrm{d}H P(h|t, g, H) p(H) \Theta(H - \tfrac{1}{2}gt^2) \\ &= \frac{1}{Z(g, t)} \int_0^\infty \mathrm{d}H p(H) \delta\left(H - h - \tfrac{1}{2}gt^2\right) \Theta(h) \\ &= \frac{p(h + \frac{1}{2}gt^2)\Theta(h)}{Z(g, t)}. \end{aligned} \tag{6.86}$$

6.9 — Otras interpretaciones del concepto de probabilidad

Para cerrar este capítulo, compararemos nuestra formulación con otros enfoques alternativos para la idea de probabilidad.

6.9.1 La probabilidad según Kolmogorov

La formulación de la probabilidad de Kolmogorov (1933) es de carácter abstracto, ya que en ella sólo se especifica que una probabilidad $p(E)$ se asigna a todo conjunto $E \subset \Omega$, donde Ω es denominado el espacio muestral. Los siguientes axiomas definen a $p(\bullet)$ como una probabilidad.

Recuadro 6.6 — Los axiomas de Kolmogorov

$$\begin{aligned} &\text{La probabilidad } p(E) \text{ es un número real no negativo,} && \text{(K1)}\\ &\text{La probabilidad del espacio muestral } \Omega \text{ es } p(\Omega) = 1, && \text{(K2)}\\ &\text{Si } E_1 \text{ y } E_2 \text{ son disjuntos, } p(E_1 \cup E_2) = p(E_1) + p(E_2), && \text{(K3)} \end{aligned}$$

Para nosotros (K3) es análogo al caso particular de la regla extendida de la suma (6.9) cuando $A \wedge B = \mathbb{F}$. Además de estos axiomas, se define la probabilidad condicional

$$p_A(B) := \frac{p(A \cap B)}{p(A)}, \qquad \text{(K4)}$$

y esta relación constituye el análogo a nuestra regla del producto (6.4). Es importante notar aquí que no se especifica el significado de p ni qué representan los conjuntos E y Ω, y esto hace que el formalismo sea sumamente general, pero por lo mismo vacío de semántica. Sin embargo, la probabilidad definida así es siempre *absoluta*, en el sentido de que $p(E)$ depende únicamente de E, e incluso las probabilidades condicionales definidas en (K4) no se interpretan en términos de información conocida.

> **Importante:** No debemos confundir las operaciones unión $\cup$ e intersección $\cap$ entre conjuntos con las operaciones lógicas disyunción $\vee$ y conjunción $\wedge$. Aunque existe un paralelo claro entre ellas, conceptualmente no representan lo mismo.

6.9.2 Probabilidad desde la expectación de Whittle

El tratamiento de la probabilidad que hemos desarrollado aquí, en cuanto al formalismo matemático, es muy cercano al expuesto en el libro de Whittle (2000), en el que la expectación es el elemento fundamental del cual la probabilidad se deduce usando funciones indicador, y de hecho algunos de sus

postulados son equivalentes a los nuestros, como **Aditividad de la estimación** y **Conservación del orden**. Sin embargo, Whittle llega a una definición **frecuentista** de probabilidad inspirada en la operación promedio, donde el estado de conocimiento no juega papel, y por esto al igual que Kolmogorov al conjunto de sus cinco postulados debe añadir como una definición adicional la expectación condicional, que para nosotros (como para cualquier «bayesianista») es un elemento ineludible.

6.9.3 Probabilidad bayesiana según Cox

En el trabajo fundacional de Richard T. Cox (1946), se proponen los siguientes axiomas que definen la probabilidad, entendida $P(A|I)$ del mismo modo que nosotros, como un grado de creencia plausible en la proposición A bajo la hipótesis I.

Recuadro 6.7 — Los axiomas de Cox

La probabilidad $P(A \wedge B|I)$ es función de $P(A|I)$ y $P(B|A \wedge I)$, (C1)

La probabilidad $P(\neg A|I)$ es función de $P(A|I)$. (C2)

El axioma (**C1**) se traduce en la búsqueda de una función F tal que

$$P(A \wedge B|I) = F\left(P(B|A \wedge I), P(A|I)\right), \tag{6.87}$$

mientras que (**C2**) implica la existencia de la función S tal que

$$P(\neg A|I) = S(P(A|I)). \tag{6.88}$$

Cox demuestra que el exigir compatibilidad con las reglas de la lógica, es decir, exigir que si $A = B$ entonces $P(A|I) = P(B|I)$ para todo I, lleva a que sin pérdida de generalidad las funciones F y S pueden ser tomadas como

$$F(x, y) = xy, \tag{6.89a}$$

$$S(x) = 1 - x, \tag{6.89b}$$

vale decir, cualquier otra elección de F y S compatible con los axiomas puede ser reducida a (**6.89**), que por supuesto son las reglas del producto (**6.4**) y de la suma (**6.3**), respectivamente.

► Para más detalles sobre los orígenes de la teoría bayesiana de la probabilidad, se recomienda el libro (póstumo) de Jaynes (2003) y el libro de Sivia y Skilling (2006).

PROBLEMAS

Problema 6.1. *Deduzca la regla extendida de la suma (6.9) a partir de la regla de la suma (6.3), la regla del producto (6.4) y las leyes de De Morgan.*

Problema 6.2. *Demuestre que si $A \Rightarrow B$ es cierto, se tiene que*

(a) a mayor probabilidad de A, mayor probabilidad de B,

(b) a menor probabilidad de B, menor probabilidad de A.

Problema 6.3. *En un juego de azar comenzamos con cero pesos, y a cada lanzamiento de un dado, que supondremos balanceado, ganamos 500 pesos si el dado sale 3 o 5, y perdemos 500 pesos si sale 1, 2, 4, o 6. Si perdemos 500 pesos cuando tenemos cero en nuestro total, el juego termina. Cuál es la probabilidad de que nuestro juego dure 4 lanzamientos?*

Problema 6.4. *Las cartas Zener se utilizan para poner a prueba supuestas habilidades paranormales (percepción extrasensorial, o ESP). Estas cartas tienen 5 posibles diseños: círculo, cruz, líneas onduladas, cuadrado, estrella. El sujeto a prueba debe adivinar un cierto número de veces la cara de la carta sin mirar. Si ud. observa a una persona adivinar 5 de 5 intentos, y su probabilidad previa de que esta persona presente habilidades perfectas de adivinación es de 0.1, cuál debiera ser su nueva probabilidad? Considere sólo dos posibles hipótesis, T si el individuo presenta habilidades paranormales, y $\neg T$ si sólo es resultado del azar.*

Problema 6.5. *En un casino existen dos máquinas de juegos, una en la cual la probabilidad de ganar es de 0.1, y la otra en la cual la probabilidad de ganar es de 0.2. Por supuesto, no sabemos cuál máquina es cuál, y para averiguarlo, decidimos que* a priori *la probabilidad de elegir la mejor máquina es 1/2. Elegimos una máquina cualquiera y perdemos. ¿Cuál es la probabilidad de haber elegido la mejor máquina?*

Problema 6.6. *Supongamos tres cartas, idénticas en forma, excepto que la primera carta tiene ambas caras rojas, la segunda tiene ambas caras negras, y la tercera tiene una cara roja y otra negra. Se nos muestra una de las cartas extraída al azar, cuyo lado visible es rojo. ¿Cuál es la probabilidad de que la cara oculta sea negra?*

Problema 6.7. *Use el teorema de Bayes para explicar por qué no debemos creer en alguien que hace predicciones y acierta, si estas predicciones son vagas, por ejemplo que ocurrirá un terremoto en algún lugar del mundo con magnitud mayor a 7 durante 2023.*

CAPÍTULO 7

Distribuciones de probabilidad

> One's ideas must be as broad as Nature if they are to interpret Nature
>
> Sherlock Holmes, A Study in Scarlet

Gracias a la formulación de la probabilidad en términos bayesianos que vimos en el **Capítulo 6**, ahora tenemos todas las herramientas y el lenguaje apropiado para poder formular nuestras ideas del **Capítulo 1**, donde los modelos que crearemos bajo información incompleta serán *modelos probabilísticos*, conocidos comúnmente en el lenguaje estándar de la probabilidad y estadística como *distribuciones de probabilidad*. Estos modelos describen todo aquello que conocemos acerca de una cantidad o conjunto de cantidades, que por ejemplo podrían corresponder a las variables relevantes que describen un fenómeno, o los grados de libertad de un sistema de interés.

La estadística tradicional se basa en importante medida en reducir el conocimiento contenido en estos modelos probabilísticos a un conjunto de *descriptores*, usualmente no más de dos. En particular, estamos interesados en descriptores que nos permitan capturar la información de una variable en términos de un *valor central* en torno al cual creemos que se encuentra, y una *incerteza* que nos indique el ancho o volumen de la región que concentra la mayor probabilidad. Conceptos como la media y la varianza aparecen profusamente en la estadística, y ahora veremos cómo llegar a ellos.

7.1 — MEDIDAS CENTRALES

A continuación nos enfocaremos en el problema de definir un valor representativo de una variable $X \sim I$. Llamaremos un estimador $\hat{x}$ a uno de estos posibles valores representativos de X, y definiremos criterios para encontrar tales estimadores. Estos criterios se basan en la minimización de la expectación de cierta **función pérdida** $L(X, \hat{x})$, esto es,

$$\hat{x} := \arg\min_{y} \left\langle L(X, y) \right\rangle_I. \tag{7.1}$$

Esto implica que $\hat{x}$ debe cumplir la condición de extremo

$$\begin{aligned} 0 = \frac{\partial}{\partial \hat{x}} \langle L(X,\hat{x}) \rangle_I &= \frac{\partial}{\partial \hat{x}} \int \mathrm{d}x P(X=x|I) L(x,\hat{x}) \\ &= \int \mathrm{d}x P(X=x|I) \frac{\partial L(x,\hat{x})}{\partial \hat{x}} \\ &= \left\langle \frac{\partial L(X,\hat{x})}{\partial \hat{x}} \right\rangle_I . \end{aligned} \tag{7.2}$$

Comenzaremos definiendo la **media** o valor medio de una distribución de probabilidad.

Definición 7.1 — Media de una distribución
Definiremos la media de una variable discreta $x \sim I \in \{x_1, \ldots, x_n\}$ como

$$\langle X \rangle_I := \sum_{i=1}^{n} P(X=x_i|I) x_i. \tag{7.3}$$

Similarmente, para una variable continua $x \sim I \in [a,b]$ la media será definida como

$$\langle X \rangle_I := \int_a^b \mathrm{d}x P(X=x|I) x. \tag{7.4}$$

Es importante notar que, en general, para una variable discreta la media podría no corresponder a ninguno de los valores permitidos de la variable, ya que es una suma ponderada de éstos. Por ejemplo, para $X \in \{0,2\}$ con probabilidades $P(X=0|I) = 1/2$ y $P(X=2|I) = 1/2$ se tendrá

$$\langle X \rangle_I = 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = 1,$$

pero $P(X=1|I) = 0$.

Esta definición de la media puede obtenerse a partir de tomar como función pérdida el *cuadrado del error* cometido al usar $\hat{x}$ como substituto del valor X, esto es,

$$L(X,\hat{x}) = (X-\hat{x})^2, \tag{7.5}$$

con lo que la condición de extremo nos entrega

$$0 = \left\langle \frac{\partial}{\partial \hat{x}} (X-\hat{x})^2 \right\rangle_I = -2\langle X-\hat{x} \rangle_I = -2\Big(\langle X \rangle_I - \hat{x} \Big), \tag{7.6}$$

es decir, nuestro estimador óptimo $\hat{x}$ coincide con la media $\langle X \rangle_I$.

Para otras funciones pérdida tendremos distintos estimadores óptimos. Por ejemplo, si sólo nos interesa el que $\hat{x}$ coincida con el verdadero valor X, y no nos preocupa la magnitud del error cometido, entonces podemos emplear la función pérdida

$$L(X,\hat{x}) = \mathrm{Q}(X \neq \hat{x}) = \begin{cases} 0 \text{ si } X = \hat{x}, \\ \\ 1 \text{ en caso contrario}, \end{cases} \tag{7.7}$$

la cual es mínima si $X = \hat{x}$. En ese caso la condición de extremo nos entrega

$$\begin{aligned} 0 = \frac{\partial}{\partial \hat{x}} \langle \mathrm{Q}(X \neq \hat{x}) \rangle_I &= \frac{\partial}{\partial \hat{x}} \Big(1 - \big\langle \mathrm{Q}(X = \hat{x}) \big\rangle_I \Big) \\ &= \frac{\partial}{\partial \hat{x}} \Big(1 - P(X = \hat{x}|I) \Big) \\ &= -\frac{\partial}{\partial \hat{x}} P(X = \hat{x}|I), \end{aligned} \tag{7.8}$$

es decir,

$$\left(\frac{\partial}{\partial x} P(X = x|I) \right)_{x=\hat{x}} = 0, \tag{7.9}$$

lo cual nos lleva a definir la **moda** de una distribución como sigue.

Definición 7.2 — Moda de una distribución
Definiremos la moda x^* de una variable $X \sim I$ como

$$x^* := \arg\max_x P(X = x|I), \tag{7.10}$$

esto es, corresponde al valor de X que tiene probabilidad máxima.

A diferencia de la media, la moda siempre es uno de los valores permitidos de la variable. Finalmente, consideremos el caso en que la función pérdida es la distancia entre el estimador y el verdadero valor,

$$L(X, \hat{x}) = |X - \hat{x}|. \tag{7.11}$$

En este caso se tiene

$$\frac{\partial}{\partial \hat{x}} \big\langle |X - \hat{x}| \big\rangle_I = -\left\langle \frac{X - \hat{x}}{|X - \hat{x}|} \right\rangle_I = -\big\langle \operatorname{sgn}(X - \hat{x}) \big\rangle_I = 0, \tag{7.12}$$

donde sgn es la función signo, que podemos escribir en términos de funciones indicador como

$$\operatorname{sgn}(z) = 2\,\mathrm{Q}(z > 0) - 1. \tag{7.13}$$

Reemplazando en (7.12), tenemos

$$0 = \big\langle \operatorname{sgn}(X - \hat{x}) \big\rangle_I = 2\big\langle \mathrm{Q}(X > \hat{x}) \big\rangle_I - 1, \tag{7.14}$$

y por lo tanto

$$\Big\langle \mathrm{Q}(X > \hat{x}) \Big\rangle_I = P(X > \hat{x}|I) = \frac{1}{2}, \tag{7.15}$$

lo cual nos lleva a definir la **mediana** de una distribución como el valor que divide el espacio de valores permitidos en dos regiones, cada una con probabilidad $1/2$.

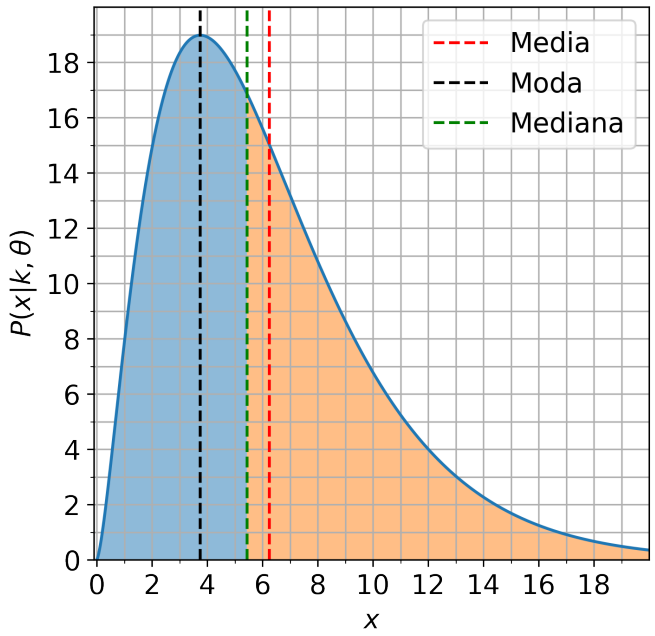


Figura 7.1: Una distribución con valores diferentes de media, moda y mediana. Aunque parezca más pequeña, el área en azul que representa la región de valores menores que la mediana, es exactamente igual al área en naranjo.

Definición 7.3 — Mediana de una distribución
Se define la mediana x_M de una variable $X \sim I$ como el valor de X tal que

$$P(x \leq x_M|I) = P(x > x_M|I) = \frac{1}{2}. \tag{7.16}$$

Un ejemplo de media, moda y mediana para una distribución asimétrica se puede ver en la **Figura 7.1**. El concepto de mediana nos lleva a su vez a definir la **distribución acumulada** para una variable.

Definición 7.4 — Distribución acumulada
Definiremos la distribución acumulada para una variable $X \sim I$ como

$$C(x_0|I) := P(X \leq x_0|I) = \langle \mathrm{Q}(X \leq x_0) \rangle_I \,. \tag{7.17}$$

De esta forma podemos decir que la mediana x_M es el punto donde la distribución acumulada es exactamente $1/2$. Claramente la distribución acumulada $P(X \leq x_0|I)$ tenderá a cero cuando x_0 tiende al límite inferior de X, y tenderá a uno cuando x_0 va al límite superior de X, típicamente teniendo la forma de una *función sigmoide*, como se ve en el panel inferior de la **Figura 7.2**.

La derivada de la distribución acumulada es la densidad de probabilidad, esto es,

$$\frac{\partial C(x_0|I)}{\partial x_0} = P(X = x_0|I). \tag{7.18}$$

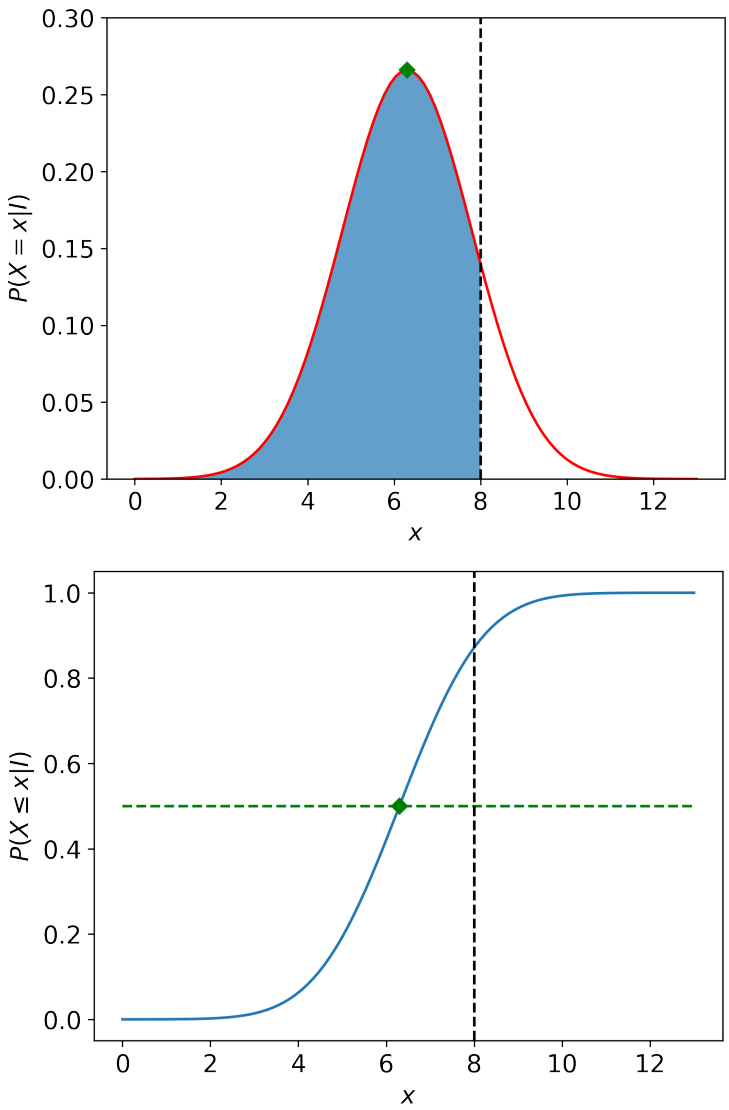


Figura 7.2: En el panel superior se muestra el valor de la distribución acumulada para una distribución normal en $x_0 = 8$ como el área en azul. El panel inferior muestra la distribución acumulada, la cual describe una curva sigmoide. En ambos paneles el diamante en verde representa la mediana, que en este caso coincide con la media y la moda.

Demostración. De la definición de $C(x_0|I)$, se cumple que

$$\frac{\partial}{\partial x_0} P(X \leq x_0|I) = \frac{\partial}{\partial x_0} \langle \mathrm{Q}(X \leq x_0) \rangle_I = \frac{\partial}{\partial x_0} \langle \Theta(x_0 - X) \rangle_I \tag{7.19}$$

Pero

$$\begin{aligned} \frac{\partial}{\partial x_0} \langle A(\bullet; x_0) \rangle_I &= \frac{\partial}{\partial x_0} \int_{-\infty}^{\infty} \mathrm{d}x' P(X = x'|I) A(x'; x_0) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x' P(X = x|I) \frac{\partial A(x'; x_0)}{\partial x_0} \\ &= \left\langle \frac{\partial A(\bullet; x_0)}{\partial x_0} \right\rangle_I \end{aligned} \tag{7.20}$$

luego se tiene

$$\frac{\partial}{\partial x_0} P(X \leq x_0|I) = \left\langle \frac{\partial}{\partial x_0} \Theta(x_0 - X) \right\rangle_I = \langle \delta(x_0 - X) \rangle_I \tag{7.21}$$

que es (7.18)

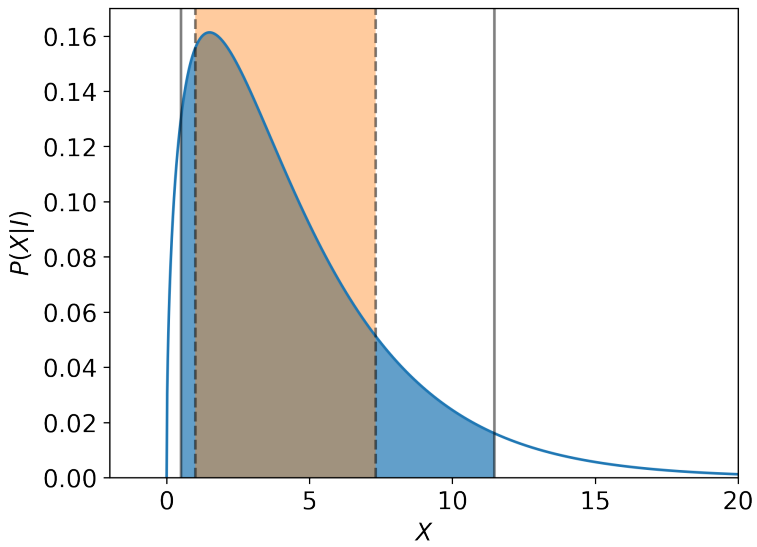


Figura 7.3: La región en naranjo muestra un intervalo de credibilidad $[a, b]$ donde $P(X \in [a, b]|I) = 0.7$, mientras que la región en azul, que incluye a la anterior, representa un intervalo $[a', b']$ tal que $P(X \in [a', b']|I) = 0.9$.

7.2 — Incerteza e intervalos de credibilidad

Por medio de la distribución acumulada es posible definir un *intervalo de credibilidad* para una variable $X \in \Omega$, definido como un intervalo $[a, b]$ tal que el valor de X está en él con probabilidad p, esto es,

$$P(X \in [a, b]|I) = \int_a^b \mathrm{d}x P(X = x|I) = p. \tag{7.22}$$

Recordando la definición de la distribución acumulada,

$$C(x|I) := P(X \leq x|I) = \int_{-\infty}^{x} \mathrm{d}x P(X = x|I), \tag{7.23}$$

entonces (7.22) se reduce a

$$C(b|I) - C(a|I) = p. \tag{7.24}$$

Esto significa que la distribución acumulada es suficiente para definir cualquier intervalo de credibilidad. Si formalmente definimos la *desviación respecto a la media* como

$$\delta X := X - \langle X \rangle_I, \tag{7.25}$$

la cual tiene media cero, tenemos

$$(\delta X)^2 = X^2 + \langle X \rangle_I^2 - 2X\langle X \rangle_I, \tag{7.26}$$

y luego, tomando expectación en el estado I, tenemos

$$\langle (\delta X)^2 \rangle_I = \langle X^2 \rangle_I + \langle X \rangle_I^2 - 2\langle X \rangle_I^2 = \langle X^2 \rangle_I - \langle X \rangle_I^2 \geq 0,$$

lo cual nos lleva a la siguiente definición.

Definición 7.5 — Varianza
Definiremos la varianza de una variable $X \sim I$ como

$$\langle (\delta X)^2 \rangle_I := \langle X^2 \rangle_I - \langle X \rangle_I^2. \tag{7.27}$$

A veces es útil emplear la raíz cuadrada de la varianza, conocida como la *desviación estándar*,

$$\sigma_I := \sqrt{\left\langle (\delta X)^2 \right\rangle_I}, \tag{7.28}$$

la cual nos permite definir un intervalo de credibilidad simétrico en torno a la media. La notación abreviada $X = X_0 \pm \Delta X$ indica que

$$X_0 = \left\langle X \right\rangle_I, \tag{7.29a}$$

$$p = P(|X - X_0| \leq \Delta X | I), \tag{7.29b}$$

con p un valor que se asume implícito dependiendo de ΔX. Por ejemplo, para una distribución normal, $\Delta X = \sigma_I$ corresponde a

$$p = \operatorname{erf}(1/\sqrt{2}) \approx 0.682689.$$

7.3 — Momentos de una distribución

Definiremos el momento n-ésimo de una distribución $P(X|I)$ como

$$\mu_n(I) := \left\langle X^n \right\rangle_I \qquad \text{para } n \geq 0. \tag{7.30}$$

¿Por qué son importantes estos momentos? Su importancia radica en que el conjunto completo de los momentos permite obtener la expectación de cualquier función analítica $g(X)$ ya que, usando la expansión en serie de potencias

$$g(x) = \sum_{n=0}^{\infty} C_n x^n, \tag{7.31}$$

y se tiene que

$$\left\langle g \right\rangle_I = \left\langle \sum_{n=0}^{\infty} C_n X^n \right\rangle_I = \sum_{n=0}^{\infty} C_n \mu_n(I). \tag{7.32}$$

7.3.1 Cumulantes y momentos

A veces es conveniente calcular, para una distribución $P(X|I)$, la expectación de $\exp(tX)$ para un parámetro t arbitrario, ya que esto permite definir los llamados *cumulantes* $\kappa_1, \kappa_2, \ldots$ a través de la función generadora

$$K_I(t) := \ln \left\langle \exp(tX) \right\rangle_I = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}, \tag{7.33}$$

donde la suma comienza de $n = 1$ ya que $\kappa_0 = K_I(0) = 0$. La importancia de los cumulantes es que el primer cumulante κ_1 corresponde a la media de la distribución, y el segundo cumulante κ_2 es la varianza. Esto es,

$$\left\langle X \right\rangle_I = \kappa_1 = \left[\frac{\partial}{\partial t} \ln \left\langle \exp(tX) \right\rangle_I \right]_{t=0}, \tag{7.34a}$$

$$\left\langle (\delta X)^2 \right\rangle_I = \kappa_2 = \left[\frac{\partial^2}{\partial t^2} \ln \left\langle \exp(tX) \right\rangle_I \right]_{t=0}. \tag{7.34b}$$

Demostración. Usamos las aproximaciones

$$\exp(x) \approx 1 + x + \frac{x^2}{2}, \tag{7.35}$$

$$\ln(1+x) \approx x - \frac{x^2}{2} \tag{7.36}$$

para $x \ll 1$, con las que obtenemos

$$\langle \exp(tX) \rangle_I \approx 1 + t\langle X \rangle_I + \frac{t^2}{2}\langle X^2 \rangle_I, \tag{7.37}$$

y luego

$$\begin{aligned} K_I(t) &\approx \ln\left(1 + t\langle X \rangle_I + \frac{t^2}{2}\langle X^2 \rangle_I\right) \\ &\approx t\langle X \rangle_I + \frac{t^2}{2}\langle X^2 \rangle_I - \frac{1}{2}\left(t\langle X \rangle_I + \frac{t^2}{2}\langle X^2 \rangle_I\right)^2 \\ &= t\langle X \rangle_I + \frac{t^2}{2}\langle (\delta X)^2 \rangle_I + \mathcal{O}(t^3), \end{aligned} \tag{7.38}$$

que comparando con (7.33) nos entrega (7.34)

7.3.2 Función generadora de probabilidad

Recordando la definición en (3.198) de la función generadora $F(t)$ de una secuencia $f_1, f_2, f_3, \ldots$ como

$$F(t) := \sum_{n=0}^{\infty} f_n t^n, \tag{7.39}$$

ahora interpretaremos las probabilidades $p_i := P(X = X_i|I)$ asociadas a una variable discreta como una secuencia, y definiremos la *función generadora de probabilidad* como sigue.

Definición 7.6 — Función generadora de probabilidad
Para una variable $X \sim I \in \{x_1, \ldots, x_n\}$ con probabilidades

$$p_i := P(X = x_i|I)$$

definiremos la función generadora de probabilidad como

$$G(t; I) := \sum_{k=1}^{n} p_k t^k. \tag{7.40}$$

Notemos que $G(t; I)$ puede ser escrita como una expectación si definimos

$$\text{ind}(X) \in \{1, 2, \ldots, n\}$$

como el índice entero que corresponde al valor de la variable X. Formalmente, podemos escribir

$$\mathrm{ind}(X) := \sum_{k=1}^{n} \mathrm{Q}(X = x_k)k, \tag{7.41}$$

y entonces tenemos

$$G(t; I) = \left\langle t^{\,\mathrm{ind}(X)} \right\rangle_I. \tag{7.42}$$

Rápidamente podemos notar que

$$G(1; I) = \left\langle 1 \right\rangle_I = 1, \tag{7.43}$$

ya que la distribución está correctamente normalizada. La principal utilidad de la función generadora de probabilidad es que, si es posible calcularla de forma cerrada, entonces los momentos de $\mathrm{ind}(X)$ pueden obtenerse a partir de las derivadas de G. Llamando $K := \mathrm{ind}(X)$ por simplicidad de notación, tenemos que la primera derivada de G respecto a su argumento entrega

$$\left(\frac{\partial G}{\partial t}\right)_{t=1} = \left(\left\langle K t^{K-1} \right\rangle_I\right)_{t=1} = \left\langle K \right\rangle_I, \tag{7.44}$$

mientras que la segunda derivada nos da

$$\left(\frac{\partial^2 G}{\partial t^2}\right)_{t=1} = \left(\left\langle K(K-1) t^{K-2} \right\rangle_I\right)_{t=1} = \left\langle K^2 \right\rangle_I - \left\langle K \right\rangle_I, \tag{7.45}$$

y podemos rearmar la varianza de K como

$$\left\langle (\delta K)^2 \right\rangle_I = \left\langle K^2 \right\rangle_I - \left\langle K \right\rangle_I^2 = \left(\frac{\partial^2 G}{\partial t^2} + \frac{\partial G}{\partial t} - \left[\frac{\partial G}{\partial t}\right]^2\right)_{t=1}. \tag{7.46}$$

Ejemplo 7.3.1. *La función generadora asociada a la distribución de Poisson es*

$$G(t; \lambda) = \sum_{k=0}^{\infty} P(k|\lambda) t^k = \sum_{k=0}^{\infty} \frac{\exp(-\lambda)(\lambda t)^k}{k!} = \exp(-\lambda(1-t)). \tag{7.47}$$

Luego, la expectación puede calcularse directamente como

$$\left\langle k \right\rangle_\lambda = \left(\frac{\partial G}{\partial t}\right)_{t=1} = \Big[\lambda \exp(-\lambda(1-t))\Big]_{t=1} = \lambda, \tag{7.48}$$

y la varianza es

$$\begin{aligned} \left\langle (\delta k)^2 \right\rangle_\lambda &= \left(\frac{\partial^2 G}{\partial t^2} + \frac{\partial G}{\partial t} - \left[\frac{\partial G}{\partial t}\right]^2\right)_{t=1} \\ &= \Big(\lambda^2 \exp(-\lambda(1-t)) + \lambda \exp(-\lambda(1-t)) - \lambda^2 \exp(-2\lambda(1-t))\Big)_{t=1} \\ &= \Big(\lambda \exp(-\lambda(1-t))\Big)_{t=1} = \lambda. \end{aligned} \tag{7.49}$$

7.4 — Parámetros de escala, posición y forma

Es posible clasificar los parámetros de un modelo en tres categorías: parámetros de escala, parámetros de posición y parámetros de forma.

Consideremos una variable $z \sim M(\boldsymbol{\theta})$ con distribución $P(z|\boldsymbol{\theta}) = \rho(z;\boldsymbol{\theta})$. Si a partir de z construimos una nueva variable reescalándola por un factor positivo s, esto es,

$$x := sz, \qquad s > 0,$$

entonces la distribución de x estará dada por

$$\begin{aligned} P(x|s,\boldsymbol{\theta}) = \langle \delta(x - sz) \rangle_{\boldsymbol{\theta}} &= \frac{1}{s} \langle \delta(z - (x/s)) \rangle_{\boldsymbol{\theta}} \\ &= \frac{1}{s} P(z = x/s|\boldsymbol{\theta}) = \frac{1}{s}\rho(x/s;\boldsymbol{\theta}). \end{aligned} \tag{7.50}$$

Esto nos lleva a definir el concepto de *parámetro de escala*.

Definición 7.7 — Parámetro de escala
Si un modelo es de la forma

$$P(X = x|s,\boldsymbol{\theta}) = \frac{1}{s}\rho(x/s;\boldsymbol{\theta}), \tag{7.51}$$

entonces s actúa como *parámetro de escala* para x.

Ejemplo 7.4.1. *En la distribución Gamma,*

$$P(x|k,\theta) = \frac{\exp(-x/\theta)x^{k-1}}{\Gamma(k)\theta^k}, \tag{7.52}$$

el parámetro θ es un parámetro de escala, ya que podemos escribir la distribución como

$$P(x|k,\theta) = \frac{1}{\theta}\left[\frac{1}{\Gamma(k)}\exp(-x/\theta)\left(\frac{x}{\theta}\right)^{k-1}\right] = \frac{1}{\theta}\rho(x/\theta). \tag{7.53}$$

Notemos la similitud entre el lado derecho de (7.51) y las funciones

$$\frac{1}{\varepsilon}\eta(x/\varepsilon)$$

que llevan a representaciones de la delta de Dirac según (3.36). Este hecho nos dice lo siguiente: cualquier distribución de probabilidad con un parámetro de escala tendiendo a cero se reduce a una distribución de conocimiento completo.

Ahora construyamos una nueva variable y desplazando z una cantidad r, es decir,

$$y := z + r.$$

La distribución de y es

$$\begin{aligned} P(y|r,\boldsymbol{\theta}) &= \langle \delta(y-[z+r]) \rangle_{\boldsymbol{\theta}} \\ &= \langle \delta(z-[y-r]) \rangle_{\boldsymbol{\theta}} \\ &= P(z=y-r|\boldsymbol{\theta}) \\ &= \rho(y-r;\boldsymbol{\theta}). \end{aligned} \tag{7.54}$$

Esto nos lleva a definir el concepto de *parámetro de posición*.

Definición 7.8 — Parámetro de posición
Si un modelo es de la forma

$$P(X=x|r,\boldsymbol{\theta}) = \rho(x-r;\boldsymbol{\theta}), \tag{7.55}$$

entonces r actúa como *parámetro de posición* para x.

Cualquier otro tipo de parámetro, que no sea de escala o posición, lo denominaremos *parámetro de forma*.

Notemos aquí una propiedad útil en términos de la existencia de parámetros de forma, expresada en el siguiente lema.

Lema 7.1. *Si una distribución $P(X|s,r,I)$ sólo tiene un parámetro de escala s y un parámetro de posición r, es decir, no tiene parámetros de forma, y luego la distribución puede escribirse como*

$$P(X=x|s,r,I) = \frac{1}{s}\rho\Big(\frac{x-r}{s}\Big), \tag{7.56}$$

entonces la variable reducida

$$Z := \frac{X-r}{s}$$

tiene una distribución universal

$$P(Z=z|s,r,I) = \rho(z). \tag{7.57}$$

La demostración es directa.

Demostración.

$$\begin{aligned} P(Z=z|s,r,I) &= \left\langle \delta\Big(z-\frac{X-r}{s}\Big) \right\rangle_{s,r,I} \\ &= \int_{-\infty}^{\infty} \mathrm{d}x \left[\frac{1}{s}\rho\Big(\frac{x-r}{s}\Big)\right] \delta\Big(z-\frac{x-r}{s}\Big) \qquad (7.58) \\ \text{(usando } z' := \frac{x-r}{s}\text{)} \quad &= \int_{-\infty}^{\infty} \mathrm{d}z' \rho(z')\delta(z-z') = \rho(z) \end{aligned}$$

7.5 — Independencia y correlación

Si X, Y son dos variables, entonces diremos que su *distribución conjunta* en el estado de conocimiento I es

$$P(x, y|I) = \langle \delta(X - x)\delta(Y - y)\rangle_I . \tag{7.59}$$

A partir de la distribución conjunta obtenemos las distribuciones de X y de Y usando la regla de la marginalización,

$$P(x|I) = \int dy P(x, y|I), \tag{7.60a}$$

$$P(y|I) = \int dx P(x, y|I). \tag{7.60b}$$

A estas distribuciones las llamaremos *distribuciones marginales*.

Si X e Y son independientes en el estado de conocimiento I, entonces se tiene

$$P(X|Y, I) = P(X|I), \tag{7.61}$$

es decir, el conocer Y no afecta la probabilidad que asignamos a X. De la misma manera, usando el teorema de Bayes vemos que

$$P(Y|X, I) = \frac{P(Y|I)\cancel{P(X|Y, I)}}{\cancel{P(X|I)}} = P(Y|I), \tag{7.62}$$

luego el conocer X tampoco afecta la probabilidad que asignamos a Y. En otras palabras, la independencia de dos cantidades es mutua. En ese caso, la distribución conjunta es simplemente el producto de las distribuciones marginales,

$$P(X, Y|I) = P(X|I)P(Y|I). \tag{7.63}$$

Una medida del grado de dependencia o *correlación* es la *covarianza*, definida a continuación.

Definición 7.9 — Covarianza
Definiremos la covarianza entre dos variables X, Y como

$$\langle \delta X \delta Y\rangle_I := \langle XY\rangle_I - \langle X\rangle_I \langle Y\rangle_I . \tag{7.64}$$

Si X e Y son independientes, entonces su covarianza es cero, ya que

$$\begin{aligned}\langle XY\rangle_I &= \int \mathrm{d}x \int \mathrm{d}y P(x, y|I)xy = \int \mathrm{d}x \int \mathrm{d}y P(x|I)P(y|I)xy \\ &= \left[\int \mathrm{d}x P(x|I)x\right]\left[\int \mathrm{d}y P(y|I)y\right] \\ &= \langle X\rangle_I \langle Y\rangle_I ,\end{aligned} \tag{7.65}$$

sin embargo, covarianza cero no implica independencia, como se ve en el siguiente ejemplo.

Ejemplo 7.5.1. *Si $\phi \sim U(0, 2\pi)$, entonces las variables*

$$X := \cos(\phi), \tag{7.66a}$$
$$Y := \sin(\phi), \tag{7.66b}$$

a pesar de ser obviamente correlacionadas ya que $X^2 + Y^2 = 1$, tienen covarianza cero puesto que

$$\langle XY \rangle_{I_0} = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\phi \cos(\phi) \sin(\phi) = 0 \tag{7.67}$$

y a la vez

$$\langle X \rangle_{I_0} = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\phi \cos(\phi) = 0, \tag{7.68a}$$
$$\langle Y \rangle_{I_0} = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\phi \sin(\phi) = 0. \tag{7.68b}$$

Para un conjunto de n variables, la covarianza se generaliza a la *matriz de covarianza* $\mathbf{\Sigma}$, que tiene componentes

$$\Sigma_{ij} := \left\langle \delta X_i \delta X_j \right\rangle_I, \tag{7.69}$$

tal que los elementos de la diagonal Σ_{ii} corresponden a las varianzas $\langle (\delta X_i)^2 \rangle_I$.

7.6 — Transformación de distribuciones

En este punto tenemos todas las herramientas para atacar el problema de cómo propagar en general la información que poseemos acerca de una variable X hacia información respecto a una nueva variable Z, conectada con X a través de un mapa o función tal que $Z = f(X)$. Esto nos permitirá además construir nuevas distribuciones a partir de las ya existentes.

Comencemos analizando el caso de una variable discreta $X \in \{x_1, \ldots, x_n\}$ tal que $X \sim I$, donde $f(X)$ es una función arbitraria de X. ¿Cuál es el modelo que corresponde a la nueva variable f? Nuestra respuesta inmediata, dado que $f = F$ es una proposición válida y recordando la conexión entre probabilidades y funciones indicador, es que

$$P(f = F|I) = \langle \mathrm{Q}(f = F) \rangle_I . \tag{7.70}$$

De todas maneras, podemos demostrar esto más formalmente.

Demostración. Usando la regla de la marginalización, podemos descomponer la probabilidad en el lado izquierdo como

$$\begin{aligned} P(f = F|I) &= \sum_{i=1}^{n} P(f = F, X = x_i|I) \\ &= \sum_{i=1}^{n} P(f = F|X = x_i, I)P(X = x_i|I). \end{aligned} \tag{7.71}$$

Sin embargo, dado que el conocer el valor de X fija automáticamente el valor de $f(X)$, tenemos que la probabilidad de f dado X debe ser una distribución de conocimiento completo (**Sección 6.8**), por lo que se obtiene

$$P(f = F|X = x_i, I) = \mathrm{Q}(f(x_i) = F) \tag{7.72}$$

para todo I. Aplicando la regla de marginalización, tenemos finalmente

$$P(f = F|I) = \sum_{i=1}^{n} \mathrm{Q}(f(x_i) = F)P(X = x_i|I) = \langle \mathrm{Q}(f = F) \rangle_I \tag{7.73}$$

En la práctica podemos utilizar (7.70) escribiéndola de la forma

$$\boxed{P(f = F|I) = \sum_{j=1}^{m} P(X = x_j^*|I),} \tag{7.74}$$

donde la suma recorre los m valores $\{x_1^*, \ldots, x_m^*\}$ tal que $f(x_j^*) = F$. Esto se puede entender más intuitivamente ya que la proposición $f = F$ es equivalente a

$$(X = x_1^*) \vee (X = x_2^*) \vee \ldots \vee (X = x_m^*), \tag{7.75}$$

esto es, equivalente a decir que X está en el conjunto $\{x_1^*, \ldots, x_m^*\}$. Aplicando función indicador a ambos lados y recordando que las proposiciones $X = x_m^*$ son mutuamente excluyentes, tenemos

$$\mathrm{Q}(f = F) = \sum_{j=1}^{m} \mathrm{Q}(X = x_j^*), \tag{7.76}$$

y finalmente aplicando expectación bajo I,

$$P(f = F|I) = \sum_{j=1}^{m} P(X = x_j^*|I),$$

con lo que se recupera (7.74).

Ejemplo 7.6.1. *Sea n una variable discreta $n \in \{-2, -1, 0, 1, 2\}$ con distribución*

$$P(n|I) = \frac{p_0}{1 + n^2}, \tag{7.77}$$

y sea $f(n) = 2n(n-1)$. ¿Qué distribución sigue f?

Solución. *Las soluciones de $f(n) = F$, esto es, $2n(n-1) = F$ son dos,*

$$n_1^* = \frac{1}{2}\left(1 + \sqrt{1+2F}\right), \tag{7.78a}$$

$$n_2^* = \frac{1}{2}\left(1 - \sqrt{1+2F}\right), \tag{7.78b}$$

luego tenemos, para todo F en el recorrido de $f(n)$,

$$\begin{aligned} P(f = F|I) &= P(n_1^*|I) + P(n_2^*|I) \\ &= \frac{p_0}{1 + \frac{1}{2}(1 + F + \sqrt{1+2F})} + \frac{p_0}{1 + \frac{1}{2}(1 + F - \sqrt{1+2F})}. \end{aligned} \tag{7.79}$$

Para una variable continua $X \in [a, b]$ tal que $X \sim I$ la situación es similar. Recordando que toda densidad de probabilidad es la expectación de una delta de Dirac, proponemos

$$P(f = F|I) = \langle \delta(f - F) \rangle_I, \tag{7.80}$$

y la demostración sigue las mismas líneas.

Demostración. Usando la regla de la marginalización, escribimos $P(f = F|I)$ como

$$\begin{aligned} P(f = F|I) &= \int_{-\infty}^{\infty} \mathrm{d}x P(f = F, X = x|I) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x P(f = F|X = x, I) P(X = x|I). \end{aligned} \tag{7.81}$$

Nuevamente, como conocer el valor de X fija automáticamente el valor de $f(X)$, se cumple

$$P(f = F|X = x, I) = \delta(f(x) - F), \tag{7.82}$$

y reemplazando tenemos

$$\begin{aligned} P(f = F|I) &= \int_{-\infty}^{\infty} \mathrm{d}x \delta(f(x) - F) P(X = x|I) \\ &= \langle \delta(f - F) \rangle_I \end{aligned} \tag{7.83}$$

El equivalente continuo de (7.74) se obtiene usando la propiedad (3.48) de composición de la delta de Dirac, quedando

$$P(f = F|I) = \langle \delta(f - F) \rangle_I = \left\langle \sum_{x^*} \frac{\delta(X - x^*)}{|f'(X)|} \right\rangle_I = \sum_{x^*} \frac{\langle \delta(X - x^*) \rangle_I}{|f'(x^*)|}, \tag{7.84}$$

pero $\langle \delta(X - x^*) \rangle_I = P(X = x^*|I)$ entonces se tiene finalmente

$$\boxed{P(f = F|I) = \sum_{x^*} \frac{P(X = x^*|I)}{|f'(x^*)|}.} \tag{7.85}$$

Notemos la similitud con el caso discreto en (7.74), la única diferencia es el factor $1/|f'(x^*)|$. En el caso particular en que la función $f(X)$ es invertible, la solución de $f(x^*) = F$ es única y está dada por la función inversa

$$x^* = f^{-1}(F) = X(F),$$

por lo tanto (7.85) se reduce a

$$P(f = F|I) = \frac{P(X = X(F)|I)}{|f'(X(F))|} = P(X = X(F)|I) \left|\frac{\mathrm{d}X(F)}{\mathrm{d}F}\right|, \tag{7.86}$$

donde la última igualdad viene dada por el *teorema de la función inversa*,

$$\frac{\mathrm{d}}{\mathrm{d}F} \underbrace{f(X(F))}_{=F} = 1 = f'(X(F)) \cdot \frac{\mathrm{d}}{\mathrm{d}F} X(F). \tag{7.87}$$

Ejemplo 7.6.2. *Consideremos la función $f(X) = a + (X - r)^2$ para $X \geq 0$ con*

$$P(X = x|I) = \frac{1}{(x+1)^2}.$$

Las soluciones de $f(X) = F$ son

$$x^* = r \pm \sqrt{F - a}, \tag{7.88}$$

y la derivada es $f'(x^*) = 2(x^* - r) = \pm 2\sqrt{F-a}$, luego $|f'(x^*)| = 2\sqrt{F-a}$ y tenemos

$$\begin{aligned} P(f = F|I) &= \sum_{x^*} \frac{P(X = x^*|I)}{|f'(x^*)|} \\ &= \frac{1}{2\sqrt{F-a}} \left(P(X = r - \sqrt{F-a}|I) + P(X = r + \sqrt{F-a}|I) \right) \\ &= \frac{1}{2\sqrt{F-a}} \left(\frac{\Theta(r - \sqrt{F-a})}{(r - \sqrt{F-a} + 1)^2} + \frac{1}{(r + \sqrt{F-a} + 1)^2} \right), \end{aligned} \tag{7.89}$$

donde el primer término del lado derecho lleva la función escalón para fijar $x^* \geq 0$.

PROBLEMAS

Problema 7.1. *Determine la distribución de $Z = X + Y$, donde $X \sim Exp(\lambda)$ y $Y \sim Gamma(k, \theta)$.*

Problema 7.2. *Considere la* distribución triangular *para una variable continua $X \in [0, 1]$, definida de acuerdo a*

$$P(X = x|m, h) = \begin{cases} \alpha x \ \ si\ x \leq m, \\ \\ b - \beta x \ \ si\ x > m \end{cases} \tag{7.90}$$

con $\alpha > 0$, $\beta > 0$, $b > 0$ y tal que $P(X = m|m, h) = h$. Considerando que la distribución es continua en $x = m$ (no así su derivada), determine α, β, b y h en función de m. Exprese la media, moda y varianza también en términos de m.

Problema 7.3. *Demuestre que si s es un parámetro de escala para la variable x, se cumple*

$$\langle g(x/s)\rangle_{s,\boldsymbol{\theta}} = G(\boldsymbol{\theta}) \tag{7.91}$$

para toda función g.

Problema 7.4. *Demuestre que si, para dos variables X e Y, se cumple*

$$\langle \delta\omega\delta\sigma\rangle_I = 0 \tag{7.92}$$

para cualquier par de funciones $\omega(X)$ y $\sigma(Y)$, entonces las variables X e Y son independientes.

Problema 7.5. *Demuestre que $\langle X\rangle_I = x_0$ para una variable real $X \sim I$ con densidad de probabilidad*

$$P(X = x|I) = f(|x - x_0|). \tag{7.93}$$

Problema 7.6. *Dos variables $x \in [0, \infty)$, $y \in [0, \infty)$ son descritas por un modelo*

$$P(x, y|\mu) = \frac{1}{Z(\mu)} \exp(-\mu(x + y))(x + y)^4. \tag{7.94}$$

(a) Calcule la constante de normalización $Z(\mu)$.

(b) Calcule la covarianza $\langle \delta x \delta y\rangle_\mu$ entre ellas.

(c) Calcule las distribuciones marginales $P(x|\mu)$ y $P(y|\mu)$.

Problema 7.7. *Para un modelo*

$$P(x, y|\alpha, k) = 2\alpha k \exp(-\alpha(x + k \cdot y)^2), \tag{7.95}$$

con $x \geq 0$, $y \geq 0$, $\alpha > 0$ y $k \geq 1$, calcule

(a) la probabilidad $P(y > x|\alpha, k)$ de que y sea mayor que x,

(b) la probabilidad $P(x < 1|\alpha, k)$ de que x sea menor que 1.

Exprese sus resultado sólo en términos de los parámetros α y k.

Problema 7.8. *Calcule los momentos $\langle x^m \rangle_{k,\theta}$ de una variable $x \sim Gamma(k, \theta)$ para $m = 0, 1, 2, 3, \ldots$. Su resultado debe quedar expresado sólo en términos de k y θ.*

Problema 7.9. *Si $x \sim Gamma(k, \theta)$, calcule la probabilidad de que x sea menor o igual a la moda de la distribución.*

Problema 7.10. *Determine la distribución de $x = \cos\phi$ si $\phi \sim U(-\pi, \pi)$.*

Problema 7.11. *Una máquina pinta cuadrados al azar con áreas A que siguen una distribución exponencial, $A \sim Exp(\lambda)$ con λ = 10 cm^{-2}. ¿Qué distribución sigue el lado $L = \sqrt{A}$ de cada cuadrado?*

Problema 7.12. *Si la variable $x \in [0, 1]$ tiene densidad de probabilidad*

$$P(x|\lambda) = K\Theta(x)\Theta(1 - x)\exp(-\lambda x), \tag{7.96}$$

encuentre la densidad de probabilidad para $g(x) = \sqrt{1 - x^2}$ y el valor de K.

Problema 7.13. *Demuestre que la distribución de Weibull,*

$$P(X|k, \lambda) = \frac{k}{\lambda}\left(\frac{X}{\lambda}\right)^k \exp(-(X/\lambda)^k) \tag{7.97}$$

es la distribución de una variable X cuando $Z = (X/\lambda)^k$ es tal que $Z \sim Exp(1)$.

CAPÍTULO 8

La distribución normal

Probablemente el modelo más conocido en toda la teoría de la probabilidad y estadística es la *distribución normal* o gaussiana, coloquialmente conocida como la *campana de Gauss*. Este es un modelo de dos parámetros para una variable real X, con densidad de probabilidad como se ve a continuación.

Definición 8.1 — Distribución normal
Una variable $X \in \mathbb{R}$ sigue una distribución normal o gaussiana si

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big), \tag{8.1}$$

Diremos que $X \sim \mathcal{N}(\mu, \sigma^2)$.

Un aspecto muy cómodo de la distribución normal es que sus parámetros son directamente la media y la varianza de la distribución, como se muestra en el siguiente recuadro.

Recuadro 8.1 — Media y varianza de la distribución normal

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ entonces la media y la varianza de X están dadas por

$$\big\langle X \big\rangle_{\mu,\sigma} = \mu, \tag{8.2a}$$
$$\big\langle (\delta X)^2 \big\rangle_{\mu,\sigma} = \sigma^2. \tag{8.2b}$$

Notemos además que el parámetro μ es un parámetro de posición, mientras que el parámetro σ es un parámetro de escala, es decir no existen parámetros de forma.

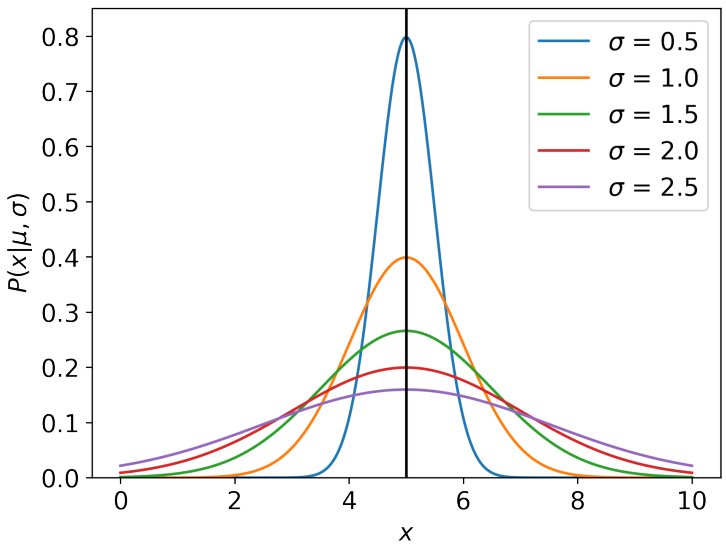


Figura 8.1: La distribución normal para μ =5 (línea negra vertical) y valores de σ entre 0.5 y 2.5.

La distribución normal se muestra para distintos valores de su varianza σ en la **Figura 8.1**.

8.1 — La distribución normal como aproximación

Supongamos un modelo $P(X = x|I) = p(x)$ para la variable continua X donde su densidad de probabilidad $p(x)$ está muy concentrada en torno al valor x_0. En ese caso podemos aproximar $\ln p(x)$ en torno a x_0 hasta segundo orden como

$$\ln p(x) \approx \ln p(x_0) + (x - x_0)\frac{\partial}{\partial x_0}\ln p(x_0) + \frac{1}{2}(x - x_0)^2\frac{\partial^2}{\partial x_0^2}\ln p(x_0) \tag{8.3}$$

Si x_0 es un máximo de $p(x)$, entonces

$$\frac{\partial}{\partial x_0}\ln p(x_0) = 0 \tag{8.4}$$

y se tiene

$$\ln p(x) \approx \ln p(x_0) + \frac{1}{2}(x - x_0)^2\frac{\partial^2}{\partial x_0^2}\ln p(x_0). \tag{8.5}$$

Definiendo

$$\sigma^2 := \left(-\frac{\partial^2}{\partial x_0^2}\ln p(x_0)\right)^{-1} \tag{8.6}$$

tenemos

$$p(x) \propto \exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right) \tag{8.7}$$

que podemos normalizar usando la integral gaussiana como

$$\int_{-\infty}^{\infty} dx \exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right) = \sqrt{2\pi}\sigma, \tag{8.8}$$

llegando de esta forma a la distribución normal según (8.1).

Para calcular la media y varianza de un modelo normal, es útil recordar el **Lema 7.1** que nos permite expresar una variable normal en términos de la variable reducida

$$Z := \frac{X-\mu}{\sigma}, \tag{8.9}$$

conocida en inglés como el *z-score*, de tal forma que

$$P(X=x|\mu,\sigma) = \frac{\exp(-z^2/2)}{\sqrt{2\pi}\sigma} = \frac{\phi(z)}{\sigma}, \tag{8.10}$$

donde

$$\phi(z) = \frac{\exp(-z^2/2)}{\sqrt{2\pi}}$$

es la función gaussiana que vimos ya en (3.40). Según el mismo **Lema 7.1** se tiene que $Z \sim \mathcal{N}(0,1)$, esto es, la densidad de probabilidad del *z-score* es universal. La media de esta distribución universal es cero por simetría, ya que

$$\begin{aligned}
\langle z\rangle_I &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}z\, \exp(-z^2/2)z \\
&= \frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{0} \mathrm{d}z\, \exp(-z^2/2)z + \int_{0}^{\infty} \mathrm{d}z\, \exp(-z^2/2)z\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[-\int_{0}^{\infty} \mathrm{d}z\, \exp(-z^2/2)z + \int_{0}^{\infty} \mathrm{d}z\, \exp(-z^2/2)z\right] = 0,
\end{aligned} \tag{8.11}$$

mientras que la varianza puede calcularse de

$$\begin{aligned}
\langle(\delta z)^2\rangle_I = \langle z^2\rangle_I - \overbrace{\langle z\rangle_I^2}^{0} &= \langle z^2\rangle_I \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{d}z\, \exp(-z^2/2)z^2 \\
&= \frac{1}{\sqrt{2\pi}}\left[-2\frac{\partial}{\partial\alpha}\int_{-\infty}^{\infty} \mathrm{d}z\, \exp(-\alpha z^2/2)\right]_{\alpha=1} \\
\text{usando (12)} \quad &= -\frac{2}{\sqrt{2\pi}}\left(\frac{\partial}{\partial\alpha}\sqrt{\frac{2\pi}{\alpha}}\right)_{\alpha=1} = \left(\alpha^{-\frac{3}{2}}\right)_{\alpha=1} = 1,
\end{aligned} \tag{8.12}$$

respectivamente, luego el *z-score* tiene media igual a 0 y varianza igual a 1.

Con esto podemos calcular la media de cualquier variable X distribuida normal como

$$\langle x\rangle_{\mu,\sigma^2} = \langle \sigma z + \mu\rangle_{\mu,\sigma^2} = \sigma\overbrace{\langle z\rangle_I}^{0} + \mu = \mu,$$

y a la vez la varianza de X como

$$\begin{aligned}
\langle(\delta x)^2\rangle_{\mu,\sigma^2} &= \langle x^2\rangle_{\mu,\sigma^2} - \mu^2 \\
&= \langle(\sigma z+\mu)^2\rangle_{\mu,\sigma^2} - \mu^2 \\
&= \langle \sigma^2 z^2 + \mu^2 + 2\sigma\mu z\rangle_{\mu,\sigma^2} - \mu^2 \\
&= \sigma^2\overbrace{\langle z^2\rangle_I}^{1} + 2\sigma\mu\overbrace{\langle z\rangle_I}^{0} = \sigma^2,
\end{aligned}$$

confirmando el resultado en (8.2).

8.2 — La distribución normal multivariable

En más dimensiones, como la expansión en Taylor a segundo orden es

$$\ln p(\boldsymbol{x}) \approx \ln p(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)\cdot\nabla \ln p(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\intercal \mathbb{H}(\boldsymbol{x}-\boldsymbol{x}_0) \tag{8.13}$$

con $\mathbb{H}$ la *matriz hessiana*, se tiene que si repetimos el mismo procedimiento anterior, el equivalente a la distribución normal es la *distribución normal multivariable*.

Definición 8.2 — Distribución normal multivariable
La distribución normal multivariable está dada por

$$P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \tag{8.14}$$

para $\boldsymbol{x} = (x_1, \ldots, x_d)$. Aquí el vector $\boldsymbol{\mu}$ contiene la media de cada componente, esto es

$$\mu_i = \langle X_i \rangle_{\boldsymbol{\mu},\boldsymbol{\Sigma}}, \tag{8.15}$$

mientras que la matriz $\boldsymbol{\Sigma} = -\mathbb{H}^{-1}$ es la *matriz de covarianza* (7.69),

$$\Sigma_{ij} = \langle \delta X_i \delta X_j \rangle_{\boldsymbol{\mu},\boldsymbol{\Sigma}}. \tag{8.16}$$

En el caso de dos dimensiones (d=2 en (8.14)) se obtiene la distribución normal bivariante,

$$P(X,Y|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2(1-\rho_{XY}^2)}\left[Z_X^2 + Z_Y^2 - 2\rho_{XY}Z_X Z_Y\right]\right)}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}}, \tag{8.17}$$

donde $\boldsymbol{\mu} = (\mu_X, \mu_Y)$, Z_X y Z_Y son los z-scores asociados a X e Y,

$$Z_X := \frac{X-\mu_X}{\sigma_X}, \tag{8.18a}$$

$$Z_Y := \frac{Y-\mu_Y}{\sigma_Y}, \tag{8.18b}$$

las componentes de la matriz de covarianza son

$$\Sigma_{11} = \sigma_X^2, \tag{8.19a}$$

$$\Sigma_{22} = \sigma_Y^2, \tag{8.19b}$$

$$\Sigma_{12} = \Sigma_{21} = \rho_{XY}\sigma_X\sigma_Y, \tag{8.19c}$$

y ρ_{XY} es el denominado coeficiente de correlación de Pearson entre X e Y, definido a continuación.

Definición 8.3 — Coeficiente de correlación de Pearson

$$\rho_{XY} := \frac{\langle XY \rangle_{\boldsymbol{\mu},\boldsymbol{\Sigma}} - \langle X \rangle_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \langle Y \rangle_{\boldsymbol{\mu}\boldsymbol{\Sigma}}}{\sigma_X \, \sigma_Y} \tag{8.20}$$

8.3 — Suma de variables

Vimos cómo toda distribución de probabilidad continua con un *peak* suficientemente agudo, es decir, muy concentrada en torno a un valor, puede aproximarse por una distribución normal. Ahora estudiaremos otra propiedad, muy relacionada, que hace que la distribución normal sea tan ubicua en distintos contextos: la distribución normal es el «atractor» al que tiende la distribución de la suma de variables independientes con varianza finita.

Este es uno de los resultados más conocidos y confiables de la teoría de la probabilidad, conocido como el *teorema central del límite*[(1)], y que hace que el problema de estimar una suma de muchas cantidades desconocidas sea relativamente sencillo: el modelo normal será en la gran mayoría de las veces la respuesta correcta. Para entender esto, deberemos primero descubrir cómo calcular la distribución asociada a la suma de variables, usando la idea de convolución y el concepto de *función característica*.

En primer lugar, demostraremos dos resultados sencillos para variables independientes y que siguen la misma distribución. Supongamos dos variables X_1 e X_2 que cumplen estas condiciones, y que tienen media y varianza dadas por

$$\langle X_i \rangle_I = x_0, \tag{8.21a}$$

$$\langle (\delta X_i)^2 \rangle_I = s^2, \tag{8.21b}$$

con i=1, 2. Con estas definiciones es inmediato ver que la media de la suma cumple

$$\langle X_1 + X_2 \rangle_I = \langle X_1 \rangle_I + \langle X_2 \rangle_I = 2x_0, \tag{8.22}$$

mientras que para la varianza se cumple

$$\begin{aligned} \left\langle \left(\delta[X_1 + X_2]\right)^2 \right\rangle_I &= \langle (X_1 + X_2)^2 \rangle_I - \langle X_1 + X_2 \rangle_I^2 \\ &= \langle {X_1}^2 \rangle_I + 2\langle X_1 \rangle_I \langle X_2 \rangle_I + \langle {X_2}^2 \rangle_I - \left(2\langle X_1 \rangle_I\right)^2 \\ &= \langle {X_1}^2 \rangle_I + 2\,x_0^2 + \langle {X_2}^2 \rangle_I - (2\,x_0)^2 \\ &= \langle {X_1}^2 \rangle_I - 2\,x_0^2 + \langle {X_2}^2 \rangle_I \\ &= \langle (\delta X_1)^2 \rangle_I + \langle (\delta X_2)^2 \rangle_I, \\ &= 2\,s^2, \end{aligned} \tag{8.23}$$

es decir, la varianza de la suma es el doble de la varianza de cualquiera de las variables.

(1) Como lo hace notar Jaynes (2003, pág. 242), en alemán *den zentralen Grenzwertsatz* no es el «teorema del límite central» sino el *teorema central* del límite: es decir, ¡es el teorema el que es central, no el límite!

Por inducción matemática es sencillo probar que, si

$$Z := \sum_{i=1}^{n} X_i \tag{8.24}$$

para $n \geq 1$ variables independientes $\{X_1, \ldots, X_n\}$, cada una con media x_0 y varianza s^2, se cumple

$$\langle Z \rangle_I = n\, x_0, \tag{8.25a}$$
$$\langle (\delta Z)^2 \rangle_I = n\, s^2. \tag{8.25b}$$

Aunque de seguro ya le es familiar el concepto, aquí introduciremos por primera vez la definición de promedio aritmético.

Definición 8.4 — Promedio aritmético
Para un conjunto de n valores $(x_1, \ldots, x_n)$ se define el promedio aritmético de ellos como

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{8.26}$$

Si tomamos ahora el promedio aritmético de las variables $X_1, \ldots, X_n$,

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{Z}{n}, \tag{8.27}$$

vemos que

$$\langle \overline{X}_n \rangle_I = \frac{\langle Z \rangle_I}{n} = x_0, \tag{8.28a}$$
$$\langle (\delta \overline{X}_n)^2 \rangle_I = \frac{\langle (\delta Z)^2 \rangle_I}{n^2} = \frac{s^2}{n}, \tag{8.28b}$$

donde hemos usado la propiedad

$$\langle (\delta[\alpha Z])^2 \rangle_I = \langle (\alpha Z)^2 \rangle_I - \langle \alpha Z \rangle_I^2 = \alpha^2 \langle (\delta Z)^2 \rangle_I \tag{8.29}$$

para cualquier constante α. De (8.28) vemos que, en el límite $n \to \infty$, el promedio aritmético se vuelve exactamente igual a la media x_0 de una de las variables dado que la varianza tiende a cero al ser s^2 finito. Esto es, el límite $n \to \infty$ de (8.28) lleva a

$$\lim_{n \to \infty} P(\overline{X}_n = x | I) = \delta(x - x_0), \tag{8.30}$$

conocida como la *ley de los grandes números*, que enunciaremos a continuación.

Teorema 8.1 — Ley de los grandes números

Si $\{X_1, X_2, X_3, \ldots\}$ es un conjunto de variables *idénticamente distribuidas e independientes*, tal que $X \sim I$, entonces el promedio aritmético de un gran número de dichas variables cumple

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \langle X \rangle_I. \tag{8.31}$$

Esto nos dice que el promedio de muchas variables pierde su incerteza a medida que crece sin límite el número de ellas.

Volveremos a discutir la ley de los grandes números en una nueva interpretación en la **Sección 9.3**. Dado que tenemos este resultado asintótico, ahora nos preguntamos, ¿cuál es la distribución que sigue $\overline{X}_n$ para n finito? En otras palabras, ¿a través de qué distribución nos aproximamos a la delta de Dirac en el límite $n \to \infty$?

8.4 — Funciones características

Para abordar este problema consideremos ahora el caso más general de dos variables X e Y independientes, pero con distribuciones distintas

$$P(X = x|I) = p(x), \tag{8.32}$$
$$P(Y = y|I) = q(y), \tag{8.33}$$

y veamos cuál es la distribución $\rho(z) := P(X + Y = z|I)$ que sigue la suma $Z = X + Y$.
Ésta estará dada por

$$\begin{aligned}
\rho(z) &= P(X + Y = z|I) \\
&= \Big\langle \delta(X + Y - z) \Big\rangle_I \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y P(x, y|I)\delta(x + y - z) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x P(x|I) \int_{-\infty}^{\infty} \mathrm{d}y P(y|I)\delta(y - [z - x]) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x p(x)q(z - x),
\end{aligned} \tag{8.34}$$

que no es otra cosa que la convolución de las funciones p y q,

$$\rho(z) = (p * q)(z) = \int_{-\infty}^{\infty} \mathrm{d}t p(t)q(z - t). \tag{8.35}$$

como definimos en la **Sección 3.11**. Claramente este tipo de integrales no serán sencillas de evaluar excepto para formas muy particulares de las distribuciones p y q, sin embargo definiendo una transformada integral lineal (y su inversa) es posible usar el **Teorema de convolución** para hacerlo.

Usando la definición de transformada integral $T[f]$ en (3.215), su inversa $T^{-1}[f]$ según (3.221) y los *kernels* en (3.225), definiremos el concepto de *función característica*.

Definición 8.5 — Función característica
Para una variable $X \sim I$ definiremos la función característica de X que denotaremos por $\varphi_X(t)$, como

$$\varphi_X(t) := \langle \exp(\mathrm{i}tX) \rangle_I. \tag{8.36}$$

Si X es continua, $\varphi_X(t)$ es una transformada integral de $P(X|I)$,

$$\varphi_X(t) = \int_{-\infty}^{\infty} \mathrm{d}x P(X = x|I) \exp(\mathrm{i}tx), \tag{8.37}$$

con inversa

$$P(X = x|I) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}t \varphi_X(t) \exp(-\mathrm{i}tx). \tag{8.38}$$

La principal utilidad de la función característica es que para un conjunto de n variables idénticamente distribuidas $X_1, \ldots, X_n$ con suma

$$Z = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i, \tag{8.39}$$

se cumple que la función característica de la suma es

$$\boxed{\varphi_Z(t) = \varphi_X(t)^n.} \tag{8.40}$$

Demostración.

$$\begin{aligned} \varphi_Z(t) = \langle \exp(\mathrm{i}tZ) \rangle_I &= \left\langle \exp\left(\mathrm{i}t \sum_{i=1}^{n} X_i\right) \right\rangle_I \\ &= \left\langle \prod_{i=1}^{n} \exp(\mathrm{i}tX_i) \right\rangle_I \\ &= \prod_{i=1}^{n} \langle \exp(\mathrm{i}tX_i) \rangle_I \\ &= \langle \exp(\mathrm{i}tX) \rangle_I^n \\ &= \varphi_X(t)^n \end{aligned} \tag{8.41}$$

Este resultado es por supuesto el mismo **Teorema de convolución**, aplicado a la convolución de n distribuciones de probabilidad idénticas,

$$T[\underbrace{p * p * \ldots * p}_{n \text{ veces}}] = \prod_{i=1}^{n} T[p] = T[p]^n. \tag{8.42}$$

Ejemplo 8.4.1 (Función característica de la distribución normal)**.** *Calculemos* $\varphi_X(t)$ *para* $X \sim \mathcal{N}(\mu, \sigma^2)$*, donde la distribución es*

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2). \tag{8.43}$$

Calculamos directamente

$$\begin{aligned}\varphi_x(t) &= \int_{-\infty}^{\infty} \mathrm{d}x \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2}(x-\mu)^2\Big) \exp(\mathrm{i}tx) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \mathrm{d}x \exp\left(-\frac{1}{2\sigma^2}\left[x^2+\mu^2\right] + \left(\mathrm{i}t + \frac{\mu}{\sigma^2}\right)x\right) \\ &= \frac{\exp(-\mu^2/(2\sigma^2))}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \mathrm{d}x \exp\Big(-\frac{1}{2\sigma^2}\left(x^2 - 2x[\mathrm{i}t\sigma^2+\mu]\right)\Big),\end{aligned} \tag{8.44}$$

y completando el cuadrado, tenemos

$$\begin{aligned}x^2 - 2x(\mathrm{i}t\sigma^2+\mu) &= \Big(x - [\mathrm{i}t\sigma^2+\mu]\Big)^2 - (\mathrm{i}t\sigma^2+\mu)^2 \\ &= \Big(x - [\mathrm{i}t\sigma^2+\mu]\Big)^2 + t^2\sigma^4 - \mu^2 - 2\mathrm{i}t\mu\sigma^2,\end{aligned} \tag{8.45}$$

luego

$$\begin{aligned}\varphi_x(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\cancel{\frac{\mu^2}{2\sigma^2}} + \cancel{\frac{\mu^2}{2\sigma^2}} - \frac{t^2\sigma^2}{2} + \mathrm{i}t\mu\Big) \\ \times \int_{-\infty}^{\infty} \mathrm{d}x \exp\Big(-\frac{1}{2\sigma^2}\Big(x - [\mathrm{i}t\sigma^2+\mu]\Big)^2\Big) = \exp\Big(\mathrm{i}t\mu - \frac{t^2\sigma}{2}\Big).\end{aligned} \tag{8.46}$$

Este es un resultado de suma importancia que utilizaremos para llegar al teorema central del límite: la función característica de la distribución normal $\mathcal{N}(\mu, \sigma^2)$ es

$$\boxed{\varphi_x(t) = \exp\Big(\mathrm{i}t\mu - \frac{t^2\sigma^2}{2}\Big).} \tag{8.47}$$

Dos propiedades de las funciones características son directas de demostrar. En primer lugar,

$$\lim_{t\to 0} \varphi_X(t) = \lim_{t\to 0} \big\langle \exp(\mathrm{i}tX)\big\rangle_I = 1, \tag{8.48}$$

y por otro lado,

$$\begin{aligned}|\varphi_X(t)| &= \big|\big\langle \exp(\mathrm{i}tX)\big\rangle_I\big| \\ &= \sqrt{\big\langle \cos(tX)\big\rangle_I^2 + \big\langle \sin(tX)\big\rangle_I^2} \\ &= \sqrt{\Big\langle \underbrace{\cos(tX)^2 + \sin(tX)^2}_{=1} \Big\rangle_I - \big\langle(\delta\cos(tX))^2\big\rangle_I - \big\langle(\delta\sin(tX))^2\big\rangle_I} \\ &= \sqrt{1 - \big\langle(\delta\cos(tX))^2\big\rangle_I - \big\langle(\delta\sin(tX))^2\big\rangle_I} \;\leq 1.\end{aligned} \tag{8.49}$$

8.5 — El teorema central del límite

Ahora que tenemos en nuestro arsenal de herramientas el concepto de función característica podemos hacernos la siguiente pregunta. Como según (8.25) para dos variables idénticamente distribuidas e independientes X_1 y X_2, cada una con media x_0 y varianza s^2 se cumple

$$\langle X_1 + X_2 \rangle_I = 2x_0, \tag{8.50a}$$
$$\langle (\delta[X_1 + X_2])^2 \rangle_I = 2s^2. \tag{8.50b}$$

vemos que la distribución de la suma $X_1 + X_2$ nunca podrá coincidir exactamente con la distribución de las variables X_i, a menos que tanto la media como la varianza sean cero, pero ese es el caso trivial en que $X_1 = X_2 = 0$ y no hay incerteza. Sin embargo, suponiendo la media $x_0 = 0$ podemos construir la cantidad

$$X' := \frac{X_1 + X_2}{\sqrt{2}} \tag{8.51}$$

tal que

$$\langle X' \rangle_I = 0, \tag{8.52a}$$
$$\langle (\delta X')^2 \rangle_I = s^2, \tag{8.52b}$$

es decir, la misma media y varianza que cada una de las variables X_i, y en general para

$$\tilde{X} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

se cumple

$$\langle \tilde{X} \rangle_I = 0, \tag{8.53a}$$
$$\langle (\delta \tilde{X})^2 \rangle_I = s^2. \tag{8.53b}$$

La pregunta es, ¿existirá una distribución

$$p(x) := P(X_1 = x|I) = P(X_2 = x|I)$$

tal que a la vez coincida con la distribución de X'? Si suponemos que dicha distribución existe y tiene función característica $f(t) := \varphi_X(t)$, entonces debe cumplirse

$$\varphi_{X'}(t) = \varphi_{X_1+X_2}\left(\frac{t}{\sqrt{2}}\right) = \varphi_X\left(\frac{t}{\sqrt{2}}\right)^2 \tag{8.54}$$

y por lo tanto, llamando $u := t/\sqrt{2}$, vemos que f debe ser solución de la ecuación

$$f(u)^2 = f(u\sqrt{2}). \tag{8.55}$$

Aplicando logaritmo a ambos lados y derivando respecto a u, tenemos

$$\frac{2f'(u)}{f(u)} = \frac{\sqrt{2}f'(u\sqrt{2})}{f(u\sqrt{2})}$$
$$\hookrightarrow \frac{\sqrt{2}f'(u)}{f(u)} = \frac{f'(u\sqrt{2})}{f(u\sqrt{2})}, \tag{8.56}$$

esto es

$$\sqrt{2}g(u) = g(u\sqrt{2}) \tag{8.57}$$

con $g(u) := f'(u)/f(u) = \frac{\mathrm{d}}{\mathrm{d}u} \ln f(u)$, y si nuevamente derivamos respecto a u, vemos que

$$\cancel{\sqrt{2}}g'(u) = \cancel{\sqrt{2}}g'(u\sqrt{2}) \tag{8.58}$$

es decir, la función $g'(u)$ no depende de su argumento y es por tanto una constante, llamémosla α. Tenemos $g'(u) = \alpha$ e integrando,

$$g(u) = \frac{\mathrm{d}}{\mathrm{d}u} \ln f(u) = \alpha u + \beta \tag{8.59}$$

pero por (8.57) para $u = 0$ se tiene que $\sqrt{2}\beta = \beta$, luego $\beta = 0$ y al integrar (8.59) tenemos

$$f(u) = F_0 \exp\Big(\frac{\alpha u^2}{2}\Big). \tag{8.60}$$

Usando (8.48) fijamos $F_0 = 1$ y para cumplir (8.49) estamos obligados a fijar $\alpha < 0$, digamos $\alpha = -\sigma^2$ y tenemos finalmente

$$f(t) = \varphi_X(t) = \exp\Big(-\frac{t^2\sigma^2}{2}\Big), \tag{8.61}$$

de lo cual comparando con (8.47) se sigue que $X_i \sim \mathcal{N}(0, \sigma^2)$ y también

$$\frac{X_1 + X_2}{\sqrt{2}} \sim \mathcal{N}(0, \sigma^2).$$

Esto contesta nuestra pregunta inicial en términos afirmativos, y podemos refrasearla de la siguiente manera: la única distribución estable ante la operación

$$a, b \mapsto \frac{a+b}{\sqrt{2}}$$

es la distribución normal de media cero. Esto es, de existir una distribución universal a la cual converge la distribución de una suma de variables idénticas, esta debe ser la distribución normal, ya que $a + b \sim \mathcal{N}(0, 2\sigma^2)$.

Este resultado se conoce como el teorema central del límite, que enunciaremos más precisamente a continuación.

Teorema 8.2 — Teorema central del límite
Sean n variables $\{X_1, X_2, \ldots, X_n\}$ independientes con distribución

$$P(X_i = x|I) = p(x)$$

para todo $i = 1, \ldots, n$. Entonces, en el límite $n \to \infty$ su suma

$$Z := \sum_{i=1}^{n} X_i \tag{8.62}$$

tiende a seguir una distribución normal,

$$P(Z = z|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{1}{2\sigma^2}(z - \mu)^2 \Big), \tag{8.63}$$

con $\mu = n\langle X \rangle_I$ y $\sigma^2 = n\langle (\delta X)^2 \rangle_I$.

Reinterpretando este teorema en términos del promedio aritmético de $\{X_1, \ldots, X_n\}$ tenemos que

$$P(\overline{X}_n = x|I) \to \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{n}{2\sigma^2}(x - \langle X \rangle_I)^2 \Big) \tag{8.64}$$

para n suficientemente grande, esto es,

$$\langle \overline{X}_n \rangle_I = \langle X \rangle_I, \tag{8.65a}$$

$$\langle (\delta \overline{X}_n)^2 \rangle_I = \frac{1}{n} \langle (\delta X)^2 \rangle_I, \tag{8.65b}$$

y es de esta forma que, en el límite $n \to \infty$ de (8.65), se recupera (8.30) como

$$\lim_{n \to \infty} P(\overline{X}_n = x|I) = \delta\left(x - \langle X \rangle_I\right). \tag{8.66}$$

> **Importante:** Estos resultados nos muestran cómo la incerteza asociada a un promedio aritmético se reduce con el número de observaciones incluidas en dicho promedio, lo cual constituye la base de la estadística tradicional.

De esta forma, la distribución normal juega el papel más importante como la distribución asociada a la combinación —a través del promedio— de mediciones independientes.

Finalmente veamos un bosquejo de la demostración del teorema central del límite usando funciones características.

Demostración. Sea Z una suma de variables

$$Z = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i, \tag{8.67}$$

donde $P(X_i = x|I) = p(x)$ para todo i y donde tanto la media $\langle X \rangle_I$ como la varianza $\langle (\delta X)^2 \rangle_I$ son finitas. La función característica de Z es

$$\varphi_Z(t) = \varphi_X(t)^n. \tag{8.68}$$

Ya que $|\varphi_x(t)| \leq 1$ con $\varphi_x(0) = 1$, para n suficientemente grande sólo la contribución de $t \approx 0$ es importante, y podemos expandir en Taylor en torno a $t = 0$,

$$\begin{aligned} \varphi_z(t) = \varphi_x(t)^n &= \exp(n \ln \varphi_x(t)) \\ &\approx \exp\left(n\left[\ln \varphi_x(0) + t\, L_1(0) + \frac{t^2}{2} L_2(0)\right]\right) \\ &= \exp\left(n\left[t\, L_1(0) + \frac{t^2}{2} L_2(0)\right]\right), \end{aligned} \tag{8.69}$$

donde

$$L_1(t) = \frac{\partial}{\partial t} \ln \varphi_x(t) = \frac{1}{\varphi_x(t)} \frac{\partial}{\partial t} \langle \exp(\mathrm{i}tx) \rangle_I = \frac{\mathrm{i} \langle \exp(\mathrm{i}tx) x \rangle_I}{\varphi_x(t)}, \tag{8.70}$$

$$\begin{aligned} L_2(t) = \frac{\partial^2}{\partial t^2} \ln \varphi_x(t) &= \frac{1}{\varphi_x(t)} \frac{\partial^2}{\partial t^2} \langle \exp(\mathrm{i}tx) \rangle_I - \left(\mathrm{i} \frac{\langle \exp(\mathrm{i}tx) x \rangle_I}{\varphi_x(t)} \right)^2 \\ &= -\frac{1}{\varphi_x(t)} \langle \exp(\mathrm{i}tx) x^2 \rangle_I + \left(\frac{\langle \exp(\mathrm{i}tx) x \rangle_I}{\varphi_x(t)} \right)^2. \end{aligned} \tag{8.71}$$

Evaluando en $t = 0$ tenemos

$$L_1(0) = \mathrm{i} \langle x \rangle_I, \tag{8.72a}$$

$$L_2(0) = -\langle x^2 \rangle_I + \langle x \rangle_I^2 = -\langle (\delta x)^2 \rangle_I, \tag{8.72b}$$

y reemplazando en (8.69) se sigue que

$$\varphi_z(t) \approx \exp\left(\mathrm{i}t \cdot n \langle x \rangle_I - \frac{1}{2} t^2 \cdot n \langle (\delta x)^2 \rangle_I \right), \tag{8.73}$$

que es (8.47) con $\mu = n \langle x \rangle_I$ y $\sigma^2 = n \langle (\delta x)^2 \rangle_I$, es decir, z sigue una distribución normal con

$$\langle z \rangle_I = n \langle x \rangle_I, \tag{8.74a}$$

$$\langle (\delta z)^2 \rangle_I = n \langle (\delta x)^2 \rangle_I \tag{8.74b}$$

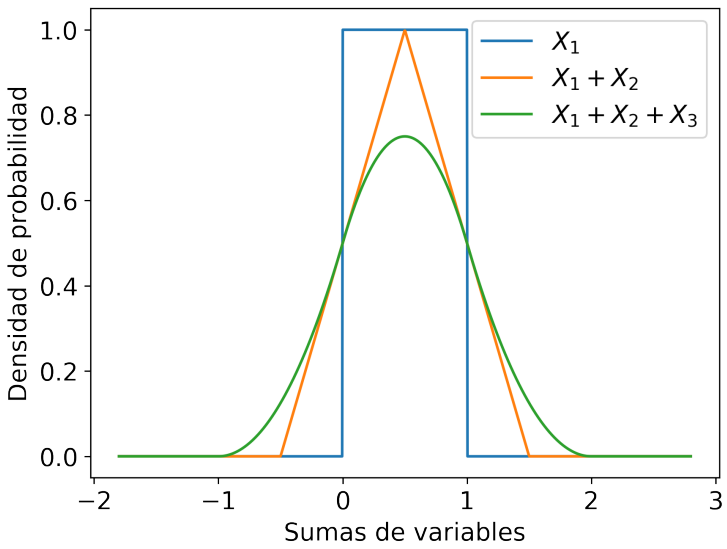


Figura 8.2: Un ejemplo del teorema central del límite. Para tres variables continuas X_1, X_2, X_3 con $X_i \sim \mathrm{U}(0,1)$ se muestra en azul la distribución de X_1, en naranjo la distribución de $X_1 + X_2$, y en verde la distribución de $X_1 + X_2 + X_3$. Todas las distribuciones han sido desplazadas para tener media igual a $1/2$.

En la **Figura 8.2** se muestra el comportamiento de la suma de una, dos y tres variables idénticas, con distribución uniforme $\mathrm{U}(0,1)$, donde vemos que ya para n=3 el comportamiento puede aproximarse muy bien por una distribución normal.

8.6 — Propagación de incerteza en modelos normales

Hemos visto que si $X \sim I$ es una variable continua con varianza tendiendo a cero, su distribución tiende a ser normal. En este caso, ¿qué distribución siguen los valores de una función $f(X)$?

De (7.85) y llamando $x_0 := \langle X \rangle_I$, $\sigma^2 := \langle (\delta X)^2 \rangle_I$, tenemos

$$P(f = F|x_0, \sigma^2) = \sum_{x_F} \frac{P(X = x_F|x_0, \sigma^2)}{|f'(x_F)|}. \tag{8.75}$$

Como X está suficientemente concentrado en torno a x_0, podemos aproximar a primer orden

$$f(x) \approx F_0 + (x - x_0) f'(x_0), \qquad \text{con } F_0 := f(x_0). \tag{8.76}$$

Dado que en esta aproximación $f(x)$ es lineal en x, es por tanto invertible y se tiene

$$x_F = x_0 + \frac{F - F_0}{f'(x_0)}, \tag{8.77}$$

y además $f'(x) \approx f'(x_0)$. Reemplazando en (8.75), tenemos

$$\begin{aligned} P(f = F|x_0, \sigma^2) &= \frac{1}{|f'(x_0)|} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x_F - x_0)^2 \Big) \\ &= \frac{1}{|f'(x_0)|} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2 |f'(x_0)|^2}(F - F_0)^2 \Big) \\ &= \frac{1}{\sqrt{2\pi}\sigma_F} \exp\Big(-\frac{1}{2\sigma_F{}^2}(F - F_0)^2 \Big) \end{aligned} \tag{8.78}$$

donde $\sigma_F := \sigma|f'(x_0)|$. Es decir, si X es una variable con varianza tendiendo a cero y por tanto normal, entonces $f(X)$ también es normal con

$$\langle f \rangle_I = f(\langle X \rangle_I), \tag{8.79}$$

$$\langle (\delta f)^2 \rangle_I = \langle (\delta x)^2 \rangle_I \left(\left| \frac{\mathrm{d}f}{\mathrm{d}x} \right|^2 \right)_{x=x_0}. \tag{8.80}$$

Usando la notación

$$\Delta X := \sqrt{\langle (\delta X)^2 \rangle_I}, \tag{8.81a}$$

$$\Delta f := \sqrt{\langle (\delta f)^2 \rangle_I} \tag{8.81b}$$

podemos escribir (8.80) como

$$\boxed{\Delta f = \Delta X \left(\left| \frac{\mathrm{d}f}{\mathrm{d}x} \right| \right)_{x=x_0},} \tag{8.82}$$

que coincide con el resultado intuitivo del cálculo diferencial si pensamos que ΔX y Δf son cantidades pequeñas comparadas con x_0 y $f(x_0)$ respectivamente. Esta es la *regla de propagación de errores* de la estadística tradicional.

Ejemplo 8.6.1. *Un ángulo θ sigue una distribución normal con $\mu = \pi/3$ y $\sigma = 0.02$. ¿Qué distribución sigue $X := \cos\theta$?*

Solución. *Sabemos por el resultado anterior que X será también normal, con media*

$$\langle X \rangle_{\mu,\sigma} = \cos(\pi/3) = \frac{1}{2}, \tag{8.83}$$

y varianza

$$\langle (\delta X)^2 \rangle_{\mu,\sigma} = (0.02)^2 \times \sin(\pi/3)^2 = 0.0003. \tag{8.84}$$

8.7 — La distribución lognormal

Consideremos una variable normal $z \sim \mathcal{N}(\mu, \sigma^2)$, y definamos una nueva variable $x \geq 0$ a través de la transformación

$$x = f(z) = \exp(z). \tag{8.85}$$

¿Cuál es la distribución que sigue x?

Como $f(z) = \exp(z)$ es una función invertible, podemos usar (7.86) para escribir

$$P(x|\mu,\sigma) = \langle \delta(\exp(z) - x) \rangle_{\mu,\sigma} = \frac{P(z = z_0|\mu,\sigma)}{|f'(z_0)|} \tag{8.86}$$

donde $z_0 = \ln x$ es la solución única de $f(z_0) = \exp(z_0) = x$. Más aún, como $f'(z) = f(z)$ se tiene

$$|f'(z_0)| = |f(z_0)| = |x| = x,$$

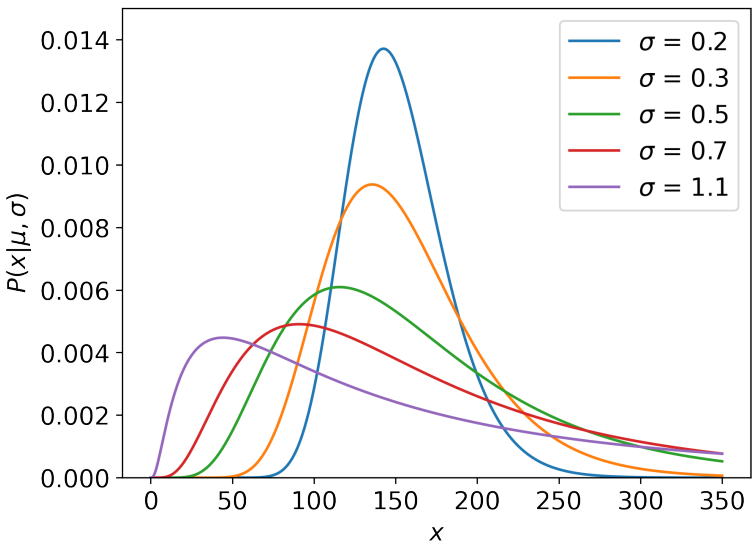


Figura 8.3: La distribución lognormal para μ =5 y valores de σ entre 0.2 y 1.1.

y finalmente

$$P(x|\mu,\sigma) = \frac{P(z = \ln x|\mu,\sigma)}{x} = \frac{1}{x}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)\right).$$

Esta es la llamada distribución lognormal, que se muestra en la **Figura 8.3** para distintos valores de σ.

Definición 8.6 — Distribución lognormal
Una variable $X \geq 0$ sigue una distribución lognormal si

$$P(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma X}\exp\left(-\frac{(\ln X - \mu)^2}{2\sigma^2}\right), \tag{8.87}$$

Diremos que $X \sim \text{LogNorm}(\mu,\sigma^2)$.

PROBLEMAS

Problema 8.1. *Calcule la función característica asociada a la distribución Gamma(k, θ).*

Problema 8.2. *Demuestre, utilizando funciones características, que si $Z = X_1 + \ldots + X_N$ es la suma de N variables exponenciales $X_i \sim Exp(\lambda)$, se cumple $Z \sim Gamma(k = N, \theta = 1/\lambda)$.*

Problema 8.3. *Utilizando el resultado demostrado en el Problema 8.2, muestre explícitamente cómo la distribución Gamma asociada a Z tiende a la distribución normal cuando $N \to \infty$.*

Problema 8.4. *Demuestre, usando la fórmula de convolución, que la suma de dos variables normales es normal y determine la media y varianza de la distribución resultante. Verifique el resultado usando funciones características.*

Problema 8.5. *Demuestre (8.25) por inducción matemática.*

Problema 8.6. *A qué distribución tiende la variable*

$$F := \prod_{i=1}^{n} X_i \tag{8.88}$$

donde $X_1, \ldots, X_n$ son variables independientes no negativas, cuando $n \to \infty$?
Hint: *piense en el logaritmo de F.*

CAPÍTULO 9

Inferencia de parámetros

> It is a capital mistake to theorize before one has the data. Insensibly one begins to twist the facts to suit theories instead of theories to fit the facts.
>
> Sherlock Holmes, A Scandal in Bohemia

Entre las aplicaciones prácticas del teorema de Bayes, probablemente la más conocida y usada, siendo una de las tareas más comunes en la Ciencia, es la de ajustar los valores de los parámetros libres de un modelo usando datos observados. Más rigurosamente diremos que se trata del problema de *inferencia de parámetros*, y lo formularemos de la siguiente manera.

Tenemos un modelo $M(\boldsymbol{\theta})$ donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ es un conjunto de m parámetros que lo definen, y una cantidad $\boldsymbol{X} \sim M$. Además tenemos en nuestro poder dos elementos: un *prior* o *distribución previa* $P(\boldsymbol{\theta}|I_0)$ para los parámetros, y un conjunto de datos D compuesto por n observaciones de X,

$$D := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n). \tag{9.1}$$

Nuestro objetivo es determinar la *distribución posterior* de $\boldsymbol{\theta}$ dados los datos D, que escribiremos como $P(\boldsymbol{\theta}|D, I_0)$ y que de acuerdo al teorema de Bayes puede calcularse como

$$P(\boldsymbol{\theta}|D, I_0) = \frac{P(\boldsymbol{\theta}|I_0)\ P(D|\boldsymbol{\theta}, I_0)}{P(D|I_0)}. \tag{9.2}$$

Para el problema de inferencia de parámetros podemos considerar la probabilidad previa de los datos[(1)] $P(D|I_0)$ como una simple constante de normalización, dada por

$$P(D|I_0) = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}|I_0)\ P(D|\boldsymbol{\theta}, I_0), \tag{9.3}$$

por lo que la única pieza faltante es la probabilidad $P(D|\boldsymbol{\theta}, I_0)$ de observar los datos D bajo un conjunto de parámetros $\boldsymbol{\theta}$, que en estadística se conoce como la *función verosimilitud* (en inglés *likelihood function*). Debido a que en la mayoría de las ocasiones es más manejable, sobre todo cuando se trata

(1) A veces a esta cantidad se le denomina la *evidencia*, con la evidente confusión con el significado de evidencia como una proposición que favorece o perjudica a una hipótesis.

de observaciones independientes, nosotros trabajaremos con el logaritmo de esta función,

$$\mathcal{L}_D(\boldsymbol{\theta}) := \ln P(D|\boldsymbol{\theta}, I_0) = \ln P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\theta}, I_0). \tag{9.4}$$

que llamaremos *log-verosimilitud*.

Cuando el estado de conocimiento previo I_0 sea el de absoluto desconocimiento, lo llamaremos $\varnothing$ y escribiremos

$$P(\boldsymbol{\theta}|D, \varnothing) = \frac{P(\boldsymbol{\theta}|\varnothing)\ P(D|\boldsymbol{\theta}, \varnothing)}{P(D|\varnothing)}, \tag{9.5}$$

y más aún, usaremos la convención de escribir $\varnothing$ sólo cuando es el único símbolo a la derecha de la barra condicional. De esta forma, (9.5) puede escribirse como

$$P(\boldsymbol{\theta}|D) = \frac{P(\boldsymbol{\theta}|\varnothing)\ P(D|\boldsymbol{\theta})}{P(D|\varnothing)}. \tag{9.6}$$

9.1 — Máximo a posteriori y máxima verosimilitud

Como ilustración del proceso de inferencia de parámetros, veamos cómo determinar la distribución posterior de μ y σ de una distribución normal cuando se tienen n observaciones independientes $D = (x_1, \ldots, x_n)$ y se comienza desde un *prior* plano, esto es, usando

$$P(\mu, \sigma|I_0) = \text{constante}. \tag{9.7}$$

La función verosimilitud (probabilidad de observar los datos bajo μ y σ) es

$$\begin{aligned} P(D|\mu, \sigma, I_0) &= \prod_{i=1}^{n} P(x_i|\mu, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big), \end{aligned} \tag{9.8}$$

con lo que el teorema de Bayes en la forma (9.2) nos entrega

$$P(\mu, \sigma|D, I_0) = \underbrace{\Big(\frac{P(\mu, \sigma|I_0)}{P(D|I_0)}\frac{1}{(\sqrt{2\pi})^n}\Big)}_{=\frac{1}{\eta}} \frac{1}{\sigma^n} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big), \tag{9.9}$$

donde hemos definido η como una constante que sólo depende de D y n. Podemos escribir ahora el logaritmo de la probabilidad posterior como

$$\ln P(\mu, \sigma|D, I_0) = -\ln \eta - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - n\ln\sigma, \tag{9.10}$$

y por tanto los valores más probables de μ y σ, que denominaremos $\hat{\mu}$ y $\hat{\sigma}$ respectivamente, estarán dados por la condición de extremo

$$\left(\frac{\partial}{\partial\mu}\ln P(\mu,\sigma|D,I_0)\right)_{\hat{\mu},\hat{\sigma}} = 0 = -\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu}), \tag{9.11a}$$

$$\left(\frac{\partial}{\partial\sigma}\ln P(\mu,\sigma|D,I_0)\right)_{\hat{\mu},\hat{\sigma}} = 0 = \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 - \frac{n}{\hat{\sigma}}. \tag{9.11b}$$

Como $\sigma > 0$, de (9.11a) obtenemos

$$\sum_{i=1}^{n} x_i - n\hat{\mu} = 0, \tag{9.12}$$

es decir

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.} \tag{9.13}$$

Por otro lado, (9.11b) nos dice que

$$\sum_{i=1}^{n}(x_i-\hat{\mu})^2 - n\hat{\sigma}^2 = 0, \tag{9.14}$$

luego se tiene

$$\boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2.} \tag{9.15}$$

Con la definición de promedio aritmético en (8.26) podemos escribir (9.13) y (9.15) simplemente como

$$\hat{\mu} = \overline{x}, \tag{9.16a}$$

$$\hat{\sigma}^2 = \overline{(x-\hat{\mu})^2}, \tag{9.16b}$$

con una inquietante similitud con (8.2),

$$\mu = \langle x\rangle_{\mu,\sigma},$$
$$\sigma^2 = \langle (x-\mu)^2\rangle_{\mu,\sigma}.$$

El significado de esta similitud, y en general de la conexión entre promedio aritmético y expectación, lo entenderemos en la **Sección 9.3**.

A través de este ejemplo hemos demostrado el método por excelencia de la estadística tradicional (no bayesiana) para la inferencia de parámetros, conocido como el *método de máxima verosimilitud* y cuya exactitud está garantizada en el límite de muchas observaciones ($n \to \infty$).

A continuación formularemos este método en su mayor generalidad.

Recuadro 9.1 — Método de máxima verosimilitud
Para un modelo $M(\boldsymbol{\theta})$ y datos $D = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, el conjunto de parámetros $\hat{\boldsymbol{\theta}}$ que mejor se ajusta a los datos es el que maximiza la función log-verosimilitud $\mathcal{L}_D(\boldsymbol{\theta}) = \ln P(D|\boldsymbol{\theta}, I_0)$, esto es,

$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \mathcal{L}_D(\boldsymbol{\theta}), \tag{9.17}$$

y por tanto $\hat{\boldsymbol{\theta}}$ debe cumplir la condición de extremo

$$\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \mathcal{L}_D(\boldsymbol{\theta})\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}. \tag{9.18}$$

> **Importante:** Para nosotros el método de máxima verosimilitud es sólo una aproximación, válida cuando se justifique una elección de un *prior* plano para los parámetros $\boldsymbol{\theta}$ o, como veremos, cuando el número de observaciones sea suficientemente grande para que la elección del *prior* no tenga efecto en la distribución posterior.

El método más general que reemplaza al de máxima verosimilitud, válido para cualquier elección de *prior* y número de observaciones es el método de *máximo a posteriori*, en el cual se obtiene el conjunto de parámetros $\boldsymbol{\theta}^*$ que maximiza (el logaritmo de) la distribución posterior $P(\boldsymbol{\theta}|D, I_0)$.

Recuadro 9.2 — Método de máximo a posteriori
Para un modelo $M(\boldsymbol{\theta})$ y datos $D = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, el conjunto de parámetros $\boldsymbol{\theta}^*$ que mejor se ajusta a los datos es el que maximiza el logaritmo de la distribución posterior $\ln P(\boldsymbol{\theta}|D, I_0)$, esto es,

$$\boldsymbol{\theta}^* := \arg\max_{\boldsymbol{\theta}} \ln P(\boldsymbol{\theta}|D, I_0), \tag{9.19}$$

y por tanto $\boldsymbol{\theta}^*$ debe cumplir la condición de extremo

$$\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln P(\boldsymbol{\theta}|D, I_0)\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbf{0}. \tag{9.20}$$

Si las n observaciones son independientes, se tendrá

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln P(\boldsymbol{\theta}|D, I_0) = \frac{\partial}{\partial \boldsymbol{\theta}} \Big(\ln P(\boldsymbol{\theta}|I_0) + \underbrace{\ln P(D|\boldsymbol{\theta}, I_0)}_{=\mathcal{L}_D(\boldsymbol{\theta})} + \cancelto{0}{\ln P(D|I_0)} \Big) \tag{9.21}$$

pero $\mathcal{L}_D(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln P(\boldsymbol{X} = \boldsymbol{x}_i|\boldsymbol{\theta}, I_0)$ crece con n y domina frente a $\ln P(\boldsymbol{\theta}|I_0)$. De esta forma el método de máximo *a posteriori* converge al método de máxima verosimilitud para $n \to \infty$.

9.2 — El prior no informativo de Jeffreys

En esta sección abordaremos el problema de encontrar distribuciones previas cuando carecemos de conocimiento basado en observaciones. Este es uno de los problemas centrales de la inferencia bayesiana, y nuestro principio guía será el de invarianza ante transformaciones.

Comenzaremos buscando una distribución previa $P(X|\varnothing)$ para una cantidad $X \geq 0$ invariante de escala. Esto es, consideraremos la condición de que las variables X y αX (con $\alpha > 0$) siguen la misma distribución, es decir,

$$P(X = x|\varnothing) = f(x), \tag{9.22a}$$
$$P(\alpha X = z|\varnothing) = f(z). \tag{9.22b}$$

¿Existe alguna función $f(x)$ que cumpla esto? Escribiendo (9.22b) en términos de la expectación de la delta de Dirac y extrayendo la constante α, tenemos

$$P(\alpha X = z|\varnothing) = \langle \delta(\alpha X - z) \rangle_{\varnothing} = \frac{1}{\alpha} P(X = \tfrac{z}{\alpha}|\varnothing). \tag{9.23}$$

Es decir, de acuerdo a (9.22a) debe cumplirse

$$f(z) = \frac{1}{\alpha} f(z/\alpha), \tag{9.24}$$

y si derivamos a ambos lados respecto a α, obtenemos la ecuación diferencial de primer orden

$$\frac{z}{\alpha} f'\left(\frac{z}{\alpha}\right) = -f\left(\frac{z}{\alpha}\right) \tag{9.25}$$

que bajo el cambio de variable $u := z/\alpha$ puede escribirse como

$$\frac{f'}{f} = -\frac{1}{u}, \quad \text{con solución general} \quad f(u) \propto \frac{1}{u}.$$

Finalmente entonces, nuestra distribución previa es de la forma

$$P(X = x|\varnothing) \propto \frac{1}{x}, \tag{9.26}$$

conocida como el *prior* de Jeffreys. Notemos que este *prior* no es normalizable, y podemos leer este resultado como el hecho de que no podemos dar una estimación razonable para X si sólo sabemos que $X \geq 0$. Por otro lado, consideremos una variable de la cual únicamente sabemos que $X \in \mathbb{R}$, buscamos invarianza traslacional, esto es

$$P(X = x|\varnothing) = f(x), \tag{9.27a}$$
$$P(X + \beta = z|\varnothing) = f(z). \tag{9.27b}$$

Nuevamente, escribimos (9.27b) como la expectación de una delta de Dirac y entonces vemos que

$$P(X + \beta = z|\varnothing) = \langle \delta(X + \beta - z) \rangle_{\varnothing} = P(X = z - \beta|\varnothing), \tag{9.28}$$

luego combinando con (9.27a) tenemos

$$f(z) = f(z - \beta) \tag{9.29}$$

para todo z, β, por lo tanto $f(z)$ = constante para todo z, y se sigue que

$$\boxed{P(X = x|\varnothing) = \text{constante.}} \tag{9.30}$$

Ejemplo 9.2.1. *Los tiempos de viaje t en un recorrido de buses desde inicio a final siguen una distribución exponencial*

$$P(t|\lambda) = \lambda \exp(-\lambda t) \tag{9.31}$$

con $\lambda > 0$. Tenemos n tiempos observados independientes $D = (t_1, t_2, \ldots, t_n)$ y usamos el prior de Jeffreys para λ,

$$P(\lambda|\varnothing) \propto \frac{1}{\lambda}. \tag{9.32}$$

El teorema de Bayes nos entrega la distribución posterior

$$P(\lambda|D) = \frac{P(\lambda|\varnothing)P(D|\lambda)}{P(D|\varnothing)}, \tag{9.33}$$

donde la función verosimilitud es

$$\begin{aligned} P(D|\lambda) &= P(t_1, t_2, \ldots, t_n|\lambda) \\ &= \prod_{i=1}^{n} P(t_i|\lambda) \\ &= \lambda^n \prod_{i=1}^{n} \exp(-\lambda t_i) \\ &= \lambda^n \exp(-n\lambda \bar{t}), \end{aligned} \tag{9.34}$$

y con

$$\bar{t} := \frac{1}{n} \sum_{i=1}^{n} t_i \tag{9.35}$$

el promedio aritmético de las observaciones $t_1, t_2, \ldots, t_n$. Por lo tanto, la distribución posterior de λ es

$$P(\lambda|D) = \frac{\lambda^{n-1} \exp(-n\lambda \bar{t})}{\eta(n, \bar{t})} \tag{9.36}$$

la cual es completamente caracterizada por el número de observaciones n y el valor promedio $\bar{t}$, por lo que podemos escribirla como $P(\lambda|n, \bar{t})$. Esta es una distribución gamma, por lo que de inmediato sabemos su normalización,

$$P(\lambda|n, \bar{t}) = \frac{\lambda^{n-1} \exp(-n\lambda \bar{t})}{\Gamma(n)(n\bar{t})^{-n}}. \tag{9.37}$$

La media y varianza de λ están dadas por

$$\langle \lambda \rangle_{n,\bar{t}} = \frac{\cancel{n}}{\cancel{n}\bar{t}} = \frac{1}{\bar{t}}, \tag{9.38a}$$

$$\langle (\delta\lambda)^2 \rangle_{n,\bar{t}} = \frac{\cancel{n}}{n\bar{t}^{\cancel{2}}} = \frac{1}{n\bar{t}^2}, \tag{9.38b}$$

y como en el límite $n \to \infty$ *se tiene*

$$\left\langle (\delta\lambda)^2 \right\rangle_{n,\bar{t}} \to 0$$

vemos que λ *toma un valor único,* $\lambda_0 = \dfrac{1}{\bar{t}}$*, y entonces la distribución de* λ *se vuelve de conocimiento completo,*

$$\lim_{n\to\infty} P(\lambda|n,\bar{t}) = \delta\left(\lambda - \frac{1}{\bar{t}}\right). \tag{9.39}$$

Notemos que la contribución del prior *de Jeffreys es corregir el factor* λ^n *a* λ^{n-1}*, con lo que dicha contribución se hace despreciable cuando* $n \to \infty$*, y entonces los métodos de máximo a posteriori y máxima verosimilitud se hacen equivalentes. Además, como sabemos que si* $t \sim Exp(\lambda)$ *se cumple*

$$\left\langle t \right\rangle_{\lambda} = \int_0^\infty \mathrm{d}t\lambda \exp(-\lambda t)t = \frac{1}{\lambda} \tag{9.40}$$

se tendrá que

$$\lim_{n\to\infty} \left\langle t \right\rangle_{n,\bar{t}} = \left\langle t \right\rangle_{\lambda_0} = \frac{1}{\lambda_0} = \bar{t}. \tag{9.41}$$

9.3 — La ley de los grandes números revisitada

En general, una de las consecuencias del teorema de Bayes para observaciones independientes es que para $n \to \infty$ las expectaciones tienden a coincidir con los valores promedio de las cantidades, como se ve en (9.41), un ejemplo de la ley de los grandes números que volveremos a obtener, en este contexto de inferencia de parámetros, a continuación.

Sea $P(\boldsymbol{x}|I)$ la densidad de probabilidad de una variable $\boldsymbol{x} \sim I \in U$. Como toda densidad de probabilidad es no negativa, es posible escribirla como

$$P(\boldsymbol{x}|I) = \frac{1}{\eta[F]} \exp(F(\boldsymbol{x})), \tag{9.42}$$

donde $F(\boldsymbol{x})$ es una función en los reales y $\eta[F]$ es un funcional que actúa como una constante de normalización, dada por

$$\eta[F] = \int_U \mathrm{d}\boldsymbol{x} \exp(F(\boldsymbol{x})). \tag{9.43}$$

Ahora supongamos una base completa de funciones $\phi_0(\boldsymbol{x}), \phi_1(\boldsymbol{x}), \ldots$ y procedamos a escribir $F(\boldsymbol{x})$ en términos de esta base, como

$$F(\boldsymbol{x}) = \sum_{n=0}^{\infty} \lambda_n \phi_n(\boldsymbol{x}), \tag{9.44}$$

donde los coeficientes $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots)$ describen completamente a la función F. Usando esta representación podemos escribir $P(\boldsymbol{x}|I)$ como

$$P(\boldsymbol{x}|I) = \frac{1}{\eta(\boldsymbol{\lambda})} \exp\left(\sum_{n=0}^{\infty} \lambda_n \phi_n(\boldsymbol{x})\right) \tag{9.45}$$

donde

$$\eta(\boldsymbol{\lambda}) = \int_U \mathrm{d}\boldsymbol{x} \exp\Big(\sum_{n=0}^{\infty} \lambda_n \phi_n(\boldsymbol{x})\Big) \tag{9.46}$$

es ahora una función de los coeficientes $\boldsymbol{\lambda}$. De esta forma, las distintas funciones $P(\boldsymbol{x}|I)$ normalizables quedan representadas por los distintos conjuntos de coeficientes $\boldsymbol{\lambda}$.

Ahora consideremos un conjunto de N observaciones $D = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ independientes, para las cuales queremos encontrar la mejor función $P(\boldsymbol{x}|D, I_0)$ que las representa, esto es, el mejor conjunto de coeficientes $\boldsymbol{\lambda}$. De acuerdo al teorema de Bayes, tenemos

$$\begin{aligned} P(\boldsymbol{\lambda}|D, I_0) &= \frac{P(\boldsymbol{\lambda}|I_0) \cdot P(D|\boldsymbol{\lambda}, I_0)}{P(D|I_0)} \\ &= \frac{P(\boldsymbol{\lambda}|I_0)}{P(D|I_0)} \prod_{i=1}^{N} P(\boldsymbol{x}_i|\boldsymbol{\lambda}, I_0) \\ &= \frac{P(\boldsymbol{\lambda}|I_0)}{P(D|I_0)} \prod_{i=1}^{N} \frac{1}{\eta(\boldsymbol{\lambda})} \exp\Big(\sum_{n=0}^{\infty} \lambda_n \phi_n(\boldsymbol{x})\Big) \\ &= \frac{P(\boldsymbol{\lambda}|I_0)}{P(D|I_0)} \exp\Big(\sum_{i=1}^{N} \sum_{n=0}^{\infty} \lambda_n \phi_n(\boldsymbol{x}_i) - N \ln \eta(\boldsymbol{\lambda})\Big) \\ &= \frac{1}{P(D|I_0)} \exp\left(N\Big[\sum_{n=0}^{\infty} \lambda_n \overline{\phi_n} - \ln \eta(\boldsymbol{\lambda})\Big] + \ln P(\boldsymbol{\lambda}|I_0)\right), \end{aligned} \tag{9.47}$$

donde hemos introducido el promedio aritmético

$$\overline{\phi_n} = \frac{1}{N} \sum_{i=1}^{N} \phi_n(\boldsymbol{x}_i).$$

El método de máximo a posteriori nos entrega entonces

$$\begin{aligned} \frac{\partial}{\partial \lambda_k} \ln P(\boldsymbol{\lambda}|D, I_0) &= N \frac{\partial}{\partial \lambda_k} \left[\sum_{n=0}^{\infty} \lambda_n \overline{\phi_n} - \ln \eta(\boldsymbol{\lambda})\right] + \frac{\partial}{\partial \lambda_k} \ln P(\boldsymbol{\lambda}|I_0) \\ &= N \left[\overline{\phi_k} - \frac{\partial}{\partial \lambda_k} \ln \eta(\boldsymbol{\lambda})\right] + \frac{\partial}{\partial \lambda_k} \ln P(\boldsymbol{\lambda}|I_0) = 0. \end{aligned} \tag{9.48}$$

Tomando derivada parcial de $\ln \eta(\boldsymbol{\lambda})$ respecto a λ_k vemos que

$$\frac{\partial}{\partial \lambda_k} \ln \eta(\boldsymbol{\lambda}) = \frac{1}{\eta(\boldsymbol{\lambda})} \int_U \mathrm{d}\boldsymbol{x} \exp\Big(\sum_{n=0}^{\infty} \lambda_n \phi_n(\boldsymbol{x})\Big) \phi_k(\boldsymbol{x}) = \big\langle \phi_k \big\rangle_{D, I_0}, \tag{9.49}$$

luego tenemos

$$N \left[\overline{\phi_k} - \big\langle \phi_k \big\rangle_I\right] + \frac{\partial}{\partial \lambda_k} \ln P(\boldsymbol{\lambda}|I_0) = 0. \tag{9.50}$$

Para $N \to \infty$ nos damos cuenta que el primer término domina, y luego se debe cumplir

$$\lim_{N \to \infty} \overline{\phi_k} = \big\langle \phi_k \big\rangle_{D, I_0} \qquad \text{para todo } k. \tag{9.51}$$

Como la base es completa, cualquier función de prueba $\omega(\boldsymbol{x})$ puede descomponerse en términos de ella,

$$\omega(\boldsymbol{x}) = \sum_{n=0}^{\infty} \mu_n \phi_n(\boldsymbol{x}), \tag{9.52}$$

y por lo tanto para ω también se cumplirá

$$\begin{aligned}\lim_{N\to\infty}\overline{\omega} = \lim_{N\to\infty}\left(\sum_{n=0}^{\infty}\mu_n\overline{\phi_n}\right) &= \sum_{n=0}^{\infty}\mu_n\Big(\lim_{N\to\infty}\overline{\phi_n}\Big)\\ &= \sum_{n=0}^{\infty}\mu_n\langle\phi_n\rangle_I \\ &= \left\langle\sum_{n=0}^{\infty}\mu_n\phi_n\right\rangle_{D,I_0} = \langle\omega\rangle_{D,I_0}.\end{aligned} \tag{9.53}$$

Con esto hemos recuperado la ley de los grandes números, en una nueva interpretación.

Teorema 9.1 — Ley de los grandes números (revisitada)
Para un conjunto de N observaciones independientes $D = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ de una variable $\boldsymbol{x} \in U$ con $N \to \infty$, el modelo más probable $P(\boldsymbol{x}|D, I_0)$ que representa a $\boldsymbol{x}$ es aquel que cumple

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\omega(\boldsymbol{x}_i) = \int_U \mathrm{d}\boldsymbol{x}P(\boldsymbol{x}|D, I_0)\omega(\boldsymbol{x}) = \langle\omega\rangle_{D,I_0}. \tag{9.54}$$

para toda función $\omega(\boldsymbol{x})$.

La conexión de la probabilidad con la frecuencia

Notemos que para $\omega(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}')$ con $\boldsymbol{x}'$ un punto arbitrario, (9.54) nos dice que el modelo $P(\boldsymbol{x}|D, I_0)$ puede definirse como

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}_i - \boldsymbol{x}') = \langle\delta(\boldsymbol{x} - \boldsymbol{x}')\rangle_{D,I_0} = P(\boldsymbol{x}'|D, I_0). \tag{9.55}$$

Podemos entender esto como si cada punto observado $\boldsymbol{x}_i$ contribuye de igual manera a la densidad de probabilidad, con una delta de Dirac centrada en él, y por tanto las regiones del espacio que concentran más puntos observados serán más probables según $P(\boldsymbol{x}|D, I_0)$. ¡Pero esto no es más que la idea de histograma!

Para llevarla a cabo en la práctica, subdividiremos el espacio U en m regiones disjuntas $E_1, \ldots, E_m$, y en lugar de la delta de Dirac usaremos la función indicador $\mathrm{Q}(\boldsymbol{x} \in E_j)$ que evalúa la pertenencia del punto $\boldsymbol{x}$ a una región E_j. Reemplazando $\omega(\boldsymbol{x}) = \mathrm{Q}(\boldsymbol{x} \in E_j)$ en (9.54), tenemos

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathrm{Q}(\boldsymbol{x}_i \in E_j) &= \langle\mathrm{Q}(\boldsymbol{x} \in E_j)\rangle_{D,I_0} \\ &= P(\boldsymbol{x} \in E_j|D, I_0), \quad \text{con } j = 1, \ldots, m.\end{aligned} \tag{9.56}$$

Acá es conveniente definir la *frecuencia* de A en N observaciones como

$$f_N(A) := \frac{\text{número de casos donde } A = \mathbb{T}}{\text{número de casos totales}} = \frac{\sum_{i=1}^{N}\mathrm{Q}(A_i)}{N}, \tag{9.57}$$

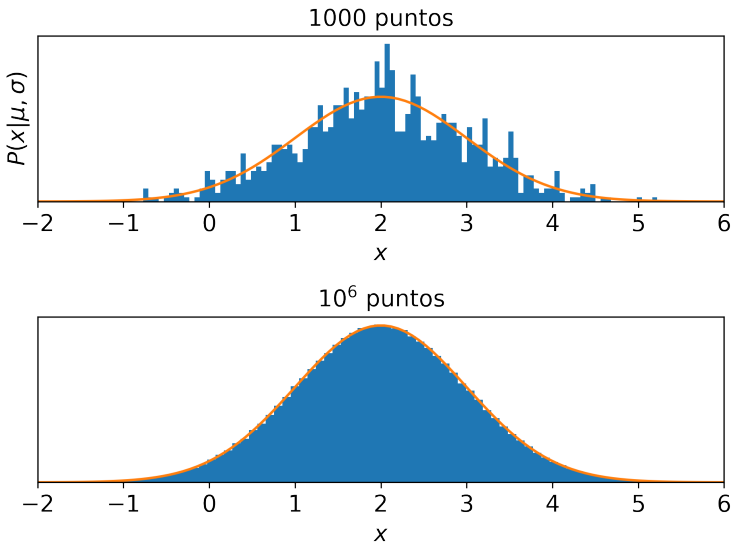


Figura 9.1: Dos histogramas de valores obtenidos de una distribución $\mathcal{N}(2,1)$. Arriba, usando 1000 valores, abajo con un millón de valores. La curva naranja es la densidad de probabilidad exacta a la cual la frecuencia converge.

y más aún,

$$f(A) := \lim_{N\to\infty} f_N(A). \tag{9.58}$$

Usando estas definiciones podemos escribir de forma compacta

$$P(\boldsymbol{x} \in E_j|D, I_0) = \lim_{N\to\infty} f_N(\boldsymbol{x} \in E_j) = f(\boldsymbol{x} \in E_j), \qquad j = 1, \ldots, m, \tag{9.59}$$

notando que $P(\boldsymbol{x} \in E_j|D, I_0)$ depende únicamente de los puntos D y no de la información previa I_0, con lo que podemos escribir

$$P(\boldsymbol{x} \in E_j|D) = f(\boldsymbol{x} \in E_j), \qquad j = 1, \ldots, m. \tag{9.60}$$

Lo que (9.60) nos revela es que en el límite de infinitas observaciones independientes de un fenómeno repetible, el mejor modelo depende únicamente de los datos y deja de depender de cualquier información previa I_0, y es en ese sentido que lo vemos como si fuera una propiedad objetiva del fenómeno.

La **Figura 9.1** muestra precisamente cómo la frecuencia f_N observada de variables normales generadas computacionalmente converge a la densidad de probabilidad correspondiente a la distribución normal.

> **Importante:** En general, $P(A|I) \neq f(A)$, ya que para nosotros la probabilidad no es una propiedad objetiva de la realidad sino un instrumento de recopilación de información.

En el **Ejemplo 9.3.1** vemos cómo usar la ley de los grandes números para estimar numéricamente el área de una región. Esta es la idea central de los llamados métodos de Monte Carlo, que veremos en mayor detalle en el **Capítulo 14**.

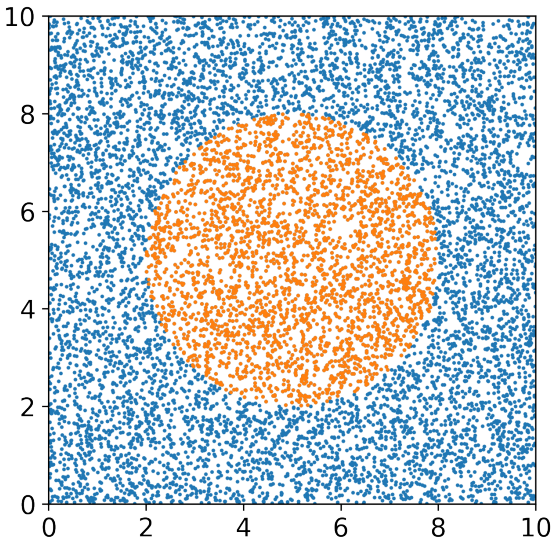


Figura 9.2: Puntos generados al azar para estimar una razón entre áreas según (9.65).

Ejemplo 9.3.1 (Estimación del área de una región)**.**
Consideremos el área A de una región S en dos dimensiones, dada por

$$A = \int_{\boldsymbol{x}\in S} \mathrm{d}\boldsymbol{x} = \int d\boldsymbol{x}\, \mathrm{Q}(\boldsymbol{x} \in S), \tag{9.61}$$

y definamos una distribución plana $P(\boldsymbol{x}|\varnothing) = p_0$ en una región cuadrada C de área L^2, que contiene a la región S, de esta forma aseguramos que $L^2 \geq A$. Imponiendo normalización vemos que

$$\int_C \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}|\varnothing) = p_0 \int_C \mathrm{d}\boldsymbol{x} = p_0\, L^2 = 1, \tag{9.62}$$

es decir, $P(\boldsymbol{x}|\varnothing) = p_0 = 1/L^2$. Si ahora escribimos A como una expectación bajo $\varnothing$, tenemos

$$\begin{aligned} A = \int d\boldsymbol{x}\, \mathrm{Q}(\boldsymbol{x} \in S) &= \int d\boldsymbol{x}\, P(\boldsymbol{x}|\varnothing) \left[\frac{\mathrm{Q}(\boldsymbol{x} \in S)}{P(\boldsymbol{x}|\varnothing)}\right] \\ &= \int d\boldsymbol{x}\, P(\boldsymbol{x}|\varnothing) \left[\mathrm{Q}(\boldsymbol{x} \in S) \cdot L^2\right] \\ &= \langle \mathrm{Q}(\boldsymbol{x} \in S)\rangle_{\varnothing} \cdot L^2, \end{aligned} \tag{9.63}$$

esto es,

$$\frac{A}{L^2} = \langle \mathrm{Q}(\boldsymbol{x} \in S)\rangle_{\varnothing} = P(\boldsymbol{x} \in S|\varnothing). \tag{9.64}$$

Al usar la ley de los grandes números (9.60) con D un conjunto de $N \to \infty$ puntos en el cuadrado C, y usando $m = 2$, $E_1 = S$, y $E_2 = C - S$ tenemos

$$P(\boldsymbol{x} \in S|D) = \frac{A}{L^2} = f(\boldsymbol{x} \in S) \tag{9.65}$$

donde $f(\boldsymbol{x} \in S)$ es la fracción de puntos en C que caen dentro de la región S, en el límite de infinitos puntos. Usando $A = \pi R^2$ esto nos da una aproximación para π,

$$\pi = \lim_{N\to\infty} \left(\frac{L}{R}\right)^2 f_N(\boldsymbol{x} \in S), \tag{9.66}$$

en el caso de que S sea un círculo de radio R contenido íntegramente dentro de la región cuadrada C, como se ve en la Figura 9.2.

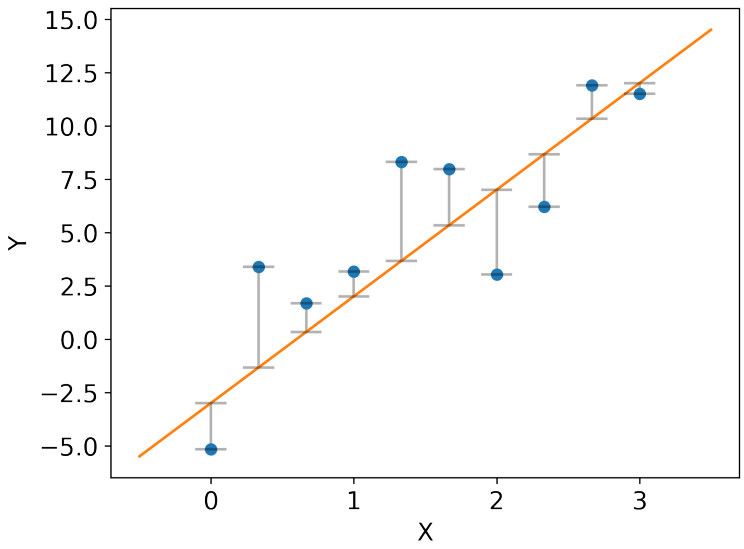


Figura 9.3: Un ejemplo de regresión lineal, donde los puntos azules son observaciones de pares (x, y) y la recta naranja es una de las rectas plausibles que explican la relación entre X e Y según (9.67). Las barras verticales en gris indican las desviaciones de los valores observados respecto a la predicción.

9.4 — El problema de la regresión

Aplicaremos las técnicas y conceptos aprendidos en este capítulo para el que probablemente es el ejemplo universalmente más aplicado de inferencia, la regresión lineal. Consideremos una relación lineal entre dos variables X e Y, esto es

$$Y(X) = \alpha X + \beta \tag{9.67}$$

con parámetros α (pendiente) y β (intercepto) desconocidos y que quisiéramos determinar a partir de n observaciones de pares (x, y),

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, \tag{9.68}$$

como se ve en la **Figura 9.3**. Si no existieran incertezas en los valores de x e y por supuesto bastarían dos puntos cualesquiera para determinar α y β exactamente. Supondremos, como se acostumbra usualmente, que la variable independiente X se conoce exactamente (porque es la que controlamos, digamos, en un experimento), mientras que la variable Y tiene una incerteza o error ε asociado, de forma que la i-ésima observación está dada por

$$y_i = \alpha x_i + \beta + \varepsilon_i \tag{9.69}$$

con $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, esto es, la incerteza ε sigue una distribución normal de media cero y varianza σ^2,

$$P(\varepsilon|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \tag{9.70}$$

Nuestras cantidades desconocidas son entonces α, β y σ, y como ya hemos visto en los casos anteriores, debemos calcular la distribución posterior $P(\alpha, \beta, \sigma|D, I_0)$ usando el teorema de Bayes,

$$P(\alpha, \beta, \sigma|D, I_0) = \frac{P(\alpha, \beta, \sigma|I_0) P(D|\alpha, \beta, \sigma, I_0)}{P(D|I_0)}. \tag{9.71}$$

Consideraremos nuestro estado de conocimiento inicial como el no informativo, luego $I_0 = \varnothing$ y nuestro *prior* para α, β será plano, dado que suponemos $\alpha, \beta \in (-\infty, \infty)$,

$$P(\alpha, \beta|\varnothing) = \text{constante}, \tag{9.72}$$

mientras que para σ usaremos el *prior* de Jeffreys

$$P(\sigma|\varnothing) \propto \frac{1}{\sigma}. \tag{9.73}$$

Como las observaciones de los pares son independientes, nuestra función verosimilitud queda de la forma

$$\begin{aligned} P(D|\alpha, \beta, \sigma) &= \prod_{i=1}^{n} P(x_i, y_i|\alpha, \beta, \sigma) \\ &= \prod_{i=1}^{n} P(y_i|x_i, \alpha, \beta, \sigma) P(x_i|\alpha, \beta, \sigma) \\ &= \prod_{i=1}^{n} P(y_i|x_i, \alpha, \beta, \sigma) P(x_i|I_0), \end{aligned} \tag{9.74}$$

pero la distribución de probabilidad de y dado x es esencialmente la distribución de ε,

$$P(Y = y|X = x, \alpha, \beta, \sigma) = P(\varepsilon = y - (\alpha x + \beta)|\sigma), \tag{9.75}$$

por lo que, definiendo $p_i := P(x_i|\varnothing)$, tenemos

$$\begin{aligned} P(D|\alpha, \beta, \sigma) &= \prod_{i=1}^{n} p_i P(\varepsilon_i = y_i - (\alpha x_i + \beta)|\sigma) \\ &= \prod_{i=1}^{n} \frac{p_i}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - [\alpha x_i + \beta]\right)^2\right) \\ &= \frac{1}{Z_D} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - [\alpha x_i + \beta]\right)^2 - n \ln \sigma\right), \end{aligned} \tag{9.76}$$

donde Z_D es una constante, por ahora sin importancia, que sólo depende de los datos D. La función log-verosimilitud para nuestro problema es entonces,

$$\mathcal{L}_D(\alpha, \beta, \sigma) = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} (y_i - [\alpha x_i + \beta])^2\right] - n \ln \sigma - \ln Z_D. \tag{9.77}$$

Antes de atacar el caso más general de este problema, consideremos la suposición usual de que σ es un valor conocido. En ese caso la función log-verosimilitud se simplifica a

$$\mathcal{L}_D(\alpha, \beta) = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} \left(y_i - [\alpha x_i + \beta]\right)^2\right] - L_0, \tag{9.78}$$

donde nuevamente L_0 es una constante sin importancia. Dado que $\sigma^2 > 0$, maximizar $\mathcal{L}_D(\alpha, \beta)$ en (9.78) es equivalente a minimizar

$$\mathcal{E}^2(\alpha, \beta) := \sum_{i=1}^{n} \left(y_i - [\alpha x_i + \beta]\right)^2 \tag{9.79}$$

como función de los parámetros de la recta, α y β, y hemos recuperado el conocido *método de los mínimos cuadrados*.

Recuadro 9.3 — Método de los mínimos cuadrados
Los parámetros más probables $\hat{\boldsymbol{\theta}}$ que ajustan una curva $y = f(x;\boldsymbol{\theta})$ a un conjunto de n puntos $\{(x_i, y_i)\}$ son tales que

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \big(y_i - f(x_i;\boldsymbol{\theta})\big)^2. \tag{9.80}$$

Este método sólo es correcto si $y = f(x;\boldsymbol{\theta}) + \epsilon$ con $\epsilon \sim \mathcal{N}(0,\sigma^2)$ y se supone σ conocido y un *prior* plano para $\boldsymbol{\theta}$,

$$P(\boldsymbol{\theta}|I_0) = \text{constante}.$$

Ahora usaremos la condición de extremo de $\varepsilon^2(\alpha,\beta)$ para buscar los valores $\hat{\alpha}$ y $\hat{\beta}$ que la minimizan. Derivando respecto a α, tenemos

$$\begin{aligned}\Big(\frac{\partial \mathcal{E}^2}{\partial \alpha}\Big)\Big|_{\alpha=\hat{\alpha},\beta=\hat{\beta}} = 0 &= -2\sum_{i=1}^{n}(y_i - [\hat{\alpha}x_i + \hat{\beta}])\cdot x_i \\ &= -2n\underbrace{\Big[\overline{xy} - \hat{\alpha}\overline{x^2} - \hat{\beta}\overline{x}\Big]}_{=0 \text{ ya que } n>0},\end{aligned} \tag{9.81}$$

y luego se tiene que

$$\hat{\alpha}\overline{x^2} + \hat{\beta}\overline{x} = \overline{xy}. \tag{9.82}$$

Por otro lado, derivando con respecto a β tenemos

$$\begin{aligned}\Big(\frac{\partial \mathcal{E}^2}{\partial \beta}\Big)\Big|_{\alpha=\hat{\alpha},\beta=\hat{\beta}} = 0 &= -2\sum_{i=1}^{n}(y_i - [\hat{\alpha}x_i + \hat{\beta}]) \\ &= -2n\underbrace{\big[\overline{y} - \hat{\alpha}\overline{x} - \hat{\beta}\big]}_{=0 \text{ ya que } n>0},\end{aligned} \tag{9.83}$$

y entonces llegamos a que

$$\hat{\alpha}\overline{x} + \hat{\beta} = \overline{y}. \tag{9.84}$$

Combinando (9.82) y (9.84) obtenemos una ecuación para $\hat{\alpha}$ que sólo involucra los datos,

$$\hat{\alpha}\overline{x^2} + (\overline{y} - \hat{\alpha}\overline{x})\overline{x} = \overline{xy}, \tag{9.85}$$

con la que finalmente podemos despejar $\hat{\alpha}$ y $\hat{\beta}$ como los conocidos estimadores de mínimos cuadrados para la regresión lineal.

Recuadro 9.4 — Mínimos cuadrados para la regresión lineal

$$\hat{\alpha} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}, \tag{9.86}$$

$$\hat{\beta} = \overline{y} - \overline{x}\left(\frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}\right). \tag{9.87}$$

Recordando la definición de la correlación de Pearson en (8.20) podemos escribir

$$\hat{\alpha} = \rho_{XY}\left(\frac{s_y}{s_x}\right). \tag{9.88}$$

donde las desviaciones estándar s_x y s_y de los datos están dadas por

$$s_x := \sqrt{\overline{x^2} - \overline{x}^2}, \tag{9.89a}$$

$$s_y := \sqrt{\overline{y^2} - \overline{y}^2}. \tag{9.89b}$$

Ahora volvamos a la solución más general que tenemos con la función log-verosimilitud en (9.77) y escribamos la distribución posterior

$$P(\alpha, \beta, \sigma|D) = \frac{1}{\eta_D} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - [\alpha x_i + \beta])^2 - (n+1)\ln\sigma\right), \tag{9.90}$$

donde $\eta_D := Z_D \cdot P(D|I_0)$ es una constante sin importancia. Primero reescribiremos la suma sobre los pares que allí aparece para hacerla más manejable, expandiendo

$$\varepsilon^2 = \sum_{i=1}^{n}(y_i - [\alpha x_i + \beta])^2 = n\left(\overline{y^2} + \alpha^2\overline{x^2} + \beta^2 + 2\alpha\beta\overline{x} - 2\alpha\overline{xy} - 2\beta\overline{y}\right). \tag{9.91}$$

Acá completamos los cuadrados de α y de β como

$$\varepsilon^2 = \gamma + K_{11}(\alpha - \alpha^*)^2 + K_{22}(\beta - \beta^*)^2 + 2K_{12}(\alpha - \alpha^*)(\beta - \beta^*) \tag{9.92}$$

con lo que por simple inspección podemos determinar

$$K_{11} = n\overline{x^2}, \tag{9.93a}$$

$$K_{22} = n, \tag{9.93b}$$

$$K_{12} = n\overline{x}, \tag{9.93c}$$

dado que son los coeficientes de los únicos términos que incluyen α^2, β^2 y $2\alpha\beta$, respectivamente. De expandir los términos con K_{11} y K_{12} vemos que debe cumplirse

$$-2K_{11}\alpha^*\alpha - 2K_{12}\beta^*\alpha = -2n(\overline{x^2}\alpha^* + \overline{x}\beta^*)\alpha = -2n\overline{xy}\alpha \tag{9.94}$$

es decir

$$\alpha^*\overline{x^2} + \beta^*\overline{x} = \overline{xy}. \tag{9.95}$$

Haciendo lo mismo para la expansión de los términos con K_{22} y K_{12} que contienen β vemos que

$$-2K_{22}\beta^*\beta - 2K_{12}\alpha^*\beta = -2n(\beta^* + \overline{x}\alpha^*)\beta = -2n\overline{y}\beta \tag{9.96}$$

es decir

$$\beta^* + \alpha^*\overline{x} = \overline{y}. \tag{9.97}$$

Pero (9.95) y (9.97) forman exactamente el mismo sistema de ecuaciones que (9.82) y (9.84), por lo que de inmediato vemos que, felizmente,

$$\alpha^* = \hat{\alpha}, \tag{9.98a}$$

$$\beta^* = \hat{\beta}. \tag{9.98b}$$

Con esto podemos escribir nuestra distribución posterior en la forma

$$P(\boldsymbol{\theta}, \sigma|D) = \frac{1}{\eta_D\, \sigma^{n+1}} \exp\Big(-\frac{1}{2\sigma^2}\Big[\gamma + (\boldsymbol{\theta} - \boldsymbol{\mu})\mathbb{K}(\boldsymbol{\theta} - \boldsymbol{\mu})^\intercal\Big]\Big), \tag{9.99}$$

con $\boldsymbol{\theta} := (\alpha, \beta)$, $\boldsymbol{\mu} := (\hat{\alpha}, \hat{\beta})$ y

$$\mathbb{K} = \begin{pmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{pmatrix} = n \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}. \tag{9.100}$$

Con esto vemos que si σ es constante, la distribución de $\boldsymbol{\theta}$ es una normal multidimensional, como fue descrita en la **Sección 8.2**. Marginalizando $\boldsymbol{\theta}$ de una vez en (9.99) y usando la integral gaussiana multidimensional, tenemos que la distribución marginal del error σ es

$$P(\sigma|D) = \frac{2\pi}{\eta_D \sigma^{n-1}\sqrt{n}s_x} \exp\Big(-\frac{\gamma}{2\sigma^2}\Big), \tag{9.101}$$

donde hemos usado

$$\det \mathbb{K} = n(\overline{x^2} - \overline{x}^2) = ns_x^2. \tag{9.102}$$

El valor de γ lo obtenemos agrupando los términos constantes en (9.92),

$$\begin{aligned} n\overline{y^2} &= \gamma + K_{11}(\alpha^*)^2 + K_{22}(\beta^*)^2 + 2K_{12}\alpha^*\beta^* \\ &= \gamma + n\Big(\overline{x^2}\hat{\alpha}^2 + \hat{\beta}^2 + 2\overline{x}\hat{\alpha}\hat{\beta}\Big), \end{aligned} \tag{9.103}$$

de tal forma que, luego de un poco de álgebra, encontramos que

$$\gamma = ns_y^2(1 - \rho_{XY}^2). \tag{9.104}$$

Podemos calcular la constante de normalización η_D usando la integral tipo Gamma inversa,

$$\eta_D = \frac{2\pi}{\sqrt{n}s_x}\int_0^\infty \mathrm{d}\sigma \exp(-\frac{\gamma}{2\sigma^2})\sigma^{-(n-1)} = \frac{\pi}{\sqrt{n-1}s_x} 2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2} - 1\right)\gamma^{1-\frac{n}{2}} \tag{9.105}$$

Marginalizando β y α por separado en la distribución (9.99) obtenemos las distribuciones marginales

$$P(\alpha,\sigma|D) = \frac{\sqrt{2\pi}}{\eta_D \sigma^n} \exp\Big(- n s_x^2 \frac{(\alpha-\hat{\alpha})^2}{2\sigma^2} - \frac{\gamma}{2\sigma^2}\Big), \tag{9.106a}$$

$$P(\beta,\sigma|D) = \frac{\sqrt{2\pi}}{\eta_D \sigma^n \sqrt{\overline{x^2}}} \exp\Big(- n \left[\frac{s_x^2}{\overline{x^2}}\right] \frac{(\beta-\hat{\beta})^2}{2\sigma^2} - \frac{\gamma}{2\sigma^2}\Big), \tag{9.106b}$$

y dividiendo cada una de ellas por $P(\sigma|D)$ en (9.101), obtenemos las distribuciones para α y β dado σ como

$$P(\alpha|\sigma,D) = \frac{s_x\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\Big(- n s_x^2 \frac{(\alpha-\hat{\alpha})^2}{2\sigma^2}\Big), \tag{9.107}$$

$$P(\beta|\sigma,D) = \frac{s_x\sqrt{n}}{\sqrt{2\pi\overline{x^2}}\sigma} \exp\Big(- n \left[\frac{s_x^2}{\overline{x^2}}\right] \frac{(\beta-\hat{\beta})^2}{2\sigma^2}\Big), \tag{9.108}$$

de forma que las varianzas de α y β dado σ son

$$\big\langle(\delta\alpha)^2\big\rangle_{\sigma,D} = \frac{\sigma^2}{n{s_x}^2}, \tag{9.109a}$$

$$\big\langle(\delta\beta)^2\big\rangle_{\sigma,D} = \frac{\sigma^2\overline{x^2}}{n{s_x}^2}. \tag{9.109b}$$

El valor más probable de σ dados los datos está dado por

$$\frac{\partial}{\partial\sigma} \ln P(\sigma|D)\Big|_{\sigma=\hat{\sigma}} = \frac{\gamma}{(\hat{\sigma})^3} - \frac{n-1}{\hat{\sigma}} = 0 \tag{9.110}$$

es decir, $\hat{\sigma}^2 = \gamma/(n-1)$, o explícitamente

$$\boxed{\hat{\sigma}^2 = \frac{n}{n-1} s_y^2 (1-\rho_{XY}^2).} \tag{9.111}$$

Podemos calcular la expectación de σ^2 a partir de (9.101) como

$$\big\langle\sigma^2\big\rangle_D = \frac{\gamma}{n-4}, \tag{9.112}$$

con lo que la varianza de σ está dada por

$$\big\langle(\delta\sigma)^2\big\rangle_D = \gamma\left(\frac{1}{n-4} - \frac{1}{2}\left[\frac{\Gamma(\frac{n-1}{2}-1)}{\Gamma(\frac{n}{2}-1)}\right]^2\right), \tag{9.113}$$

la cual tiende a cero cuando $n \to \infty$ ya que

$$\begin{aligned}\frac{\Gamma(z-\frac{1}{2})}{\Gamma(z)} &\to \exp\Big((z-\frac{1}{2})\ln(z-\frac{1}{2}) - \cancel{z} + \frac{1}{2} - z\ln z + \cancel{z}\Big)\\ &\approx \exp\Big(-\frac{1}{2}(1+\ln z) + \frac{1}{2} + \cancel{z\ln z} - \cancel{z\ln z}\Big) \approx \frac{1}{\sqrt{z}} \to 0.\end{aligned}$$

Esto implica que las varianzas de α y β también van a cero cuando $n \to \infty$ y $\sigma \to \hat{\sigma}$, luego en ese límite la recta con $\hat{\alpha}$ y $\hat{\beta}$ que entrega el método

de mínimos cuadrados se vuelve la única recta admisible, y desaparece la incerteza.

De hecho, las incertezas exactas de α y β pueden obtenerse marginalizando σ^2 en (9.109) usando (9.112), con lo que finalmente obtenemos

$$\left\langle (\delta\alpha)^2 \right\rangle_D = \left(\frac{s_y}{s_x}\right) \frac{(1-\rho_{XY}^2)}{n-4}, \tag{9.114a}$$

$$\left\langle (\delta\beta)^2 \right\rangle_D = \left(\frac{s_y}{s_x}\right) \frac{(1-\rho_{XY}^2)\overline{x^2}}{n-4}. \tag{9.114b}$$

Vemos de inmediato que las incertezas se anulan en el caso de $\rho_{XY} = \pm 1$, es decir, cuando todos los puntos están perfectamente alineados. Por otro lado, marginalizando σ en (9.99) se tiene

$$P(\boldsymbol{\theta}|D) = \frac{\sqrt{n-1}(n-2)s_x}{2\pi\gamma}\left(1 + (\boldsymbol{\theta}-\boldsymbol{\mu})\frac{\mathbb{K}}{\gamma}(\boldsymbol{\theta}-\boldsymbol{\mu})^\intercal\right)^{-\frac{n}{2}} \tag{9.115}$$

que es un caso particular de la distribución t de Student. Notemos que, debido a que γ es proporcional a n según (9.104), en el límite $n \to \infty$ es posible aplicar la propiedad

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \exp(x) \tag{9.116}$$

para demostrar que

$$P(\boldsymbol{\theta}|D) \to \frac{1}{2\pi\sqrt{\det\boldsymbol{\Sigma}}}\exp\left(-\frac{1}{2}\left[(\boldsymbol{\theta}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})^\intercal\right]\right), \tag{9.117}$$

donde $\boldsymbol{\Sigma}$ es la matriz de covarianza entre α y β, tal que

$$\boldsymbol{\Sigma}^{-1} = \frac{\mathbb{K}}{ns_y^2(1-\rho_{XY}^2)} \tag{9.118}$$

y donde hemos usado (9.102) para reemplazar

$$\sqrt{\det\boldsymbol{\Sigma}} = \frac{s_y}{s_x}\sqrt{|1-\rho_{XY}^2|}. \tag{9.119}$$

► Mucho más se puede decir acerca de la regresión, dada la diversidad de modelos y suposiciones existentes más allá de lo visto acá. Para ejemplos prácticos a distintos niveles de complejidad se recomienda encarecidamente ver el libro de von der Linden, Dose y von Toussaint (2014) y el de Bailer-Jones (2017).

PROBLEMAS

Problema 9.1. *Muestre que (9.115) tiende a $P(\boldsymbol{\theta}, \sigma = \hat{\sigma}|D)$ según (9.99) con $\hat{\sigma}$ dado en (9.111) cuando $n \to \infty$.*

Problema 9.2. *Complete los pasos para ir de (9.103) a (9.104).*

Problema 9.3. *Escriba la función a minimizar, y el sistema de ecuaciones a resolver para obtener los parámetros $\hat{A}$ y $\hat{B}$ de un modelo lineal $y = A \cdot x + B + \varepsilon$ con un error de medición $\varepsilon \sim Gamma(k, \varepsilon_0)$. Considere k y ε_0 como constantes conocidas.*

Problema 9.4. *La energía en un sistema crítico está descrita por una distribución tipo ley de potencia,*

$$P(E|\epsilon_0, b) \propto \frac{1}{(\epsilon_0 + E)^b}, \tag{9.120}$$

con parámetros $\epsilon_0 > 0$ y $b > 1$. Se tienen n mediciones de energía independientes $(E_1, \ldots, E_n)$. Utilizando el prior *de Jeffreys para ϵ_0 y un* prior *constante para b, encuentre*

(a) La distribución posterior $P(\epsilon_0, b|E_1, \ldots E_n, \varnothing)$. ¿Son independientes los parámetros ϵ_0 y b en la distribución posterior?

(b) Las ecuaciones de máximo a posteriori *para ϵ_0 y b.*

(c) La distribución posterior marginal $P(\epsilon_0|E_1, \ldots, E_n, \varnothing)$.

Problema 9.5. *En una encuesta se intenta estimar la edad del habitante de mayor edad de la ciudad, llamémosla e_m. Se muestrean N personas obteniéndose edades $e_1, \ldots, e_N$ independientemente, las cuales son números reales no negativos. Considere un modelo en que todas las edades menores a e_m son equiprobables y es imposible tener una edad $e > e_m$. Es decir,*

$$P(e|e_m) \propto \Theta(e_m - e), \tag{9.121}$$

Use el prior *de Jeffreys para e_m.*

(a) Demuestre que la probabilidad posterior para e_m dadas las muestras, está dada por una distribución de Pareto,

$$P(e_m|e_1, \ldots, e_N) \propto \frac{\Theta(e_m - E)}{(e_m)^{\alpha+1}}. \tag{9.122}$$

¿Qué valores toman α y E?

(b) ¿Cuál es el valor más probable de e_m?

(c) Obtenga la distribución predictiva $P(e|e_1, \ldots, e_N)$ y demuestre que

$$P(e > max(e_1, \ldots, e_N)|e_1, \ldots, e_N) = \frac{1}{N+1}, \tag{9.123}$$

es decir, $e_m \to max(e_1, \ldots, e_N)$ en el límite $N \to \infty$, como es de esperar.

CAPÍTULO 10

Otras propiedades de la expectación

> You know my methods, Watson. There was not one of them which I did not apply to the inquiry.
>
> Sherlock Holmes, The Crooked Man

Ahora que hemos desarrollado el lenguaje de la inferencia a través de la definición de modelos probabilísticos, usando para ello probabilidades y densidades de probabilidad, podemos volver a conectar lo aprendido con la idea de expectación. Como veremos en este capítulo, el trabajar con expectaciones a veces resulta ser más directo que recurrir a probabilidades.

Recordemos que en general la expectación de una variable discreta X se define como

$$\langle f \rangle_I = \sum_{i=1}^{n} P(X = x_i|I) f(x_i) \tag{10.1}$$

donde $X \in \{x_1, \ldots, x_n\}$, mientras que la expectación de una o más variables continuas $\boldsymbol{X}$ en términos de su densidad de probabilidad puede escribirse como

$$\langle f \rangle_I = \int_a^b \mathrm{d}\boldsymbol{x}' P(\boldsymbol{X} = \boldsymbol{x}'|I) f(\boldsymbol{x}') \tag{10.2}$$

donde $X \in [a, b]$. Veremos ahora algunas consecuencias importantes de estas definiciones, en la forma de identidades entre expectaciones de cantidades. Pero primero debemos abordar un punto formal: mostraremos aquí que toda variable discreta puede también ser descrita por una densidad de probabilidad[(1)], y por lo tanto nos basta con demostrar los resultados futuros únicamente para variables continuas.

[(1)] Esta es una de las enseñanzas que nos deja la teoría de la medida: las sumas discretas pueden considerarse como integrales bajo una medida.

Para convencernos de que esto es así, demostraremos a continuación un simple lema que nos da la prescripción para definir la densidad de probabilidad asociada a una variable discreta.

Lema 10.1. *Si $X \in \{x_1, \ldots, x_n\}$ es una variable discreta con probabilidades*

$$p_i = P(X = x_i|I),$$

entonces la variable continua $\tilde{X}$ con densidad de probabilidad

$$P(\tilde{X} = x|I) = \sum_{i=1}^{n} \delta(x - x_i)\ p_i \tag{10.3}$$

es completamente equivalente a X, en el sentido de que para toda función $\omega(X)$,

$$\langle \omega(X) \rangle_I = \langle \omega(\tilde{X}) \rangle_I. \tag{10.4}$$

En resumen, a una distribución discreta le corresponde una densidad de probabilidad construida como una suma de deltas de Dirac, ponderadas por las probabilidades de cada valor que toma la variable discreta. Es claro que esta densidad está correctamente normalizada, ya que

$$\int_{-\infty}^{\infty} \mathrm{d}x P(\tilde{X} = x|I) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathrm{d}x \delta(x - x_i) p_i = \sum_{i=1}^{n} p_i = 1. \tag{10.5}$$

Demostración. Imponemos

$$\langle \omega(\tilde{X}) \rangle_I = \langle \omega(X) \rangle_I \quad \text{para todo } \omega \tag{10.6}$$

luego usando $\omega(X) = \delta(X - x)$, tenemos

$$\begin{aligned} P(\tilde{X} = x|I) &= \langle \delta(\tilde{X} - x) \rangle_I \\ &= \langle \delta(X - x) \rangle_I \\ &= \sum_{i=1}^{n} P(X = x_i|I)\delta(x_i - x) \end{aligned} \tag{10.7}$$

10.1 — Desigualdad de Jensen

El postulado de **Aditividad de la estimación** nos dice que si

$$f(X) = aX + b$$

con a, b constantes, entonces se cumple

$$\langle f(X) \rangle_I = \langle aX + b \rangle_I = a\langle X \rangle_I + b = f(\langle X \rangle_I).$$

Sin embargo, esto no se cumple en general para la expectación de una función $f(X)$ arbitraria, ésta no será igual a la función aplicada a la expectación de X,

$$f(\langle X \rangle_I) \neq \langle f \rangle_I. \tag{10.8}$$

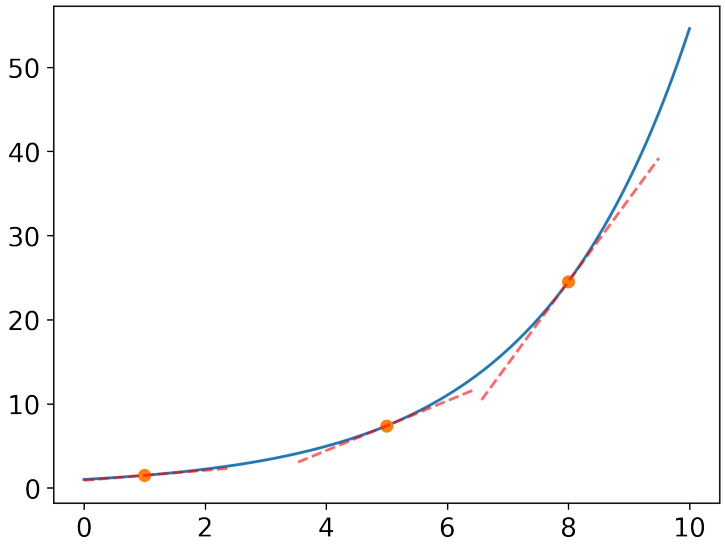


Figura 10.1: La función convexa $f(x) = \exp(\alpha x)$ con α=0.4. La expectación de una función convexa siempre será mayor o igual a la función evaluada en la expectación de su argumento, de acuerdo a la desigualdad de Jensen.

Para ver esto, escribamos la diferencia

$$S := \langle f \rangle_I - f(\langle X \rangle_I) \tag{10.9}$$

usando el postulado 5.1, de la forma

$$S = \Big\langle f(X) - f(E) \Big\rangle_I, \tag{10.10}$$

donde hemos definido

$$E := \langle X \rangle_I. \tag{10.11}$$

Si suponemos que f es una función diferenciable, podemos escribir

$$S = \Big\langle f(X) - f(E) \Big\rangle_I = \left\langle \int_E^X \mathrm{d}t f'(t) \right\rangle_I, \tag{10.12}$$

y si ahora suponemos además que es una función convexa, con lo que $f'(t)$ es monótonamente creciente con t, tendremos para el caso $X \geq E$ que

$$f'(t) \geq f'(E) \quad \text{para } t \geq E \tag{10.13}$$

y por lo tanto integrando desde E hasta X se cumplirá

$$\begin{aligned} &\int_E^X \mathrm{d}t \Big(\underbrace{f'(t) - f'(E)}_{\geq 0} \Big) \geq 0 \\ \hookrightarrow\ &\int_E^X \mathrm{d}t f'(t) \geq \int_E^X \mathrm{d}t f'(E) \\ \hookrightarrow\ &\int_E^X \mathrm{d}t f'(t) \geq f'(E)(X - E), \end{aligned} \tag{10.14}$$

es decir, hemos probado que

$$\int_E^X \mathrm{d}t f'(t) \geq f'(E)(X - E), \tag{10.15}$$

donde la igualdad ocurre para $X = E$. Por otro lado, para el caso $X < E$ se tiene

$$f'(t) \leq f'(E) \quad \text{para } t \leq E, \tag{10.16}$$

luego integrando desde X hasta E obtenemos

$$\begin{aligned} &\int_X^E \mathrm{d}t \Big(\underbrace{f'(t) - f'(E)}_{\leq 0} \Big) \leq 0 \\ \hookrightarrow\ &\int_X^E \mathrm{d}t f'(t) \leq \int_X^E \mathrm{d}t f'(E) \\ \hookrightarrow\ &\int_X^E \mathrm{d}t f'(t) \leq f'(E)(E - X). \end{aligned} \tag{10.17}$$

Invirtiendo esta última desigualdad recuperamos también la desigualdad en (**10.15**), y por lo tanto, independiente del orden entre X y E, podemos reemplazar (**10.15**) en (**10.12**) y obtener

$$\begin{aligned} S &= \left\langle \int_E^X \mathrm{d}t f'(t) \right\rangle_I \\ \text{usando (5.10) y (10.15)} \quad &\geq \Big\langle f'(E)(X - E) \Big\rangle_I \\ \text{usando (5.7)} \quad &= f'(E) \Big(\cancel{\langle X \rangle_I} - \cancel{E} \Big) \\ &= 0. \end{aligned} \tag{10.18}$$

Es decir, hemos demostrado que $S \geq 0$, o de otra forma, que

$$\big\langle f(X) \big\rangle_I \geq f(\langle X \rangle_I),$$

que es conocida como la *desigualdad de Jensen*.

Teorema 10.1 — Desigualdad de Jensen
Si $f(X)$ es una función convexa de X entonces se cumple

$$\big\langle f(X) \big\rangle_I \geq f(\langle X \rangle_I) \tag{10.19}$$

para cualquier estado de conocimiento I.

Notemos que hemos deducido esta desigualdad únicamente usando el contenido de los postulados 5.1, 5.2 y 5.3, ya que se trata siempre de expectaciones bajo el mismo estado de conocimiento I, y no hemos usado directamente el hecho que la expectación es una suma de valores ponderada por coeficientes no negativos. Adicionalmente, notemos que la demostración supone que $f(X)$ es diferenciable, aunque la desigualdad es válida para funciones convexas en general.

► Para más detalles, ver el libro de Cover y Thomas (2006) y el de MacKay (2003).

10.2 — Proyección de expectaciones

Es posible expresar el teorema de Bayes en términos de expectaciones en un estado de conocimiento I y en un estado (E, I), en la forma del siguiente teorema.

Teorema 10.2 — Propiedad de proyección de la función indicador
Si ω es una cantidad arbitraria, E es una proposición usada como evidencia e I un estado de conocimiento, entonces se cumple

$$\langle \omega \rangle_{E,I} = \frac{\langle \omega \, \mathrm{Q}(E) \rangle_I}{\langle \mathrm{Q}(E) \rangle_I}. \tag{10.20}$$

Lo que este teorema nos dice es que, si sabemos calcular expectaciones en el estado de conocimiento I, entonces automáticamente podemos calcular cualquier expectación en un nuevo estado (E, I) que incorpora nueva evidencia simplemente multiplicando por $\mathrm{Q}(E)$.

Al usar (10.20) diremos que la función indicador *proyecta*[(2)] la expectación de ω hacia un estado más restringido. Pensemos por ejemplo en $\omega = \omega(\boldsymbol{u})$, entonces $\langle \omega \rangle_I$ toma en cuenta todos los puntos $\boldsymbol{u}$ compatibles con I, mientras que la expectación proyectada $\langle \omega \rangle_{E,I}$ sólo considera entre los puntos compatibles con I aquellos tales que $E(\boldsymbol{u}) = \mathbb{T}$.

Para ver que este teorema es equivalente al teorema de Bayes, simplemente usamos $\omega = \mathrm{Q}(T)$,

$$\begin{aligned} \langle \mathrm{Q}(T) \rangle_{E,I} = P(T|E,I) &= \frac{\langle \mathrm{Q}(T)\mathrm{Q}(E) \rangle_I}{\mathrm{Q}(E) I} \\ &= \frac{P(T,E|I)}{P(E|I)} \\ &= \frac{P(T|I)P(E|T,I)}{P(E|I)}. \end{aligned} \tag{10.21}$$

Primero haremos la demostración de este teorema utilizando la forma usual del teorema de Bayes, junto a la regla de marginalización y la regla del producto.

Demostración. Comenzamos con la expectación dado (E, I) en el lado izquierdo,

$$\begin{aligned} \langle \omega \rangle_{E,I} &= \int d\omega P(\omega|E,I)\omega \\ &= \int d\omega \frac{P(\omega|I)P(E|\omega,I)}{P(E|I)}\omega \\ &= \frac{1}{P(E|I)} \int d\omega P(\omega|I) \Big(\omega P(E|\omega,I) \Big). \end{aligned} \tag{10.22}$$

Ahora escribamos $P(E|\omega, I)$ como expectación de la función indicador

(2) En el sentido matemático de eliminar componentes, como una proyección en dos dimensiones de una figura tridimensional.

$\epsilon := \mathrm{Q}(E)$, esto es

$$P(E|\omega, I) = \langle \mathrm{Q}(E) \rangle_{\omega,I} = \langle \epsilon \rangle_{\omega,I}, \tag{10.23}$$

con lo que tenemos

$$\begin{aligned}
\langle \omega \rangle_{E,I} &= \frac{1}{P(E|I)} \int d\omega P(\omega|I) \left(\omega \langle \epsilon \rangle_{\omega,I} \right) \\
&= \frac{1}{P(E|I)} \int d\omega P(\omega|I) \langle \omega\, \epsilon \rangle_{\omega,I} \\
&= \frac{1}{P(E|I)} \int d\omega P(\omega|I) \sum_{\epsilon} \int d\omega' P(\omega', \epsilon|\omega, I) \omega'\, \epsilon,
\end{aligned} \tag{10.24}$$

pero usando la regla del producto para separar

$$P(\omega', \epsilon|\omega, I) = P(\omega'|\omega, \epsilon, I) P(\epsilon|\omega, I)$$

y notando que

$$P(\omega'|\omega, \zeta) = \delta(\omega - \omega') \tag{10.25}$$

para cualquier estado de conocimiento ζ tenemos que

$$\begin{aligned}
\langle \omega \rangle_{E,I} &= \frac{1}{P(E|I)} \int d\omega P(\omega|I) \sum_{\epsilon} \int d\omega' \delta(\omega - \omega') P(\epsilon|\omega, I) \omega'\, \epsilon \\
&= \frac{1}{P(E|I)} \int d\omega \sum_{\epsilon} P(\omega|I) P(\epsilon|\omega, I) \omega\, \epsilon \\
&= \frac{1}{P(E|I)} \int d\omega \sum_{\epsilon} P(\omega, \epsilon|I) \omega\, \epsilon \\
&= \frac{\langle \omega\, \epsilon \rangle_I}{P(E|I)} = \frac{\langle \omega\, \mathrm{Q}(E) \rangle_I}{\langle \mathrm{Q}(E) \rangle_I}
\end{aligned} \tag{10.26}$$

Ahora veamos cómo la demostración es mucho más simple utilizando el postulado de **Doble estimación**. Comenzaremos escribiendo $\langle \omega\, \mathrm{Q}(E) \rangle_I$,

$$\langle \omega\, \mathrm{Q}(E) \rangle_I = \left\langle \langle \omega\, \epsilon \rangle_{\epsilon=\bullet,I} \right\rangle_I = \left\langle \bullet \langle \omega \rangle_{\epsilon=\bullet,I} \right\rangle_I, \tag{10.27}$$

pero aquí podemos ver, usando (4.26), que la función $z \mapsto z \langle \omega \rangle_{\epsilon=z,I}$ en el lado derecho de la última igualdad es idéntica a la función $z \mapsto z \langle \omega \rangle_{\epsilon=1,I}$. Por lo tanto,

$$\begin{aligned}
\langle \omega\, \mathrm{Q}(E) \rangle_I &= \left\langle \epsilon \langle \omega \rangle_{\epsilon=1,I} \right\rangle_I \\
&= \langle \omega \rangle_{\epsilon=1,I} \langle \epsilon \rangle_I \\
&= \langle \omega \rangle_{E,I} \langle \mathrm{Q}(E) \rangle_I.
\end{aligned} \tag{10.28}$$

Reemplazando E por la proposición $f = F$, con $f(X)$ una función de una

variable discreta $X \in \{x_1, \ldots, x_n\}$, obtenemos

$$\langle \omega \, \mathrm{Q}(f = F) \rangle_I = \langle \omega \rangle_{f=F,I} \, P(f = F|I). \tag{10.29}$$

Si sumamos en F a ambos lados vemos que el lado izquierdo se reduce a

$$\sum_F \langle \omega \, \mathrm{Q}(f = F) \rangle_I = \Big\langle \omega \underbrace{\sum_F \mathrm{Q}(f = F)}_{=1} \Big\rangle_I = \langle \omega \rangle_I, \tag{10.30}$$

mientras que en el lado derecho se obtiene

$$\sum_F P(f = F|I) \langle \omega \rangle_{f=F,I} = \Big\langle \langle \omega \rangle_{f=\bullet,I} \Big\rangle_I, \tag{10.31}$$

es decir, recuperamos el postulado de **Doble estimación**. El análogo continuo de (10.29) está dado por el siguiente teorema.

Teorema 10.3 — Propiedad de proyección de la delta de Dirac
Para dos cantidades arbitrarias ω y f, se cumple

$$\langle \omega \, \delta(f - F) \rangle_I = \langle \omega \rangle_{f=F,I} \, P(f = F|I), \tag{10.32}$$

donde f toma valores en un continuo y F es uno de esos valores.

La demostración usando la forma explícita de las expectaciones y el teorema de Bayes es como sigue.

Demostración.

$$\begin{aligned}
\langle \omega \, \delta(f - F) \rangle_I &= \int df' \int d\omega P(\omega, f = f'|I) \omega \, \delta(f' - F) \\
&= \int d\omega P(\omega, f = F|I) \omega \\
&= \int d\omega P(\omega|f = F, I) P(f = F|I) \omega \\
&= P(f = F|I) \int d\omega P(\omega|f = F, I) \omega \\
&= P(f = F|I) \langle \omega \rangle_{f=F,I}
\end{aligned} \tag{10.33}$$

10.3 — Desigualdad de Chebyshev

Como ejemplo de (10.20), demostraremos la *desigualdad de Chebyshev*,

$$P(|\boldsymbol{X}| \geq \alpha|I) \leq \frac{\langle g(|\boldsymbol{X}|) \rangle_I}{g(\alpha)} \tag{10.34}$$

con $g(\bullet)$ una función monótonamente creciente y para cualquier estado de conocimiento I.

Demostración. Comenzando desde la desigualdad

$$g(|\boldsymbol{X}|) \geq g(|\boldsymbol{X}|)\underbrace{\mathrm{Q}(|\boldsymbol{X}| \geq \alpha)}_{\leq 1} \tag{10.35}$$

y recordando el postulado de **Conservación del orden**, tenemos que se sigue la desigualdad

$$\langle g(|\boldsymbol{X}|)\rangle_I \geq \Big\langle g(|\boldsymbol{X}|)\mathrm{Q}(|\boldsymbol{X}| \geq \alpha)\Big\rangle_I \tag{10.36}$$

y si ahora usamos (10.20) para descomponer el lado derecho tenemos

$$\Big\langle g(|\boldsymbol{X}|)\mathrm{Q}(|\boldsymbol{X}| \geq \alpha)\Big\rangle_I = \langle g(|\boldsymbol{X}|)\rangle_{|\boldsymbol{X}|\geq\alpha,I}\, P(|\boldsymbol{X}| \geq \alpha|I). \tag{10.37}$$

Luego, teniendo en cuenta que

$$|\boldsymbol{X}| \geq \alpha \quad \rightarrow \quad g(|\boldsymbol{X}|) \geq g(\alpha) \tag{10.38}$$

el postulado de **Conservación del orden** asegura que

$$\langle g(|\boldsymbol{X}|)\rangle_{|\boldsymbol{X}|\geq\alpha,I} \geq g(\alpha) \tag{10.39}$$

y debe cumplirse que

$$\langle g(|\boldsymbol{X}|)\rangle_I \geq \Big\langle g(|\boldsymbol{X}|)\mathrm{Q}(|\boldsymbol{X}| \geq \alpha)\Big\rangle_I \geq g(\alpha)\, P(|\boldsymbol{X}| \geq \alpha|I) \tag{10.40}$$

10.4 — Derivada de una expectación

Trataremos el caso de variables continuas sin pérdida de generalidad. Al derivar la expectación de una cantidad $\omega(\boldsymbol{u},\lambda)$ respecto a λ se cumple lo siguiente,

$$\boxed{\frac{\partial}{\partial\lambda}\langle\omega\rangle_I = \left\langle\frac{\partial\omega}{\partial\lambda}\right\rangle_I}, \tag{10.41}$$

si el estado de conocimiento I no depende de λ. Simplemente intercambiando la derivada con la integral, tenemos

$$\begin{aligned}\frac{\partial}{\partial\lambda}\langle\omega\rangle_I &= \frac{\partial}{\partial\lambda}\int_\Omega \mathrm{d}\boldsymbol{u}P(\boldsymbol{u}|I)\omega(\boldsymbol{u},\lambda)\\ &= \int_\Omega \mathrm{d}\boldsymbol{u}P(\boldsymbol{u}|I)\Big(\frac{\partial\omega(\boldsymbol{u},\lambda)}{\partial\lambda}\Big)\\ &= \left\langle\frac{\partial\omega}{\partial\lambda}\right\rangle_I.\end{aligned} \tag{10.42}$$

Ahora nos corresponde generalizar el resultado en (10.41) para la derivada de una expectación respecto a un parámetro. Recordemos que teníamos

$$\frac{\partial}{\partial\lambda}\langle\omega\rangle_I = \left\langle\frac{\partial\omega}{\partial\lambda}\right\rangle_I$$

en el caso en que I no depende del parámetro λ. Tomando $\langle\omega\rangle_I$ como una función de dos argumentos, ω y I, es esperable que si $I = I(\lambda)$ exista una contribución adicional donde ω queda sin derivar y la derivada actúa sobre I. Esto es efectivamente así, y el resultado correcto está dado por una regla que por razones históricas[3] llamaremos el *teorema de fluctuación-disipación*.

Teorema 10.4 — Teorema de fluctuación-disipación
Si $\boldsymbol{X} \sim M(\theta, I)$ y $\omega = \omega(\boldsymbol{X}, \theta)$ es una función arbitraria diferenciable con respecto a θ, entonces se tiene

$$\frac{\partial}{\partial\theta}\langle\omega\rangle_{\theta,I} = \left\langle\frac{\partial\omega}{\partial\theta}\right\rangle_{\theta,I} + \left\langle\omega\frac{\partial}{\partial\theta}\ln P(\boldsymbol{X}|\theta,I)\right\rangle_{\theta,I}. \tag{10.43}$$

La utilidad de este teorema es que permite muchas veces calcular expectaciones sin emplear sumas o integrales de forma explícita. La demostración es directa si empleamos, sin pérdida de generalidad, la definición explícita de la expectación como una integral sobre variables continuas.

Demostración.

$$\begin{aligned}
\frac{\partial}{\partial\theta}\langle\omega\rangle_{\theta,I} &= \frac{\partial}{\partial\theta}\int_U \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}|\theta,I)\omega(\boldsymbol{x},\theta)\\
&= \int_U \mathrm{d}\boldsymbol{x}\frac{\partial}{\partial\theta}\Big(P(\boldsymbol{x}|\theta,I)\omega(\boldsymbol{x},\theta)\Big)\\
&= \int_U \mathrm{d}\boldsymbol{x}\Big(P(\boldsymbol{x}|\theta,I)\frac{\partial\omega(\boldsymbol{x},\theta)}{\partial\theta} + \omega(\boldsymbol{x},\theta)\frac{\partial}{\partial\theta}P(\boldsymbol{x}|\theta,I)\times\frac{P(\boldsymbol{x}|\theta,I)}{P(\boldsymbol{x}|\theta,I)}\Big)\\
&= \int_U \mathrm{d}\boldsymbol{x}\Big(P(\boldsymbol{x}|\theta,I)\frac{\partial\omega(\boldsymbol{x},\theta)}{\partial\theta} + \omega(\boldsymbol{x},\theta)P(\boldsymbol{x}|\theta,I)\frac{\partial}{\partial\theta}\ln P(\boldsymbol{x}|\theta,I)\Big)\\
&= \left\langle\frac{\partial\omega}{\partial\theta}\right\rangle_{\theta,I} + \left\langle\omega\frac{\partial}{\partial\theta}\ln P(\boldsymbol{X}|\theta,I)\right\rangle_{\theta,I} \qquad \blacksquare
\end{aligned} \tag{10.44}$$

Ejemplo 10.4.1. *Para la distribución de Poisson con $\omega = \omega(k,\lambda)$ se tiene*

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\langle\omega\rangle_\lambda &= \left\langle\frac{\partial\omega}{\partial\lambda}\right\rangle_\lambda + \left\langle\omega\frac{\partial}{\partial\lambda}\ln P(k|\lambda)\right\rangle_\lambda\\
&= \left\langle\frac{\partial\omega}{\partial\lambda}\right\rangle_\lambda + \left\langle\omega\left[\frac{k}{\lambda}-1\right]\right\rangle_\lambda.
\end{aligned} \tag{10.45}$$

Usando $\omega(k,\lambda) = 1$ tenemos

$$\frac{\partial}{\partial\lambda}\langle 1\rangle_\lambda = \left\langle\frac{\partial(1)}{\partial\lambda}\right\rangle_\lambda + \frac{\langle k\rangle_\lambda}{\lambda} - 1 \tag{10.46}$$

por lo tanto $\langle k\rangle_\lambda = \lambda$. Por otro lado, si usamos $\omega(k,\lambda) = k$ vemos que

$$\frac{\partial}{\partial\lambda}\langle k\rangle_\lambda = \left\langle\frac{\partial k}{\partial\lambda}\right\rangle_\lambda + \left\langle\frac{k^2}{\lambda} - k\right\rangle_\lambda, \tag{10.47}$$

[3] En física este teorema aparece en una forma particular al estudiar *fluctuaciones* térmicas de un sistema en equilibrio y su relación con coeficientes de respuesta lineal (*disipación*).

luego

$$1 = \frac{\langle k^2 \rangle_\lambda}{\lambda} - \lambda \tag{10.48}$$

y se tiene $\langle k^2 \rangle_\lambda = \lambda(\lambda+1)$. *La varianza de k estará dada por*

$$\langle (\delta k)^2 \rangle_\lambda = \langle k^2 \rangle_\lambda - \langle k \rangle_\lambda^2 = \lambda(\lambda+1) - \lambda^2 = \lambda. \tag{10.49}$$

Vemos que hemos obtenido tanto la media como la varianza de una distribución discreta sin calcular ninguna suma.

Notemos que la generalización para el caso en que tenemos m parámetros $(\theta_1, \ldots, \theta_m) = \boldsymbol{\theta}$ y derivamos respecto a uno de ellos, digamos θ_k, es directa. Simplemente definimos un estado de conocimiento de la forma

$$(\theta_k, \underbrace{\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_m, I}_{\text{parte de } I \text{ que no depende de } \theta_k}) = (\boldsymbol{\theta}, I)$$

y entonces podemos reescribir (10.43) como

$$\boxed{\frac{\partial}{\partial \theta_k} \langle \omega \rangle_{\boldsymbol{\theta}, I} = \left\langle \frac{\partial \omega}{\partial \theta_k} \right\rangle_{\boldsymbol{\theta}, I} + \left\langle \omega \frac{\partial}{\partial \theta_k} \ln P(\boldsymbol{X}|\boldsymbol{\theta}, I) \right\rangle_{\boldsymbol{\theta}, I}.} \tag{10.50}$$

Tomando m funciones $\omega_1, \ldots, \omega_m$ podemos armar un sistema de m ecuaciones usando (10.50) para cada ω_i, y escribir este sistema como la ecuación vectorial

$$\nabla_{\boldsymbol{\theta}} \cdot \langle \boldsymbol{\omega} \rangle_{\boldsymbol{\theta}, I} = \left\langle \nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{\omega} \right\rangle_{\boldsymbol{\theta}, I} + \left\langle \boldsymbol{\omega} \cdot \nabla_{\boldsymbol{\theta}} \ln P(\boldsymbol{X}|\boldsymbol{\theta}, I) \right\rangle_{\boldsymbol{\theta}, I}, \tag{10.51}$$

donde

$$\nabla_{\boldsymbol{\theta}} := \left(\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_m} \right), \tag{10.52}$$

y $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$.

Ejemplo 10.4.2. *Construyamos el teorema de fluctuación-disipación (10.50) para la distribución normal respecto al parámetro* μ. *Usando la derivada del logaritmo de la distribución respecto a* μ, *tenemos*

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln P(X|\mu, \sigma^2) &= \frac{\partial}{\partial \mu} \left(-\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \right) \\ &= \frac{1}{\sigma^2}(x - \mu) \end{aligned} \tag{10.53}$$

y por tanto podemos escribir

$$\frac{\partial}{\partial \mu} \langle \omega \rangle_{\mu, \sigma^2} = \left\langle \frac{\partial \omega}{\partial \mu} \right\rangle_{\mu, \sigma^2} + \frac{1}{\sigma^2} \left\langle \omega (X - \mu) \right\rangle_{\mu, \sigma^2}. \tag{10.54}$$

Si ahora escogemos $\omega = 1$ *obtenemos de inmediato*

$$\frac{\partial}{\partial \mu}(1) = \left\langle \frac{\partial (1)}{\partial \mu} \right\rangle_{\mu, \sigma^2} + \frac{1}{\sigma^2} \left\langle (X - \mu) \right\rangle_{\mu, \sigma^2} = \frac{\langle X \rangle_{\mu, \sigma^2} - \mu}{\sigma^2} \tag{10.55}$$

es decir, $\langle X \rangle_{\mu,\sigma^2} = \mu$*, mientras que usando* $\omega = (X - \mu)$ *tenemos*

$$\frac{\partial}{\partial \mu} \cancelto{0}{\langle X - \mu \rangle_{\mu,\sigma^2}} = \Big\langle \underbrace{\frac{\partial (X-\mu)}{\partial \mu}}_{=-1} \Big\rangle_{\mu,\sigma^2} + \frac{1}{\sigma^2} \Big\langle (X-\mu)^2 \Big\rangle_{\mu,\sigma^2}. \tag{10.56}$$

esto es,

$$\big\langle (X-\mu)^2 \big\rangle_{\mu,\sigma^2} = \sigma^2. \tag{10.57}$$

10.5 — Expectaciones vía integración por partes

No sólo podemos usar el teorema de fluctuación-disipación según (10.43) y (10.50) para establecer identidades entre expectaciones, sino que también existe una familia de teoremas que son consecuencias de aplicar integración por partes —en el caso de una variable— o, en el caso más general, el teorema de la divergencia.

En primer lugar, para una distribución $P(X = x|I) = p(x)$ de una variable continua $X \in \mathbb{R}$ tal que $p(\pm\infty) = 0$, usemos la integración por partes para resolver la expectación de la derivada de una función arbitraria $\omega(X)$ en términos de otras expectaciones. Escribimos

$$\begin{aligned}
\left\langle \frac{\mathrm{d}\omega(X)}{\mathrm{d}X} \right\rangle_I &= \int_{-\infty}^{\infty} \mathrm{d}x\, p(x) \frac{\mathrm{d}\omega(x)}{\mathrm{d}x} \\
&= \cancelto{0}{p(x)\omega(x)\Big|_{-\infty}^{\infty}} - \int_{-\infty}^{\infty} \mathrm{d}x\, \omega(x) \frac{\mathrm{d}p(x)}{\mathrm{d}x} \\
&= -\int_{-\infty}^{\infty} \mathrm{d}x\, \omega(x) p(x) \frac{\mathrm{d}}{\mathrm{d}x} \ln p(x) \\
&= -\left\langle \omega(X) \frac{\mathrm{d}}{\mathrm{d}X} \ln p(X) \right\rangle_I
\end{aligned} \tag{10.58}$$

por lo tanto tenemos nuestra primera versión del *teorema de variables conjugadas* (Davis y Gutiérrez 2012), cuyo nombre se entenderá mejor en el **Capítulo 12**.

Teorema 10.5 — Teorema de variables conjugadas
Si $X \sim I \in \mathbb{R}$ con distribución $p(x) := P(X = x|I)$ tal que $p(\pm\infty) = 0$, entonces para una función arbitraria diferenciable $\omega(X)$ se cumple

$$\left\langle \frac{\mathrm{d}\omega(X)}{\mathrm{d}X} \right\rangle_I + \left\langle \omega(X) \frac{\mathrm{d}}{\mathrm{d}X} \ln P(X|I) \right\rangle_I = 0. \tag{10.59}$$

Rápidamente podemos ver que la restricción de que $X \in \mathbb{R}$ con probabilidades que son cero en $\pm\infty$ puede ser reemplazada por $X \in [a, b]$ con

$$P(X = a|I) = P(X = b|I) = 0$$

y el teorema sigue siendo válido, como veremos en el siguiente ejemplo, donde también notamos que podemos reemplazar las derivadas totales respecto a X por derivadas parciales como explicaremos más adelante.

Ejemplo 10.5.1. *Consideremos la distribución beta,*

$$P(X = x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \tag{10.60}$$

y construyamos su ***Teorema de variables conjugadas****. Para ello, calculamos la derivada del logaritmo de la distribución*

$$\frac{\partial}{\partial x} \ln P(X = x|\alpha, \beta) = \frac{\alpha - 1}{x} - \frac{\beta - 1}{1 - x}, \tag{10.61}$$

y sustituimos en (10.59), obteniendo

$$\left\langle \frac{\partial \omega}{\partial X} \right\rangle_{\alpha,\beta} = \left\langle \omega \left(\frac{\beta - 1}{1 - X} - \frac{\alpha - 1}{X} \right) \right\rangle_{\alpha,\beta}. \tag{10.62}$$

Con la elección

$$\omega(x) = x(1 - x) \tag{10.63}$$

obtenemos

$$\begin{aligned} \langle 1 - 2X \rangle_{\alpha,\beta} &= \langle (X - 1)(\alpha - 1) + X(\beta - 1) \rangle_{\alpha,\beta} \\ \hookrightarrow \cancel{1} - 2\langle X \rangle_{\alpha,\beta} &= \cancel{1} - \alpha + (\alpha - 1 + \beta - 1)\langle X \rangle_{\alpha,\beta} \\ \hookrightarrow \alpha &= (\alpha + \beta)\langle X \rangle_{\alpha,\beta}, \end{aligned} \tag{10.64}$$

luego

$$\langle X \rangle_{\alpha,\beta} = \frac{\alpha}{\alpha + \beta}. \tag{10.65}$$

En el caso en que la probabilidad en los bordes $[a, b]$ sea distinta de cero, debemos generalizar nuestro resultado, pero afortunadamente para esto sólo necesitamos hacer uso de funciones indicador. Esto es, si nuestra variable $X \in [a, b]$ con densidad de probabilidad $P(X = x|I) = p(x)$, entonces la extendemos a $X \in \mathbb{R}$ haciendo que la densidad de probabilidad sólo sea distinta de cero en nuestro intervalo original. Esto es, ahora definimos

$$P(X = x|I) = \text{rect}(x; a, b)p(x) \qquad \text{para } x \in \mathbb{R}. \tag{10.66}$$

Reemplazando en (10.59) tenemos

$$\left\langle \frac{\mathrm{d}\omega(X)}{\mathrm{d}X} \right\rangle_I + \left\langle \omega(X) \frac{\mathrm{d}}{\mathrm{d}X} \Big(\ln p(X) + \ln \text{rect}(X; a, b) \Big) \right\rangle_I = 0, \tag{10.67}$$

pero usando (3.11), podemos reducir el término con la derivada de $\ln \text{rect}(\bullet; a, b)$ a

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln \text{rect}(x; a, b) = \frac{\delta(a - x) - \delta(b - x)}{\cancel{\text{rect}(x; a, b)}^{\,1}} = \delta(a - x) - \delta(b - x). \tag{10.68}$$

El teorema de variables conjugadas queda entonces, incluyendo los términos de borde, como

$$\begin{aligned}\left\langle \frac{\mathrm{d}\omega(X)}{\mathrm{d}X} \right\rangle_I + \left\langle \omega(X) \frac{\mathrm{d}}{\mathrm{d}X} \ln P(X|I) \right\rangle_I \\ = \left\langle \omega(X) \Big(\delta(b-x) - \delta(a-x) \Big) \right\rangle_I \\ = \omega(b) P(X=b|I) - \omega(a) P(X=a|I),\end{aligned} \tag{10.69}$$

que es completamente equivalente a conservar los términos de borde en (10.58). Para generalizar al caso de n dimensiones, usaremos un conjunto de variables continuas $\boldsymbol{X}$, un campo vectorial diferenciable

$$\boldsymbol{\omega}(\boldsymbol{X}) := \sum_{i=1}^{n} \omega_i(\boldsymbol{X}) \hat{\boldsymbol{e}}_i \tag{10.70}$$

y emplearemos el teorema de la divergencia para calcular la expectación de la divergencia de $\boldsymbol{\omega}$,

$$\begin{aligned}\left\langle \nabla \cdot \boldsymbol{\omega} \right\rangle_I &= \int_\Omega \mathrm{d}\boldsymbol{x} p(\boldsymbol{x}) \nabla \cdot \boldsymbol{\omega}(\boldsymbol{x}) \\ &= \cancelto{0}{\int_{\partial\Omega} \mathrm{d}s \boldsymbol{n}(\boldsymbol{x}) \cdot p(\boldsymbol{x}) \boldsymbol{\omega}(\boldsymbol{x})} - \int_\Omega \mathrm{d}\boldsymbol{x} \boldsymbol{\omega}(\boldsymbol{x}) \cdot \nabla p(\boldsymbol{x}) \\ &= -\int_\Omega \mathrm{d}\boldsymbol{x} \boldsymbol{\omega}(\boldsymbol{x}) p(\boldsymbol{x}) \cdot \nabla \ln p(\boldsymbol{x}) \\ &= -\left\langle \boldsymbol{\omega}(\boldsymbol{x}) \cdot \nabla \ln p(\boldsymbol{x}) \right\rangle_I\end{aligned} \tag{10.71}$$

donde hemos supuesto que $P(\boldsymbol{X}|I) = 0$ en el borde $\partial\Omega$ de la región Ω, condición equivalente a suponer probabilidad cero en $\pm\infty$ para una variable en $\mathbb{R}$. De esta forma el teorema de variables conjugadas en su versión vectorial es el siguiente.

Teorema 10.6 — Teorema vectorial de variables conjugadas
Si $\boldsymbol{X} \sim I \in \Omega$ con distribución $p(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|I)$ tal que $p(\boldsymbol{x}) = 0$ si $\boldsymbol{x} \in \partial\Omega$, entonces para un campo arbitrario diferenciable $\boldsymbol{\omega}(\boldsymbol{X})$ se cumple

$$\left\langle \nabla \cdot \boldsymbol{\omega}(\boldsymbol{X}) \right\rangle_I + \left\langle \boldsymbol{\omega}(\boldsymbol{X}) \cdot \nabla \ln P(\boldsymbol{X}|I) \right\rangle_I = 0. \tag{10.72}$$

Escogiendo un campo $\boldsymbol{\omega}$ tal que sólo tiene una componente distinta de cero, digamos

$$\omega_i(\boldsymbol{X}) = \delta(i,k)\omega(\boldsymbol{X}),$$

vemos que

$$\nabla \cdot \boldsymbol{\omega} \to \frac{\partial \omega}{\partial X_k}, \qquad \boldsymbol{\omega} \cdot \nabla \ln p \to \omega \frac{\partial}{\partial X_k} \ln p$$

y podemos escribir

$$\left\langle \frac{\partial \omega(\boldsymbol{X})}{\partial X_k} \right\rangle_I + \left\langle \omega(\boldsymbol{X}) \frac{\partial}{\partial X_k} \ln P(\boldsymbol{X}|I) \right\rangle_I = 0. \tag{10.73}$$

Si queremos incluir términos de borde en la versión vectorial, podemos definir un campo $B(\boldsymbol{x})$ tal que los puntos con $B(\boldsymbol{x}) < B_0$ coinciden con los puntos de la región Ω, y por tanto los puntos con $B(\boldsymbol{x}) = B_0$ serán los puntos del borde $\partial\Omega$. En ese caso extendemos nuestra densidad de probabilidad $P(\boldsymbol{X} = \boldsymbol{x}|I) = p(\boldsymbol{x})$ a

$$P(\boldsymbol{X} = \boldsymbol{x}|I) = p(\boldsymbol{x})\Theta(B_0 - B(\boldsymbol{x})), \tag{10.74}$$

y el teorema de variables conjugadas queda como

$$\langle \nabla \cdot \boldsymbol{\omega}(\boldsymbol{X}) \rangle_I + \Big\langle \boldsymbol{\omega}(\boldsymbol{X}) \cdot \nabla \Big(\ln p(\boldsymbol{X}) + \ln \Theta(B_0 - B(\boldsymbol{x})) \Big) \Big\rangle_I = 0. \tag{10.75}$$

Desarrollando la derivada de $\ln \Theta$ tenemos

$$\nabla \ln \Theta(B_0 - B(\boldsymbol{x})) = -\frac{\delta(B_0 - B(\boldsymbol{x}))\nabla B(\boldsymbol{x})}{\Theta(B_0 - B(\boldsymbol{x}))} = -\delta(B_0 - B(\boldsymbol{x}))\nabla B(\boldsymbol{x}) \tag{10.76}$$

y luego

$$\langle \nabla \cdot \boldsymbol{\omega} \rangle_I + \langle \boldsymbol{\omega} \cdot \nabla \ln P(\boldsymbol{X}|I) \rangle_I = \Big\langle (\boldsymbol{\omega} \cdot \nabla B)\delta(B_0 - B) \Big\rangle_I, \tag{10.77}$$

que puede escribirse de forma más legible usando la propiedad (10.32) de proyección de la delta de Dirac como

$$\langle \nabla \cdot \boldsymbol{\omega} \rangle_I + \langle \boldsymbol{\omega} \cdot \nabla \ln P(\boldsymbol{X}|I) \rangle_I = \Big\langle \boldsymbol{\omega} \cdot \nabla B \Big\rangle_{B=B_0,I} \cdot P(B = B_0|I). \tag{10.78}$$

Para poder ver con mayor claridad que el lado derecho de (10.77) coincide con el término de borde que despreciamos en la demostración en (10.71), utilizamos la propiedad (3.53) de composición de la delta de Dirac para escribir

$$\delta(B_0 - B(\boldsymbol{x})) = \sum_{\boldsymbol{x}^*} \frac{\delta(\boldsymbol{x} - \boldsymbol{x}^*)}{|\nabla B|}, \tag{10.79}$$

donde $\boldsymbol{x}^*$ son las soluciones de $B(\boldsymbol{x}) = B_0$, es decir, los puntos del borde, gracias a lo cual podemos reemplazar

$$\sum_{\boldsymbol{x}^*} \to \int_{\partial\Omega} \mathrm{d}s$$

y entonces escribir

$$\begin{aligned} \Big\langle (\boldsymbol{\omega} \cdot \nabla B)\delta(B_0 - B) \Big\rangle_I &= \int_{\partial\Omega} \mathrm{d}s \Big\langle \delta(\boldsymbol{X} - \boldsymbol{x})\boldsymbol{\omega}(\boldsymbol{X}) \cdot \underbrace{\Big(\frac{\nabla B}{|\nabla B|} \Big)}_{=\boldsymbol{n}} \Big\rangle_I \\ &= \int_{\partial\Omega} \mathrm{d}s \boldsymbol{\omega}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}) \big\langle \delta(\boldsymbol{X} - \boldsymbol{x}) \big\rangle_I \\ &= \int_{\partial\Omega} \mathrm{d}s \boldsymbol{\omega}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x})\, p(\boldsymbol{x}). \end{aligned} \tag{10.80}$$

► Para más detalles sobre el uso de los teoremas (10.50) y (10.72), ver Davis y Gutiérrez (2012) y Davis y Gutiérrez (2016).

PROBLEMAS

Problema 10.1. *¿Cuál es mayor, la expectación $\langle V\rangle_I$ del volumen de una esfera, o el volumen de una esfera que tiene el radio esperado $\langle R\rangle_I$? Considere*

$$V(R) = \frac{4\pi}{3}R^3.$$

Problema 10.2. *Para una variable $A > 0$ y una proposición E tal que $E \Rightarrow A \geq A_0$, demuestre la desigualdad*

$$P(E|I) \leq \frac{\langle A\rangle_I}{A_0}, \tag{10.81}$$

más general que (10.34).

Problema 10.3. *Si definimos el* promedio aritmético *de una función $f(X)$ sobre un conjunto de n valores $\{x_1, \ldots, x_n\}$ como*

$$\overline{f} := \frac{1}{n}\sum_{i=1}^{n} f(x_i), \tag{10.82}$$

encuentre el estado de conocimiento D, es decir, encuentre $P(X = x|D)$, tal que

$$\langle\omega\rangle_D = \overline{\omega} \tag{10.83}$$

para toda función continua $\omega(X)$. ¿Cómo interpretaría dicho estado de conocimiento?

Problema 10.4. *Usando el teorema de variables conjugadas, deduzca una relación de recurrencia para los momentos de la distribución Beta, esto es, $\langle x^m\rangle_{\alpha,\beta}$ en función de $\langle x^{m-1}\rangle_{\alpha,\beta}$.*

Problema 10.5. *Complete el Ejemplo 10.5.1 para calcular la varianza.*

Problema 10.6. *Demuestre que para $x \sim G(k,\theta)$ y una función $g(x)$ cualquiera se cumple*

$$\theta\frac{\partial}{\partial\theta}\langle g\rangle_{k,\theta} = \frac{\langle g\ x\rangle_{k,\theta}}{\theta} - k\langle g\rangle_{k,\theta}. \tag{10.84}$$

Usando (10.84) demuestre que $\langle x\rangle_{k,\theta} = k\theta$ y que $\langle(\delta x)^2\rangle_{k,\theta} = k\theta^2$.

Problema 10.7. *Utilizando integración por partes, demuestre que si $x \sim Gamma(k,\theta)$ se cumple para una función $g(x)$ cualquiera que*

$$\left\langle\frac{\mathrm{d}g}{\mathrm{d}x}\right\rangle_{k,\theta} = \frac{1}{\theta}\langle g\rangle_{k,\theta} + (1-k)\left\langle\frac{g}{x}\right\rangle_{k,\theta}. \tag{10.85}$$

Usando (10.85) encuentre una relación de recurrencia entre $\langle x^m\rangle_{k,\theta}$ y $\langle x^{m-1}\rangle_{k,\theta}$.

Problema 10.8. *Muestre que para una región Ω arbitraria con borde $\partial\Omega$ la generalización de (10.77) es*

$$\Big\langle\nabla\cdot\boldsymbol{\omega}\Big\rangle_I + \Big\langle\boldsymbol{\omega}\cdot\nabla\ln P(\boldsymbol{X}|I)\Big\rangle_I = -\Big\langle\boldsymbol{\omega}\cdot\nabla\mathrm{Q}(\boldsymbol{X}\in\Omega)\Big\rangle_I, \tag{10.86}$$

Problema 10.9. *Si $x \sim Gamma(k, \theta)$, determine k y θ en función de*

$$\begin{aligned} x_0 &= \langle x \rangle_{k,\theta}, \\ L_0 &= \langle \ln x \rangle_{k,\theta}, \\ \alpha &= \langle x \ln x \rangle_{k,\theta}. \end{aligned}$$

CAPÍTULO 11

Comparación de modelos

> All models are wrong, but some models are useful.
>
> George E. P. Box

Una vez que hemos hecho uso del teorema de Bayes para realizar inferencia, y explorado la manera de realizar estimación de los parámetros de un modelo, queda pendiente otro de los problemas comunes en inferencia, el de decidir entre dos modelos M_1 y M_2 cuando ambos son enfrentados a un conjunto de datos D observados. ¿Qué criterio deberíamos usar para decidir cuál de los dos modelos favorecer a la luz de los datos?

11.1 — El factor de Bayes

Dado todo lo aprendido en el **Capítulo 6**, sabemos que podemos asignar probabilidades a los modelos, y lo natural para nuestro problema será considerar las probabilidades $P(M_1|D, I_0)$ y $P(M_2|D, I_0)$ de los modelos M_1 y M_2 considerando los datos D y la información previa I_0. El teorema de Bayes nos entrega

$$P(M_1|D, I_0) = \frac{P(M_1|I_0)P(D|M_1, I_0)}{P(D|I_0)}, \tag{11.1a}$$

$$P(M_2|D, I_0) = \frac{P(M_2|I_0)P(D|M_2, I_0)}{P(D|I_0)}, \tag{11.1b}$$

y al formar el cuociente entre $P(M_1|D, I_0)$ y $P(M_2|D, I_0)$ vemos que la probabilidad previa de los datos se cancela,

$$\begin{aligned}\frac{P(M_1|D, I_0)}{P(M_2|D, I_0)} &= \frac{P(M_1|I_0)P(D|M_1, I_0)}{\cancel{P(D|I_0)}} \frac{\cancel{P(D|I_0)}}{P(M_2|I_0)P(D|M_2, I_0)} \\ &= \frac{P(M_1|I_0)}{P(M_2|I_0)} \frac{P(D|M_1, I_0)}{P(D|M_2, I_0)}\end{aligned} \tag{11.2}$$

esto es,

$$\boxed{\frac{P(M_1|D, I_0)}{P(M_2|D, I_0)} = \left[\frac{P(M_1|I_0)}{P(M_2|I_0)}\right] K(M_1, M_2; D)} \tag{11.3}$$

donde en corchetes escribimos el cuociente previo entre las probabilidades de M_1 y M_2, y el factor $K(M_1, M_2; D)$ es conocido como el *factor de Bayes*.

Definición 11.1 — Factor de Bayes
Definiremos el factor de Bayes entre el modelo M_1 y el modelo M_2 para los datos D como el cuociente

$$K(M_1, M_2; D) := \frac{P(D|M_1, I_0)}{P(D|M_2, I_0)}. \tag{11.4}$$

De la misma manera que el cuociente $R(T; E)$ en (6.31) nos indica si una evidencia E favorece o perjudica a una hipótesis, el factor de Bayes es la cantidad que nos indica si un conjunto de datos D favorece a M_1 o a M_2. Si el modelo M no depende de parámetros, el cálculo de $P(D|M, I_0)$ es directo, mientras que si $M = M(\boldsymbol{\theta})$ entonces debemos usar la regla de marginalización para eliminar dichos parámetros, y para esto necesitamos un *prior* $P(\boldsymbol{\theta}|I_0)$. Obtenemos entonces,

$$P(D|M, I_0) = \int d\boldsymbol{\theta} P(D, \boldsymbol{\theta}|M, I_0) = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}|I_0) P(D|\boldsymbol{\theta}, M). \tag{11.5}$$

Incidentalmente, la probabilidad de los datos según nuestro modelo $M(\boldsymbol{\theta})$ es precisamente el denominador del teorema de Bayes en un problema de estimación de los parámetros $\boldsymbol{\theta}$,

$$P(\boldsymbol{\theta}|D, M, I_0) = \frac{P(\boldsymbol{\theta}|I_0) \cdot P(D|\boldsymbol{\theta}, M)}{P(D|M, I_0)}. \tag{11.6}$$

Si D consiste de n observaciones $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ independientes que siguen el modelo $P(\boldsymbol{x}|\boldsymbol{\theta}, M)$ entonces se tiene

$$P(D|\boldsymbol{\theta}, M) = \prod_{i=1}^{n} P(\boldsymbol{x}_i|\boldsymbol{\theta}, M). \tag{11.7}$$

Ejemplo 11.1.1. *Para una variable positiva X de la cual tenemos observaciones $D = (x_1, \ldots, x_n)$ proponemos dos modelos. El primero es uniforme con un valor máximo L,*

$$P(X = x|M_1, L) = \frac{\Theta(L - x)}{L}, \tag{11.8}$$

y un prior de Jeffreys

$$P(L|I_0) = \frac{a_0}{L}, \tag{11.9}$$

donde dejaremos sin especificar la constante a_0. El segundo es un modelo exponencial,

$$P(X = x|M_2, \lambda) = \lambda \exp(-\lambda x), \tag{11.10}$$

con un prior de Jeffreys

$$P(\lambda|I_0) = \frac{b_0}{\lambda}, \tag{11.11}$$

donde tampoco especificaremos b_0. ¿A cuál modelo favorecen los datos?

Solución. *Para el modelo M_1 tenemos*

$$\begin{aligned} P(D|M_1, I_0) &= \int_0^\infty \mathrm{d}L P(L|I_0) \prod_{i=1}^n \frac{\Theta(L - x_i)}{L} \\ &= a_0 \int_0^\infty \mathrm{d}L \frac{1}{L^{n+1}} \prod_{i=1}^n \Theta(L - x_i), \end{aligned} \tag{11.12}$$

y aquí ocupamos la propiedad

$$\begin{aligned} \prod_{i=1}^n \Theta(L - x_i) &= \mathrm{Q}((L \geq x_1) \wedge \ldots (L \geq x_n)) \\ &= \mathrm{Q}(L \geq L^*) = \Theta(L - L^*), \end{aligned} \tag{11.13}$$

donde $L^* := \mathrm{máx}(x_1, \ldots, x_n)$ es el máximo de los valores observados. Con este resultado podemos obtener

$$P(D|M_1, I_0) = a_0 \int_{L^*}^\infty \mathrm{d}L \frac{1}{L^{n+1}} = \frac{a_0}{n(L^*)^n}. \tag{11.14}$$

Por otro lado, para el modelo M_2 tenemos

$$\begin{aligned} P(D|M_2, I_0) &= \int_0^\infty \mathrm{d}\lambda P(\lambda|I_0) \prod_{i=1}^n \Big(\lambda \exp(-\lambda x_i)\Big) \\ &= b_0 \int_0^\infty \mathrm{d}\lambda \lambda^{n-1} \exp(-\lambda n \overline{x}) \\ &= b_0 (n-1)! (n\overline{x})^{-n}, \end{aligned} \tag{11.15}$$

y podemos entonces escribir el factor de Bayes como

$$K(M_1, M_2; D) = \frac{P(D|M_1)}{P(D|M_2)} = \frac{a_0 (n\overline{x})^n}{b_0\, n(n-1)!(L^*)^n} = \frac{a_0}{b_0\, n!} \Big(\frac{n\overline{x}}{L^*}\Big)^n. \tag{11.16}$$

Para n finito no podemos evaluar $K(M_1, M_2; D)$ debido a que las constantes a_0 y b_0 no han sido especificadas, y esto es una consecuencia de utilizar priors no informativos. Sin embargo, en el límite $n \to \infty$ podemos usar la aproximación de Stirling (3.197) para transformar

$$\ln K(M_1, M_2; D) = \ln(a_0/b_0) - \ln(n!) + n \ln n + n \ln \left(\frac{\overline{x}}{L^*}\right) \tag{11.17}$$

en la forma más manejable

$$\ln K(M_1, M_2; D) \approx n \left(\ln \left(\frac{\overline{x}}{L^*}\right) - 1\right), \tag{11.18}$$

ya que $\ln(a_0/b_0)$ se puede despreciar frente a los términos que crecen con n. Esto es,

$$K(M_1, M_2; D) \approx \exp\left(n \left[\ln \left(\frac{\overline{x}}{L^*}\right) - 1\right]\right). \tag{11.19}$$

Más aún, aquí recordamos que L^* es el máximo de los valores observados, luego el promedio aritmético $\overline{x}$ es menor o igual que dicho máximo y se debe cumplir

$$\ln\left(\frac{\overline{x}}{L^*}\right) \leq 0, \tag{11.20}$$

con lo que determinamos que

$$\lim_{n\to\infty} K(M_1, M_2; D) = 0. \tag{11.21}$$

Es decir, el modelo exponencial (M_2) debe ser preferido para un número muy grande de observaciones.

11.2 — Criterio de información bayesiano (BIC)

Podemos obtener una aproximación asintóticas para $P(D|M, I_0)$ en el límite $n \to \infty$ de forma general. En primer lugar, si D consiste en n observaciones independientes, tenemos

$$\begin{aligned} P(D|M, I_0) &= \int d\boldsymbol{\theta} P(\boldsymbol{\theta}|I_0) \prod_{i=1}^{n} P(\boldsymbol{x}_i|\boldsymbol{\theta}, M) \\ &= \int d\boldsymbol{\theta} \exp\left(\sum_{i=1}^{n} \ln P(\boldsymbol{x}_i|\boldsymbol{\theta}, M) + \ln P(\boldsymbol{\theta}|I_0)\right) \\ &= \int d\boldsymbol{\theta} \exp\left(nL(\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta}|I_0)\right), \end{aligned} \tag{11.22}$$

donde

$$L(\boldsymbol{\theta}) := \frac{\mathcal{L}_D(\boldsymbol{\theta})}{n} = \frac{1}{n}\sum_{i=1}^{n} \ln P(\boldsymbol{x}_i|\boldsymbol{\theta}, M) \tag{11.23}$$

es la función log-verosimilitud del modelo M dividida en el número n de observaciones. Para n suficientemente grande, el término $\ln P(\boldsymbol{\theta}|I_0)$ es despreciable y podemos aproximar

$$P(D|M, I_0) = \int d\boldsymbol{\theta} \exp\left(nL(\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta}|I_0)\right) \approx \int d\boldsymbol{\theta} \exp\left(nL(\boldsymbol{\theta})\right). \tag{11.24}$$

Más aún, como $\exp(nL(\boldsymbol{\theta}))$ se concentra en torno al conjunto de parámetros de máxima verosimilitud $\hat{\boldsymbol{\theta}}$, por la aproximación de Laplace (3.189) y usando la propiedad $\det(n\mathbb{H}) = n^m \det \mathbb{H}$, se tiene

$$P(D|M, I_0) \approx \int \mathrm{d}\boldsymbol{\theta} \exp\left(-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\intercal}\mathbb{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right) = \frac{\exp(nL(\hat{\boldsymbol{\theta}}))\sqrt{(2\pi)^m}}{\sqrt{n^m \det \mathbb{H}(\hat{\boldsymbol{\theta}})}}, \tag{11.25}$$

con m el número de parámetros y $H_{ij}(\boldsymbol{\theta}) := -\dfrac{\partial^2 L(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}$.

En este resultado asintótico se ha perdido toda dependencia con el *prior* $P(\boldsymbol{\theta}|I_0)$, y es más, podemos escribir $P(D|M, I_0)$ como

$$P(D|M, I_0) \approx \exp\left(\mathcal{L}_D(\hat{\theta}) - \frac{m}{2}\ln n + \left[\frac{m}{2}\ln(2\pi) - \tfrac{1}{2}\ln\det \mathbb{H}(\hat{\boldsymbol{\theta}})\right]\right) \tag{11.26}$$

y reconociendo que el término en corchetes no crece con n, tenemos

$$P(D|M, I_0) \approx \exp(-2\,\mathrm{BIC}) \tag{11.27}$$

donde hemos definido BIC como el *criterio de información bayesiano* (*Bayesian information criterion*) (Konishi y Kitagawa 2008, pág. 211).

Definición 11.2 — Criterio de información bayesiano (BIC)

Para un modelo de m parámetros se define el BIC como

$$\mathrm{BIC} := m \ln n - 2\mathcal{L}_D(\hat{\boldsymbol{\theta}}), \tag{11.28}$$

de forma que, al comparar dos modelos para un conjunto de n observaciones, deberíamos preferir el modelo con menor BIC.

Ejemplo 11.2.1. *Calculemos el BIC para un modelo gamma,*

$$P(X = x|k, \theta) = \frac{\exp(-x/\theta)x^{k-1}}{\Gamma(k)\theta^k}. \tag{11.29}$$

Como este modelo tiene 2 parámetros libres, k y θ, tendremos $m = 2$ y sólo necesitamos calcular la función verosimilitud. Ésta estará dada por

$$\begin{aligned} P(D|k, \theta) &= \prod_{i=1}^{n} \frac{\exp(-x_i/\theta)x_i^{k-1}}{\Gamma(k)\theta^k} \\ &= \exp\left(-n\left[\frac{\overline{x}}{\theta} - (k-1)\overline{\ln x} - \ln\Gamma(k) - k\ln\theta\right]\right) \end{aligned} \tag{11.30}$$

y su máximo ocurre para los parámetros $\hat{k}$ y $\hat{\theta}$ dados por el sistema de ecuaciones

$$\left(\frac{\partial}{\partial k}\ln P(D|k,\theta)\right)_{k=\hat{k},\theta=\hat{\theta}} = n\left(\overline{\ln x} - \psi(\hat{k}) - \ln\hat{\theta}\right) = 0, \tag{11.31a}$$

$$\left(\frac{\partial}{\partial \theta}\ln P(D|k,\theta)\right)_{k=\hat{k},\theta=\hat{\theta}} = \frac{n}{\hat{\theta}}\left(\frac{\overline{x}}{\hat{\theta}} - \hat{k}\right) = 0, \tag{11.31b}$$

con solución

$$\hat{k}\hat{\theta} = \overline{x}, \tag{11.32a}$$

$$\overline{\ln x} - \ln\hat{\theta} = \psi(\hat{k}). \tag{11.32b}$$

La función log-verosimilitud evaluada en $(\hat{k}, \hat{\theta})$ entonces se reduce a

$$\mathcal{L}_D(\hat{k}, \hat{\theta}) = -n\left(\hat{k} + \overline{\ln x} - \ln\Gamma(\hat{k}) - \hat{k}\psi(\hat{k})\right), \tag{11.33}$$

y por tanto tendremos

$$BIC(\hat{k}) = 2\ln n + 2n\left(\hat{k} + \overline{\ln x} - \ln\Gamma(\hat{k}) - \hat{k}\psi(\hat{k})\right) \tag{11.34}$$

que, curiosamente, no depende del parámetro de escala $\hat{\theta}$.

11.3 — La divergencia de Kullback-Leibler

Veamos el problema de comparación de modelos desde otro ángulo. Nuevamente supongamos que tenemos una variable $\boldsymbol{X}$, n observaciones independientes $D = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ y dos modelos, M y M'. Sin embargo, ahora sabemos que la variable $\boldsymbol{X}$ es correctamente descrita por el modelo M, y nos preguntamos por la magnitud del error que cometeríamos al reemplazar el modelo M, que es complicado de evaluar, por el modelo M' que es más tratable.
Escribimos el factor de Bayes como

$$\frac{P(D|M)}{P(D|M')} = \frac{\prod_{i=1}^{n} P(\boldsymbol{x}_i|M)}{\prod_{i=1}^{n} P(\boldsymbol{x}_i|M')} = \exp\left(\sum_{i=1}^{n} \ln \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}\right) \tag{11.35}$$

donde hemos definido $p(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|M)$ y $q(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|M')$ por simplicidad de notación. En el límite $n \to \infty$ y dado que $\boldsymbol{X} \sim M$, podemos usar la ley de los grandes números para reescribir

$$\lim_{n\to\infty}\left[\frac{1}{n}\ln\frac{P(D|M)}{P(D|M')}\right] = \lim_{n\to\infty}\overline{\ln\left(\frac{p}{q}\right)} = \left\langle \ln\left(\frac{p}{q}\right)\right\rangle_M . \tag{11.36}$$

El lado derecho de la última igualdad nos lleva a definir la divergencia de Kullback-Leibler.

Definición 11.3 — Divergencia de Kullback-Leibler

$$D_{KL}(p||q) := \int d\boldsymbol{x} p(\boldsymbol{x}) \ln\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right]. \tag{11.37}$$

Con esta definición podemos escribir el importante resultado,

$$K(M, M'; D) = \frac{P(D|M)}{P(D|M')} = \exp\Big(n\, D_{KL}(p||q)\Big). \tag{11.38}$$

Ejemplo 11.3.1. *La divergencia de Kullback-Leibler entre dos distribuciones exponenciales con parámetros λ y λ_0 es*

$$\begin{aligned} D_{KL}(\lambda||\lambda_0) &= \left\langle \ln\left(\frac{\lambda \exp(-\lambda X)}{\lambda_0 \exp(-\lambda_0 X)}\right)\right\rangle_\lambda \\ &= \ln\left(\frac{\lambda}{\lambda_0}\right) + (\lambda_0 - \lambda)\underbrace{\langle X\rangle_\lambda}_{=1/\lambda} \\ &= \ln\left(\frac{\lambda}{\lambda_0}\right) + \left(\frac{\lambda_0}{\lambda} - 1\right), \end{aligned} \tag{11.39}$$

cuya gráfica se muestra en la Figura 11.1. Vemos que $D_{KL}(\lambda||\lambda_0)$ siempre es positiva, y es cero únicamente cuando $\lambda = \lambda_0$.

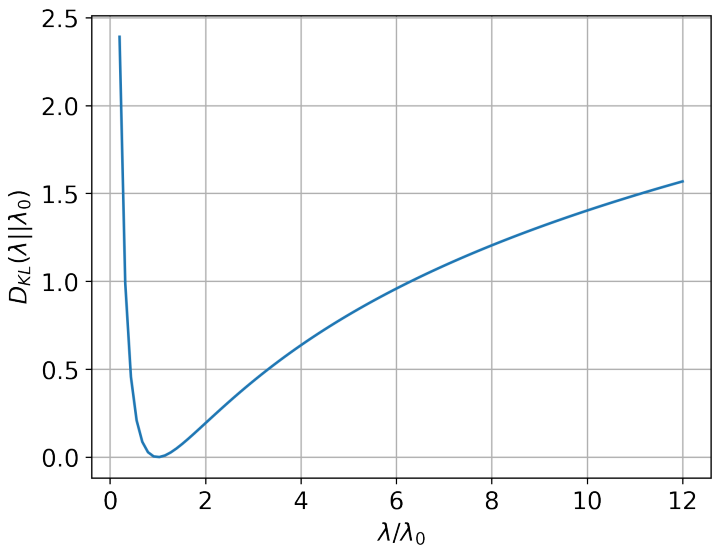


Figura 11.1: Divergencia de Kullback-Leibler entre dos distribuciones exponenciales, de acuerdo a (11.39).

Este resultado, que la divergencia de Kullback-Leibler no es negativa, se conoce como la *desigualdad de Gibbs*, y podemos demostrarla utilizando la desigualdad de Jensen (10.19).

Teorema 11.1 — Desigualdad de Gibbs

$$D_{KL}(p||q) = \int d\boldsymbol{x} p(\boldsymbol{x}) \ln\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] \geq 0. \tag{11.40}$$

Demostración. Como la función $\ln(\bullet)$ es cóncava (esto es, su derivada es monótonamente decreciente), se tiene que $-\ln(\bullet)$ es convexa, y usando la desigualdad de Jensen,

$$\begin{aligned} \langle -\ln X\rangle_I &\geq -\ln\langle X\rangle_I \\ \hookrightarrow \langle \ln X\rangle_I &\leq \ln\langle X\rangle_I \end{aligned} \tag{11.41}$$

y usando $X := q(\boldsymbol{X})/p(\boldsymbol{X})$, transformamos esta desigualdad en

$$\left\langle \ln\left(\frac{q}{p}\right)\right\rangle_p \leq \ln\left\langle\left(\frac{q}{p}\right)\right\rangle_p = \ln\left(\int d\boldsymbol{x} q(\boldsymbol{x})\right) = 0, \tag{11.42}$$

(la integral $\int d\boldsymbol{x} q(\boldsymbol{x})$ es igual a 1)

por lo tanto,

$$\left\langle \ln\left(\frac{p}{q}\right)\right\rangle_p = \int d\boldsymbol{x} p(\boldsymbol{x}) \ln\left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] \geq 0 \tag{11.43}$$

Con esta desigualdad, suponiendo $P(M|I_0) = P(M'|I_0)$, podemos verificar que

$$P(M'|D) \leq P(M|D), \tag{11.44}$$

consistente con nuestra suposición inicial de que M es el modelo «correcto».

Podemos interpretar la divergencia de Kullback-Leibler como una medida de qué tan distintas son dos distribuciones de probabilidad, aunque teniendo en cuenta que, según (11.38), D_{KL} es asimétrica, esto es,

$$D_{KL}(p||q) \neq D_{KL}(q||p) \tag{11.45}$$

a menos que $M = M'$, en cuyo caso $D_{KL}(p||p) = 0$. Otra consecuencia directa de (11.38) es que si queremos buscar un buen sustituto del modelo M, lo que debemos hacer es hacer crecer $P(D|M')$ hasta que se acerque por abajo lo más posible a $P(D|M)$ —ya que nunca podremos superarlo— y esto es equivalente a escoger q^* tal que

$$q^* = \arg\min_q D_{KL}(p||q). \tag{11.46}$$

11.4 — Información y Entropía

Ahora nos dedicaremos a conectar el concepto de divergencia de Kullback-Leibler con la idea de información, y en particular con una medida cuantitativa del *contenido de información*. El primero en establecer estas ideas en un marco matemático riguroso fue Claude Shannon (1948), en el contexto de formular una teoría de la comunicación de mensajes entre un emisor y un receptor en el caso donde existe pérdida de información [(1)].

Como ejemplo, imaginemos que Amelia elige un número entero, digamos, entre 1 y 16, y que Beatriz debe adivinarlo, realizando sólo preguntas de tipo *sí o no*. Por ejemplo, Beatriz podría preguntar «¿Es tu número impar?», «¿Es tu número menor o igual que 7?» o incluso «¿Es tu número el 13?». Es posible mostrar que si el número fue elegido de forma equiprobable, en promedio (en un sentido que veremos pronto) se necesitarán 4 preguntas para acertar. Este número corresponde al logaritmo en base 2 del número de alternativas, esto es, $2^4 = 16$.

Veremos ahora una versión simplificada del razonamiento de Shannon para cuantificar la información $\mathcal{I}(A)$ que ganamos al conocer la respuesta A a una pregunta. Supongamos que tenemos una pregunta con n posibles respuestas, que denotaremos por las proposiciones lógicas $A_1, \ldots, A_n$, y para las cuales hemos asignado probabilidades

$$p_i := P(A_i|I).$$

Deseamos asociar un contenido de información $\mathcal{I}_k$ al conocer que la respuesta correcta es A_k, donde claramente cuán informativo es este hecho sólo dependerá de la probabilidad p_k que hayamos asignado, esto es, postularemos el siguiente requerimiento.

[(1)] La teoría de Shannon fue la que hizo posibles todos nuestros avances en telecomunicaciones y almacenamiento de información en la era digital.

Postulado 11.1 — La información sólo depende de la probabilidad
La información $\mathcal{I}(A_k)$ ganada al conocer una respuesta A_k depende únicamente de la probabilidad $p_k = P(A_k|I)$.

Gracias a este postulado podemos enfocarnos en la búsqueda de una función universal $\mathcal{I}(p)$ con $p \in [0,1]$. A continuación, quisiéramos incluir el hecho de que si ya esperábamos la respuesta A_k con probabilidad 1 —esto es, era imposible cualquier otra— entonces no hemos ganado información alguna, y por tanto postularemos lo siguiente.

Postulado 11.2 — Certeza corresponde a información nula
Si $p = 1$, entonces $\mathcal{I}(p) = 0$.

Por otro lado, en el extremo opuesto esperamos que el conocer que la respuesta era una que creíamos muy improbable nos entregue mucha información, y por lo tanto cuanto menor sea la probabilidad que habíamos asignado, mayor es la información que ganamos. Esto se traduce en el siguiente postulado.

Postulado 11.3 — A menor probabilidad, mayor información
La información $\mathcal{I}(p)$ es una función decreciente de p, esto es,

$$\frac{\mathrm{d}\mathcal{I}(p)}{\mathrm{d}p} < 0.$$

Nuestro último postulado es el más importante, y el que finalmente nos permitirá descubrir la forma que toma la función $\mathcal{I}(p)$. Requeriremos que la información que ganamos al conocer las respuestas A_1 y A_2 a dos preguntas independientes tales que $P(A_1, A_2|I) = P(A_1|I)P(A_2|I)$ sea la suma de las informaciones que ganaríamos al conocer A_1 y A_2 por separado.

Postulado 11.4 — Aditividad de la información independiente
La información asociada a conocer dos respuestas independientes es la suma de las informaciones individuales. Esto es,

$$\mathcal{I}(p_1\ p_2) = \mathcal{I}(p_1) + \mathcal{I}(p_2)$$

Con esto tenemos todos los elementos necesarios para determinar la función $\mathcal{I}(p)$ de una vez por todas. Si escribimos el postulado **Aditividad de la información independiente** como

$$\mathcal{I}(xy) = \mathcal{I}(x) + \mathcal{I}(y)$$

y derivando con respecto a x, tenemos

$$\frac{\partial}{\partial x}\mathcal{I}(xy) = \mathcal{I}'(xy)y = \mathcal{I}'(x) \tag{11.47}$$

que no depende de y, luego podemos escribir

$$\mathcal{I}'(xy) = \frac{\alpha(x)}{y}. \tag{11.48}$$

donde α es una función aún no determinada. Para el caso $x = 1$ tenemos

$$\mathcal{I}'(y) = \frac{\alpha(1)}{y}, \tag{11.49}$$

e integrando llegamos a

$$\mathcal{I}(y) = \alpha(1)\ln y + h, \tag{11.50}$$

pero usando (11.50) con $y = 1$ nos dice que

$$\mathcal{I}(1) = \alpha(1)\ln 1 + h = h,$$

luego del postulado **Certeza corresponde a información nula** se sigue que $h = 0$. El postulado **A menor probabilidad, mayor información** puede satisfacerse si elegimos $\alpha(1) < 0$, luego la única función que satisface todos nuestros postulados es

$$\mathcal{I}(p) = -k\ln p, \tag{11.51}$$

donde $k > 0$ es una constante, en principio arbitraria, que nos fija las unidades en que mediremos la información. Como $k > 0$ automáticamente tenemos que $\mathcal{I}(p) \geq 0$, y para simplificar la notación, vamos a elegir $k = 1$, con lo que tenemos la definición de la *información de Shannon*.

Definición 11.4 — Información de Shannon
Si A es una proposición lógica, entonces la información de Shannon $\mathcal{I}(A|I)$ asociada a A en el estado de conocimiento I es

$$\mathcal{I}(A|I) := -\ln P(A|I). \tag{11.52}$$

Como $\mathcal{I}(A|I)$ es la información que ganamos al saber que A es cierto si estamos en el estado I, es por lo tanto información que no poseemos, es decir, *información faltante* que no está incluida en I. Con esta nueva herramienta, volvamos ahora a la definición de divergencia de Kullback-Leibler. Podemos escribirla como

$$\begin{aligned} D_{KL}(p||q) &= \big\langle \ln p(\boldsymbol{X}) - \ln q(\boldsymbol{X})\big\rangle_M \\ &= \big\langle \ln P(\boldsymbol{X}|M) - \ln P(\boldsymbol{X}|M')\big\rangle_M \\ &= \big\langle \mathcal{I}(\boldsymbol{X}|M') - \mathcal{I}(\boldsymbol{X}|M)\big\rangle_M \end{aligned} \tag{11.53}$$

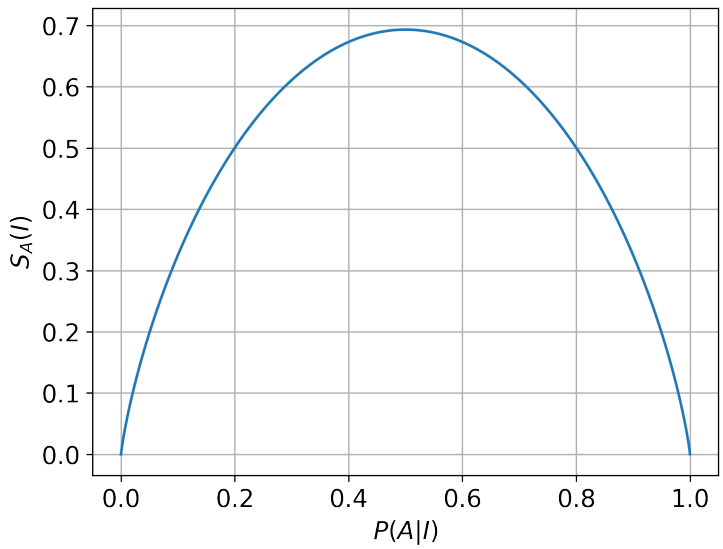


Figura 11.2: Entropía binaria $\mathcal{S}_A(I)$ asignada a una proposición A en el estado de conocimiento I, dada por (11.56).

esto es, la estimación de la diferencia entre la información faltante para el modelo M' y la información faltante para el modelo M. Por la desigualdad de Gibbs vemos que

$$\langle \mathcal{I}(\boldsymbol{X}|M') \rangle_M \geq \langle \mathcal{I}(\boldsymbol{X}|M) \rangle_M.$$

La estimación de la información faltante acerca de un aspecto que nuestro modelo describe es un concepto de suma importancia, conocido como la entropía de información, entropía de Shannon o simplemente entropía.

Definición 11.5 — Entropía de Shannon
Para un conjunto de proposiciones $A_1, \ldots, A_n$ mutuamente excluyentes y exhaustivas definiremos la entropía de Shannon $\mathcal{S}_A(I)$ en el estado I como la expectación de la información no contenida en I que ganaríamos al conocer cuál de las proposiciones $\{A_i\}$ es cierta,

$$\mathcal{S}_A(I) := \left\langle \mathcal{I}(A|I) \right\rangle_I = -\sum_{i=1}^{n} p_i \ln p_i \tag{11.54}$$

para $p_i = P(A_i|I)$.

Al evaluar la entropía de un conjunto de proposiciones hay que tener en cuenta el límite

$$\lim_{p \to 0} (p \ln p) = 0, \tag{11.55}$$

que hace que la suma en (11.54) sea bien comportada para probabilidades cercanas a cero.

> **Importante:** La entropía es una propiedad de un modelo, ya que se construye a partir de las probabilidades que dicho modelo asigna a las posibles respuestas a una pregunta. Esto significa que dos personas con modelos distintos de un aspecto de la realidad en general asignarán entropías distintas a dicho aspecto.

Por ejemplo, si una pregunta tiene sólo las posibles respuestas A y $\neg A$, con $p := P(A|I)$, entonces

$$\mathcal{S}_A(I) = -p \ln p - (1-p)\ln(1-p) \tag{11.56}$$

cuya gráfica en función de p se muestra en la **Figura 11.2**. Aquí vemos claramente que la entropía es cero en los extremos $p = 0$ y $p = 1$ donde tenemos más certeza, y es máxima en el caso de mayor incerteza, $p = \frac{1}{2}$. En este último caso, el valor máximo de la entropía es

$$-\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln\frac{1}{2} = \ln 2$$

que corresponde a *un bit de información*. En otras palabras, cuando somos completamente ignorantes acerca de la respuesta a una pregunta de tipo **sí** o **no**, significa que nos falta un *bit* de información.

Cuando se trata de un conjunto de n proposiciones, o equivalentemente, una variable discreta que toma uno de n valores posibles, la definición de entropía es clara y libre de ambigüedades. Sin embargo, al momento de generalizar a variables continuas hay algunos puntos a considerar. En primer lugar, para una variable discreta $X \in \{x_1, \ldots, x_n\}$ cuyos valores son equiprobables en un estado $\varnothing$, esto es,

$$P(X = x_i|\varnothing) = \frac{1}{n}$$

se tiene que la entropía es

$$\mathcal{S}_X(\varnothing) = -\sum_{i=1}^{n} \frac{1}{n}\ln\frac{1}{n} = \ln n, \tag{11.57}$$

y si quisiéramos tomar el límite continuo la entropía crecería sin límite para $n \to \infty$. Por otro lado, si ingenuamente tomamos la propuesta

$$\mathcal{S}_X(I) \stackrel{?}{=} -\int_\Omega \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \tag{11.58}$$

como la generalización de (**11.54**) donde $p(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|I)$, entonces es directo ver que una transformación de coordenadas de $\boldsymbol{x}$ a $\boldsymbol{y} = \boldsymbol{Y}(\boldsymbol{x})$ para la cual

$$q(\boldsymbol{y}) := P(\boldsymbol{Y} = \boldsymbol{y}|I) = \tilde{p}(\boldsymbol{y})\mathcal{J}_{xy}(\boldsymbol{y}) \tag{11.59}$$

de acuerdo a (5.46) entregaría una entropía

$$\mathcal{S}_Y(I) \stackrel{?}{=} -\int_{\Omega'} \mathrm{d}\boldsymbol{y}\, q(\boldsymbol{y}) \ln \tilde{p}(\boldsymbol{y}) = -\int_{\Omega'} \mathrm{d}\boldsymbol{y}\, q(\boldsymbol{y}) \ln \left[\frac{q(\boldsymbol{y})}{\mathcal{J}_{xy}(\boldsymbol{y})}\right], \tag{11.60}$$

es decir, la forma de la entropía sería dependiente de la elección arbitraria de coordenadas. La solución a estos problemas es, en variables continuas, siempre usar la *entropía relativa* de un estado de conocimiento a otro, también conocida como entropía de Shannon-Jaynes, y que definimos a continuación.

Definición 11.6 — Entropía relativa

$$\mathcal{S}_X(I_0 \to I) := -\int d\boldsymbol{x} P(x|I) \ln \left[\frac{P(\boldsymbol{x}|I)}{P(\boldsymbol{x}|I_0)}\right]. \tag{11.61}$$

Con esta nueva definición, si para el sistema de coordenadas $\boldsymbol{x}$ la entropía relativa de I_0 a I es

$$\mathcal{S}_X(I_0 \to I) = -\int_{\Omega} \mathrm{d}\boldsymbol{x}\, p(\boldsymbol{x}) \ln \left[\frac{p(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right], \tag{11.62}$$

al transformar se tendrá

$$\begin{aligned} \mathcal{S}_Y(I_0 \to I) &= -\int_{\Omega'} \mathrm{d}\boldsymbol{y}\, q(\boldsymbol{y}) \ln \left[\frac{q(\boldsymbol{y})\cancel{\mathcal{J}_{xy}(\boldsymbol{y})}}{q_0(\boldsymbol{y})\cancel{\mathcal{J}_{xy}(\boldsymbol{y})}}\right] \\ &= -\int_{\Omega'} \mathrm{d}\boldsymbol{y}\, q(\boldsymbol{y}) \ln \left[\frac{q(\boldsymbol{y})}{q_0(\boldsymbol{y})}\right] \end{aligned} \tag{11.63}$$

ya que tanto la distribución en I_0 como la distribución en I transforman de la misma manera y sus efectos se cancelan. Tal vez no tan sorprendentemente, la entropía relativa coincide con el negativo de la divergencia de Kullback-Leibler, y por lo tanto

$$S_X(I_0 \to I) = -D_{KL}(p||p_0) \leq 0 \tag{11.64}$$

si $p(\boldsymbol{x}) = P(\boldsymbol{x}|I)$ y $p_0(\boldsymbol{x}) = P(\boldsymbol{x}|I_0)$, con la diferencia de que ahora los estados I_0 e I son estados de conocimiento en principio arbitrarios. Podría parecer extraño que la entropía relativa sea siempre negativa, pero si volvemos a mirar a la desigualdad de Gibbs tenemos que

$$\mathcal{S}_X(I_0 \to I) = \left\langle \mathcal{I}(\boldsymbol{X}|I) \right\rangle_I - \left\langle \mathcal{I}(\boldsymbol{X}|I_0) \right\rangle_I \leq 0 \tag{11.65}$$

lo cual nos indica que el estado I_0 contiene menos información que el estado I y justifica nuestra elección de notación $I_0 \to I$, ya que vamos de un estado previo a otro posterior.

Otro argumento para llegar a la definición de entropía relativa es el que da Jaynes (2003, pág. 375), comenzando desde la entropía de Shannon (11.54)

y pasando al límite continuo. Para esto, consideremos $X \in \{a, x_2, \ldots, x_{n-1}, b\}$ y las distribuciones de probabilidad

$$p_i := P(X = x_i|I), \tag{11.66a}$$

$$p_0 := P(X = x_i|\varnothing) = \frac{1}{n}. \tag{11.66b}$$

Definiremos densidades de probabilidad $p(x)$ y $m(x)$ asociadas a los estados I y $\varnothing$, respectivamente, usando (10.3) en la forma

$$m(x) = \frac{1}{n}\sum_{j=1}^{n}\delta(x - x_j), \tag{11.67a}$$

$$p(x) = \sum_{j=1}^{n} p_i\,\delta(x - x_j). \tag{11.67b}$$

Si integramos en el intervalo $[x_i - \varepsilon, x_{i+1} - \varepsilon]$ con ε suficientemente pequeño para que dicho intervalo únicamente contenga a x_i, tendremos

$$\frac{1}{n} = \int_{x_i-\varepsilon}^{x_{i+1}-\varepsilon} \mathrm{d}x\, m(x) \approx m(x_i)(\Delta x)_i, \tag{11.68a}$$

$$p_i = \int_{x_i-\varepsilon}^{x_{i+1}-\varepsilon} \mathrm{d}x\, p(x) \approx p(x_i)(\Delta x)_i. \tag{11.68b}$$

con $(\Delta x)_i := x_{i+1} - x_i$. Ahora podemos tomar el límite de la entropía de Shannon como

$$\begin{aligned}\lim_{n\to\infty} -\sum_{i=1}^{n} p_i \ln p_i &= \lim_{n\to\infty} -\sum_{i=1}^{n}(\Delta x)_i\, p(x_i)\ln[p(x_i)(\Delta x)_i] \\ &= \lim_{n\to\infty} -\sum_{i=1}^{n}(\Delta x)_i\, p(x_i)\ln\frac{p(x_i)}{nm(x_i)}\end{aligned} \tag{11.69}$$

que podemos reescribir como

$$\begin{aligned}\lim_{n\to\infty} -\sum_{i=1}^{n} p_i \ln\left[\frac{p_i}{p_0}\right] &= \lim_{n\to\infty} -\sum_{i=1}^{n}(\Delta x)_i\, p(x_i)\ln\frac{p(x_i)}{m(x_i)} \\ &= -\int_a^b \mathrm{d}x p(x)\ln\left[\frac{p(x)}{m(x)}\right].\end{aligned} \tag{11.70}$$

En esta derivación $m(x)$ coincide con la *densidad de puntos en el límite* tal que

$$f_\varepsilon(x) = \int_x^{x+\varepsilon} \mathrm{d}\zeta m(\zeta) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{Q}(x \le x_j \le x + \varepsilon) \tag{11.71}$$

es la fracción de puntos de la variable discreta original que se encuentran entre x y $x + \varepsilon$.

11.5 — Entropía y correlación de variables

Sean $P(X, Y|I_0)$ y $P(X, Y|I)$ las distribuciones conjuntas previa y posterior, respectivamente, para dos variables continuas X e Y. La entropía relativa conjunta es

$$\mathcal{S}_{XY}(I_0 \to I) = \left\langle -\ln\left[\frac{P(X, Y|I)}{P(X, Y|I_0)}\right]\right\rangle_I \tag{11.72}$$

que puede descomponerse usando la regla del producto como

$$\begin{aligned}
\mathcal{S}_{XY}(I_0 \to I) &= \left\langle -\ln\left[\frac{P(X|I)P(Y|X,I)}{P(X|I_0)P(Y|X,I_0)}\right]\right\rangle_I \\
&= \left\langle -\ln\left[\frac{P(X|I)}{P(X|I_0)}\right]\right\rangle_I + \left\langle -\ln\left[\frac{P(Y|X,I)}{P(Y|X,I_0)}\right]\right\rangle_I \\
&= \mathcal{S}_X(I_0 \to I) + \left\langle -\ln\left[\frac{P(Y|X,I)}{P(Y|X,I_0)}\right]\right\rangle_I \qquad (11.73) \\
\text{usando (5.17)} \quad &= \mathcal{S}_X(I_0 \to I) + \left\langle \left\langle -\ln\left[\frac{P(Y|X,I)}{P(Y|X,I_0)}\right]\right\rangle_{X=\bullet,I}\right\rangle_I \\
&= \mathcal{S}_X(I_0 \to I) + \left\langle \mathcal{S}_{Y|X}(\bullet; I_0 \to I)\right\rangle_I,
\end{aligned}$$

donde $\mathcal{S}_{Y|X}(x; I_0 \to I)$ es la entropía relativa condicional de Y dado que $X = x$, dada por

$$\begin{aligned}
\mathcal{S}_{Y|X}(x; I_0 \to I) &= \left\langle -\ln\left[\frac{P(Y|X=x,I)}{P(Y|X=x,I_0)}\right]\right\rangle_{X=x,I} \\
&= -\int dy P(Y=y|X=x,I)\ln\left[\frac{P(Y=y|X=x,I)}{P(Y=y|X=x,I_0)}\right].
\end{aligned} \qquad (11.74)$$

En el caso en que X e Y son independientes, se tiene

$$P(Y|X=x,\zeta) = P(Y|\zeta) \qquad (11.75)$$

para cualquier estado ζ, y entonces

$$\mathcal{S}_{Y|X}(x; I_0 \to I) \to \mathcal{S}_Y(I_0 \to I), \qquad (11.76)$$

y de esta forma (**11.73**) se reduce a

$$\mathcal{S}_{XY}(I_0 \to I) = \mathcal{S}_X(I_0 \to I) + \mathcal{S}_Y(I_0 \to I). \qquad (11.77)$$

Esto tiene dos consecuencias importantes. En primer lugar, si la entropía relativa conjunta no es igual a la suma de las entropías «marginales», entonces las variables X, Y no son independientes, y por tanto están correlacionadas. En segundo lugar, dado que $\mathcal{S}_{Y|X} \leq 0$ al ser una entropía relativa, su expectación también lo será y se tiene, en general, que

$$\mathcal{S}_X(I_0 \to I) \geq \mathcal{S}_{XY}(I_0 \to I), \qquad (11.78)$$

esto es, marginalizar una variable nunca disminuye la entropía relativa.

Otra manera de medir correlación entre dos variables es a través de la información mutua, que es simplemente la divergencia de Kullback-Leibler entre el modelo que supone X, Y independientes y el modelo conjunto.

Definición 11.7 — Información mutua

$$\mathcal{I}(X;Y) := \left\langle \ln\left[\frac{P(X,Y|I)}{P(X|I)P(Y|I)}\right]\right\rangle_I \geq 0. \qquad (11.79)$$

Vemos que a mayor información mutua, peor se vuelve la suposición de variables independientes, es decir, las variables en la realidad presentan mayor correlación.

Ejemplo 11.5.1. *Considere el modelo conjunto*

$$P(X = x, Y = y|\lambda) = \frac{\lambda}{G_\lambda} \exp(-\lambda(xy)). \tag{11.80}$$

con $X \geq 1, Y \geq 1$*, donde*

$$G_\lambda := \int_\lambda^\infty \mathrm{d}t \exp(-t)t^{-1} \tag{11.81}$$

y con $\lambda > 0$*. Las distribuciones marginales son*

$$P(X = x|\lambda) = \frac{\exp(-\lambda x)}{xG_\lambda}, \tag{11.82a}$$

$$P(Y = y|\lambda) = \frac{\exp(-\lambda y)}{yG_\lambda}, \tag{11.82b}$$

con lo que la información mutua entre X *e* Y *como función de* λ *es*

$$\begin{aligned} \mathcal{I}(X;Y) &= \left\langle \ln \left[\frac{\lambda xy\ G_\lambda \exp(-\lambda xy)}{\exp(-\lambda(x+y))} \right] \right\rangle_\lambda \\ &= \ln(\lambda G_\lambda) - \lambda \langle xy - x - y \rangle_\lambda + \langle \ln xy \rangle_\lambda \\ &= \ln(\lambda G_\lambda) - 1 + \frac{\exp(-\lambda)}{G_\lambda} + \frac{K_\lambda}{6G_\lambda} \end{aligned} \tag{11.83}$$

donde

$$K_\lambda := 6\gamma^2 + \pi^2 - 12\lambda \cdot {}_pF_q(1,1,1;2,2,2;-\lambda) + 6(\ln\lambda)(2\gamma + \ln\lambda) \tag{11.84}$$

con ${}_pF_q(\{a\};\{b\};z)$ *la función hipergeométrica generalizada, y* γ *la constante de Euler-Mascheroni,* $\gamma \sim 0.57721567$*.*

► Para más información acerca de los conceptos de entropía relativa, divergencia de Kullback-Leibler, información mutua y en general acerca de la teoría de información de Shannon la referencia por excelencia es el libro de Cover y Thomas (2006). También es ilustrativo revisar el libro de MacKay (2003).

PROBLEMAS

Problema 11.1. *Calcule la entropía relativa para una variable X desde un modelo $X \sim Exp(\lambda)$ hasta un modelo $X \sim Gamma(k, \theta)$, ambos modelos con igual media. Verifique que su resultado*

(a) puede expresarse sólo en términos del parámetro k,

(b) es igual a cero para $k = 1$, es decir, cuando ambos modelos coinciden.

Problema 11.2. *Calcule el BIC de un modelo $M = Exp(\lambda)$ y utilícelo para determinar las condiciones bajo las cuales es preferible usar un modelo gamma en lugar del modelo exponencial.*

Problema 11.3. *Calcule el BIC de un modelo normal.*

Problema 11.4. *Para un conjunto de 10 datos con promedio aritmético $\overline{x}$ =5.0 y desviación estándar $s = \sqrt{\overline{x^2} - \overline{x}^2}$ =0.5, compare usando el factor de Bayes*

(a) un modelo exponencial $P(x|x_0) = (1/x_0)\exp(-x/x_0)$ con $x_0 = \overline{x}$ versus un modelo normal con $\mu = x_0$ y σ =0.5,

(b) un modelo exponencial $P(x|\lambda) = \lambda \exp(-\lambda x)$ considerando todos los valores posibles de λ versus el mismo modelo normal de la parte (a). Utilice el prior *de Jeffreys $P(\lambda|\varnothing) \propto 1/\lambda$.*

¿Cuál modelo es más probable en cada caso y cuánto es la razón entre las probabilidades?

Problema 11.5. *Se quiere desarrollar un modelo para la concentración x en partes por millón (ppm) de un contaminante en un lago. Para esto se ha medido el valor promedio de la concentración, $\overline{x} = x_0$. A esta información la denominaremos I_1. Posteriormente se mide la desviación estándar de la concentración, $\sqrt{\overline{x^2} - \overline{x}^2} = s$. A esta nueva información la denominaremos I_2. Determine la cantidad de información que se gana al incluir tanto I_1 como I_2 en el modelo, respecto al caso donde sólo se incluye I_1. Explique cualitativamente la dependencia de la información ganada incluir la desviación estándar.*

Problema 11.6. *Calcule la información mutua para dos variables $X \geq 0$, $Y \geq 0$ tal que su suma $Z = X + Y$ cumple $Z \sim U(0, L)$ con $L > 0$. ¿Qué se puede decir de la correlación entre las variables a medida que L aumenta?*
Hint. *Necesitará calcular las distribuciones marginales de X y de Y.*

CAPÍTULO 12

El principio de máxima entropía

> The first principle is that you must not fool yourself and you are the easiest person to fool.
>
> Richard P. Feynman

En este capítulo nos enfocaremos en el uso de la entropía de Shannon y la entropía relativa, ya no para comparar dos modelos sino para crear o actualizar modelos incorporando nueva información. Rápidamente veremos que el uso de la entropía relativa es análogo al uso del teorema de Bayes y de hecho en muchos casos puede producir las mismas respuestas.

12.1 — ¿POR QUÉ MAXIMIZAR ENTROPÍA?

Nuestro punto de partida será el siguiente. De acuerdo a lo visto en la sección 11.4, la entropía de Shannon

$$\mathcal{S}_X(I) = -\sum_{i=1}^{n} p_i \ln p_i$$

con $p_i = P(X = x_i|I)$ representa la información faltante en nuestro modelo respecto a X, y que ganaríamos al realizar observaciones de dicha variable, mientras que la entropía relativa

$$\mathcal{S}_X(I_0 \to I) = -\int_\Omega \mathrm{d}\boldsymbol{x} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{p_0(\boldsymbol{x})} \leq 0$$

donde

$$p(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|I),$$
$$p_0(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|I_0),$$

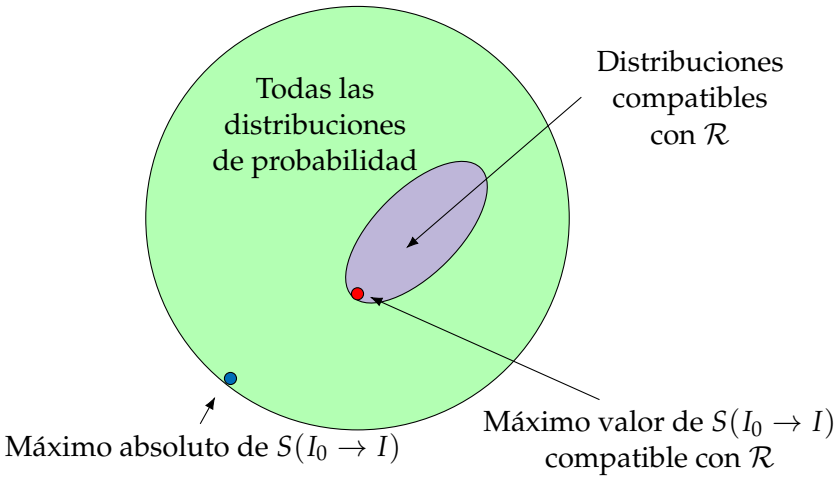


Figura 12.1: El principio de máxima entropía. La región en violeta representa las distribuciones compatibles con cierta información $\mathcal{R}$, y dentro de ésta, el punto rojo indica la distribución de mayor entropía relativa respecto a I_0 (punto azul).

cuantifica la diferencia en información faltante respecto a X entre los estados de conocimiento I_0 e I, donde I_0 es el menos informativo de los dos. Recordemos que la entropía relativa —al igual que la divergencia de Kullback-Leibler— es cero si y sólo si p es igual a p_0.

Si desde un estado de conocimiento I_0 inicial deseamos incluir una nueva información $\mathcal{R}$, deberíamos situarnos en un estado de conocimiento

$$I = I_0 \wedge \mathcal{R}$$

que incluya tanto la información en I_0 como la contenida en $\mathcal{R}$ **pero sin agregar ninguna información extra**, ya que de lo contrario estaríamos introduciendo *sesgos* indeseados.

Luego **buscamos el estado I compatible con $\mathcal{R}$ más similar a I_0** en cuanto a contenido de información, y de aquí se sigue que I es tal que minimiza la divergencia de Kullback-Leibler $D_{KL}(I||I_0)$ o equivalentemente, maximiza la entropía relativa $\mathcal{S}_X(I_0 \to I)$ dentro de la familia de estados compatibles con $\mathcal{R}$. A esta regla la denominaremos el *principio de máxima entropía*.

Recuadro 12.1 — El principio de máxima entropía
Para un conjunto de variables $\boldsymbol{X}$, la distribución de probabilidad menos sesgada $P(\boldsymbol{X} = \boldsymbol{x}|\mathcal{R}, I_0)$ que incluye como base cierta información I_0 y es compatible con una nueva información $\mathcal{R}$ es aquella función $p(\boldsymbol{x})$ que maximiza la entropía relativa

$$\mathcal{S}_X(I_0 \to I) = -\int_\Omega \mathrm{d}\boldsymbol{x} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{p_0(\boldsymbol{x})} \tag{12.1}$$

entre todas las funciones $p(\boldsymbol{x})$ compatibles con $\mathcal{R}$. Aquí se ha definido $p_0(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|I_0)$.

Esta es una versión más moderna del principio de máxima entropía, originalmente enunciado por Edwin T. Jaynes (1957) en base a maximizar la

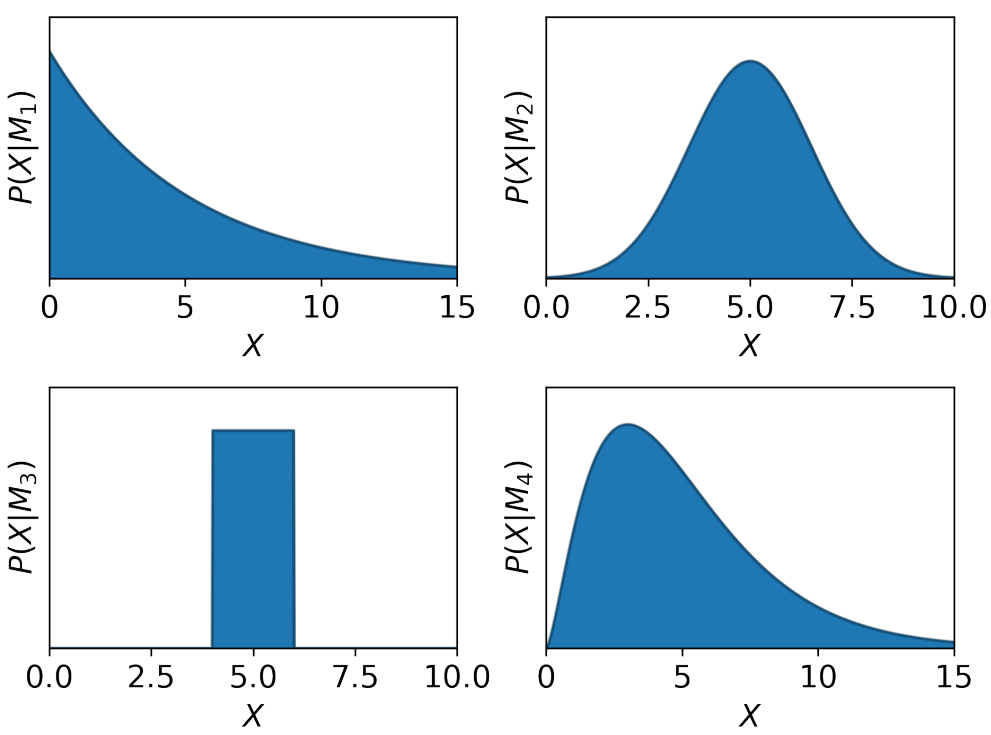


Figura 12.2: Cuatro modelos para $X > 0$ con media $\langle X \rangle_{M_\nu} = 5$ para $\nu = 1, 2, 3, 4$. Si sólo poseemos dicha información, ¿cuál es el modelo menos sesgado?

entropía de Shannon. En el esquema de la **Figura 12.1**, la región en violeta representa los estados compatibles con $\mathcal{R}$, el punto en rojo representa el estado I y el punto azul el estado I_0. De esta forma el punto rojo es el más cercano al azul dentro de la región en violeta.

La **Figura 12.2** muestra cuatro modelos distintos para una cantidad X positiva, todos con media igual a 5. Si la media es la única información que poseemos para describir a la variable X, ¿cuál es el modelo menos sesgado?

Por supuesto, el modelo de mayor entropía. Si nuestro *prior* es plano, el modelo que maximiza $\mathcal{S}_X(I_0 \to I)$ es el modelo exponencial, M_1 en la **Figura 12.2** como descubriremos en las siguientes secciones.

12.2 — Solución general para variables discretas

Como primera aplicación del principio de máxima entropía, consideremos una variable discreta $X \in \{x_1, x_2, \ldots, x_n\}$, para la cual tenemos un modelo inicial $q_i := P(x_i|I_0)$ y queremos que nuestro modelo actualizado $p_i := P(x_i|F, I_0)$ reproduzca el valor de la expectación

$$\langle f \rangle_I = \sum_{i=1}^{n} p_i f(x_i) = F. \tag{12.2}$$

La entropía relativa es

$$\mathcal{S}(I_0 \to I) = -\sum_{i=1}^{n} p_i \ln \left[\frac{p_i}{q_i}\right] = \mathcal{S}(p_1, \ldots, p_n) \tag{12.3}$$

entendiendo I como $I_0 \wedge F$. Aquí los valores de q_i están fijos y luego la entropía relativa es una función de los p_i. Maximizaremos esta función respecto a cada uno de los p_i pero sujeto a la restricción de normalización

$$\sum_{i=1}^{n} p_i = 1, \tag{12.4}$$

y a la restricción sobre la expectación de f dada en (12.2), la cual escribiremos como

$$\sum_{i=1}^{n} p_i f(x_i) = F. \tag{12.5}$$

Para maximizar bajo restricciones emplearemos el método de los multiplicadores de Lagrange, con lo cual maximizamos libremente la función aumentada

$$\tilde{\mathcal{S}} := \mathcal{S} - \lambda \left(\sum_{i=1}^{n} p_i \, f(x_i) - F \right) - \mu \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{12.6}$$

donde μ y λ son los multiplicadores de Lagrange que imponen las restricciones en (12.4) y (12.5), respectivamente. Como la derivada parcial de la entropía relativa respecto a la k-ésima probabilidad es

$$\frac{\partial \mathcal{S}}{\partial p_k} = -(1 + \ln p_k - \ln q_k), \tag{12.7}$$

tenemos

$$\frac{\partial \tilde{\mathcal{S}}}{\partial p_k} = -(1 + \ln p_k - \ln q_k) - \lambda f(x_k) - \mu = 0 \text{ para } k = 1, \ldots, n, \tag{12.8}$$

y luego despejando obtenemos

$$p_i = \frac{q_i \exp(-\lambda f(x_i))}{Z} \text{ para todo } i = 1, \ldots, n \tag{12.9}$$

donde además hemos definido por conveniencia la constante de normalización

$$Z = \exp(\mu + 1),$$

que puede ser calculada como

$$Z = \sum_{i=1}^{n} q_i \exp(-\lambda f(x_i)) \tag{12.10}$$

y es por tanto una función de λ, que llamaremos la **función partición**. Finalmente podemos escribir nuestra distribución actualizada como

$$P(x_i|I) = P(x_i|I_0) \frac{\exp(-\lambda f(x_i))}{Z(\lambda)}. \tag{12.11}$$

El valor del multiplicador de Lagrange λ es tal que consigue imponer la restricción, es decir, es el valor que asegura que $\langle f \rangle_\lambda = F$. Por lo tanto, reconociendo que

$$\langle f \rangle_\lambda = \frac{1}{Z(\lambda)} \sum_{i=1}^{n} \exp(-\lambda f(x_i)) f(x_i) = -\frac{\partial}{\partial \lambda} \ln Z(\lambda), \tag{12.12}$$

la ecuación que determina el valor de λ es

$$-\frac{\partial}{\partial\lambda}\ln Z(\lambda) = F. \tag{12.13}$$

Las ecuaciones (12.11) y (12.13) forman la solución general para el problema discreto de máxima entropía con una única restricción. Notemos que cuando $\lambda = 0$ recuperamos $P(x_i|I) = P(x_i|I_0)$ como debe ser, ya que en ese caso el estado de conocimiento I coincide con I_0 (no hay restricciones adicionales a la de normalización).

Ejemplo 12.2.1. *En un sitio de compras en línea es posible calificar a los productos con puntajes que van desde una estrella (★) hasta cinco estrellas (★★★★★).*

El puntaje promedio histórico se muestra en números con un dígito decimal, por ejemplo 4.3. Si para un producto vemos que tiene puntaje promedio de 3.7, ¿cuál es la probabilidad de que haya obtenido dos estrellas o menos?

***Solución**. Llamaremos n a la calificación como número entero de estrellas, luego $n \in \{1,2,3,4,5\}$. El promedio conocido $\overline{n} = 3.7$ lo denotaremos como n_0, y entonces pediremos que la expectación de n reproduzca dicho promedio,*

$$\langle n\rangle_{\mathcal{R}} = n_0. \tag{12.14}$$

Usando la solución general (12.11) con prior

$$P(n|I_0) = \frac{1}{5}, \tag{12.15}$$

obtenemos

$$p_n := P(n|n_0, I_0) = \frac{\exp(-\lambda n)}{Z(\lambda)}, \tag{12.16}$$

donde λ es el multiplicador de Lagrange que impone la restricción (12.14). Usamos (12.13) en la forma

$$-\frac{\partial}{\partial\lambda}\ln Z(\lambda) = n_0. \tag{12.17}$$

La función partición $Z(\lambda)$ puede calcularse como

$$\begin{aligned} Z(\lambda) = \sum_{n=1}^{5}\exp(-\lambda n) = \sum_{n=1}^{5}\{\exp(-\lambda)\}^n \\ = q + q^2 + q^3 + q^4 + q^5, \end{aligned} \tag{12.18}$$

donde $q := \exp(-\lambda)$. Como

$$\frac{\partial q}{\partial\lambda} = -\exp(-\lambda) = -q, \tag{12.19}$$

tenemos

$$-\frac{\partial Z(\lambda)}{\partial\lambda} = q + 2q^2 + 3q^3 + 4q^4 + 5q^5 \tag{12.20}$$

y entonces q puede obtenerse como solución de

$$q(1 + 2q + 3q^2 + 4q^3 + 5q^4) = n_0 q(1 + q + q^2 + q^3 + q^4). \tag{12.21}$$

Numéricamente, la única solución real positiva es q=1.44825, que corresponde a λ=-0.370356. El valor de la función partición es Z=17.3536, y entonces las probabilidades p_1 y p_2 son

$$p_1 = \frac{q}{Z} = 0.0834553, \tag{12.22}$$

$$p_2 = \frac{q^2}{Z} = 0.120864. \tag{12.23}$$

Luego la probabilidad de haber obtenido dos estrellas o menos de acuerdo a la información (12.14) y el prior *(12.15) es*

$$p_1 + p_2 = 0.204319,$$

aproximadamente de un 20 %.

Podemos extender fácilmente nuestra solución al caso de m restricciones de la forma

$$\langle f_j \rangle_{\mathcal{R}} = F_j \qquad \text{para } j = 1, \ldots, m \tag{12.24}$$

donde $f_1(X), \ldots, f_m(X)$ son funciones de $X \in \{x_1, \ldots, x_n\}$. Simplemente reemplazamos (12.6) por

$$\tilde{\mathcal{S}} := \mathcal{S} - \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} p_i \, f_j(x_i) - F \right) - \mu \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{12.25}$$

con lo que la condición de extremo queda

$$\frac{\partial \tilde{\mathcal{S}}}{\partial p_k} = -(1 + \ln p_k - \ln q_k) - \sum_{j=1}^{m} \lambda_j f_j(x_k) - \mu = 0 \text{ para } k = 1, \ldots, n, \tag{12.26}$$

y finalmente obtenemos

$$\boxed{P(x_i|I) = \frac{P(x_i|I_0)}{Z(\boldsymbol{\lambda})} \exp\Big(- \sum_{j=1}^{m} \lambda_j f_j(x_i) \Big)} \tag{12.27}$$

con $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ el vector de los multiplicadores de Lagrange y

$$Z(\boldsymbol{\lambda}) = \sum_{i=1}^{n} P(x_i|I_0) \exp\Big(- \sum_{j=1}^{m} \lambda_j f_j(x_i) \Big) \tag{12.28}$$

la función partición. Los valores de los multiplicadores de Lagrange λ_j se obtienen del sistema de ecuaciones

$$\boxed{-\frac{\partial}{\partial \lambda_j} \ln Z(\boldsymbol{\lambda}) = F_j \qquad \text{para } j = 1, \ldots, m.} \tag{12.29}$$

Ejemplo 12.2.2 (Razonamiento probabilístico)**.** *Consideremos un problema de inferencia con las premisas*

$$P(A|I) = 0.4, \tag{R_1}$$
$$P(A,B|I) = 0.26. \tag{R_2}$$

¿Cuál es la probabilidad $P(B|I)$? ¿Cuál es $P(A \vee B|I)$? Claramente no tenemos información suficiente para determinar estas probabilidades usando las reglas de la suma y el producto, sin embargo podemos usar el principio de máxima entropía. En primer lugar, convertimos las premisas a restricciones de tipo expectación,

$$\big\langle \mathrm{Q}(A) \big\rangle_I = 0.4, \tag{R_1'}$$
$$\big\langle \mathrm{Q}(A)\mathrm{Q}(B) \big\rangle_I = 0.26. \tag{R_2'}$$

Llamando por simplicidad de notación $a := \mathrm{Q}(A)$ y $b := \mathrm{Q}(B)$, tenemos

$$P(a,b|I) = P(a,b|I_0)\frac{\exp(-\lambda a - \mu ab)}{Z(\lambda,\mu)}, \qquad a,b \in \{0,1\} \tag{12.30}$$

donde hemos considerado un prior *plano*

$$P(a,b|I_0) = \frac{1}{4},$$

equivalente a suponer que $P(A|I_0) = P(B|I_0) = \frac{1}{2}$ y que A y B son independientes bajo I_0. Bajo estas condiciones, la función partición está dada por

$$\begin{aligned} Z(\lambda,\mu) &= \sum_{a=0}^{1}\sum_{b=0}^{1} \exp(-\lambda a - \mu ab) \\ &= 1 + 1 + \exp(-\lambda - \mu) + \exp(-\lambda) \\ &= 2 + \exp(-\lambda)\Big[1 + \exp(-\mu)\Big], \end{aligned} \tag{12.31}$$

y con ésta podemos fijar los parámetros λ y μ mediante el sistema de ecuaciones

$$-\frac{\partial}{\partial\lambda}\ln Z(\lambda,\mu) = \frac{\exp(-\lambda)\,[1+\exp(-\mu)]}{2+\exp(-\lambda)\,[1+\exp(-\mu)]} = 0.4, \tag{12.32a}$$
$$-\frac{\partial}{\partial\mu}\ln Z(\lambda,\mu) = \frac{\exp(-\lambda-\mu)}{2+\exp(-\lambda)\,[1+\exp(-\mu)]} = 0.26. \tag{12.32b}$$

Resolviendo este sistema numéricamente, obtenemos λ =0.76214 y μ =-0.619039, y con esto ahora podemos calcular $P(B|I)$ como la expectación

$$\begin{aligned} P(B|I) &= \big\langle \mathrm{Q}(B) \big\rangle_I \\ &= \sum_{a=0}^{1}\sum_{b=0}^{1} bP(a,b|I) \\ &= \frac{1}{Z(\lambda,\mu)}\sum_{a=0}^{1}\exp(-\lambda a - \mu a) \\ &= \frac{1}{Z(\lambda,\mu)}\Big(1+\exp(-\lambda-\mu)\Big), \\ &= \frac{1+\exp(-\lambda-\mu)}{2+\exp(-\lambda)\,[1+\exp(-\lambda)]} = 0.56. \end{aligned} \tag{12.33}$$

Por otro lado, podemos calcular $P(A \vee B|I)$ como la expectación

$$\begin{aligned} P(A \vee B|I) &= \big\langle \mathrm{Q}(A \vee B) \big\rangle_I \\ &= \frac{1}{Z(\lambda,\mu)} \left[1 \cdot 0 + 1 \cdot 1 + \exp(-\lambda) \cdot 1 + \exp(-\lambda - \mu) \cdot 1\right] \\ &= \frac{1 + \exp(-\lambda)\left[1 + \exp(-\mu)\right]}{2 + \exp(-\lambda)\left[1 + \exp(-\mu)\right]} = 0.7, \end{aligned} \tag{12.34}$$

que precisamente coincide con la regla extendida de la suma,

$$P(A|I) + P(B|I) - P(A, B|I) = 0.4 + 0.56 - 0.26 = 0.7.$$

Notemos además que

$$P(A|I)P(B|I) = 0.4 \cdot 0.56 = 0.224 \neq P(A, B|I),$$

luego el resultado de máxima entropía nos dice que A y B dejan de ser independientes luego de las restricciones (R_1) y (R_2).

12.3 — Variables y restricciones continuas

En este punto podemos hacer dos tipos de generalizaciones de nuestra solución en (12.27), (12.28) y (12.29). El primer nivel de generalización es promover X a una variable continua, y esto lo podemos lograr tomando (12.27), multiplicando a ambos lados por $\delta(X - x_i)$ y sumando en i,

$$\begin{aligned} &\sum_{i=1}^{n} p_i \delta(X - x_i) = \frac{1}{Z(\boldsymbol{\lambda})} \sum_{i=1}^{n} q_i \exp\Big(-\sum_{j=1}^{m} \lambda_j f_j(x_i)\Big) \delta(X - x_i) \\ &\hookrightarrow p(X) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\Big(-\sum_{j=1}^{m} \lambda_j f_j(X)\Big) q(X), \end{aligned} \tag{12.35}$$

con lo que obtenemos para una variable continua X la llamada *familia exponencial de distribuciones*.

Definición 12.1 — Familia exponencial de distribuciones

$$P(X = x|\mathcal{R}, I_0) = \frac{P(X = x|I_0)}{Z(\boldsymbol{\lambda})} \exp\left(-\sum_{j=1}^{m} \lambda_j f_j(x)\right). \tag{12.36}$$

Los multiplicadores λ_j se obtienen de la misma manera que en (12.29),

$$-\frac{\partial}{\partial \lambda_j} \ln Z(\boldsymbol{\lambda}) = F_j \qquad \text{para } j = 1, \ldots, m. \tag{12.37}$$

Veamos a continuación algunos ejemplos de distribuciones que pertenecen a la familia exponencial y que pueden ser recuperadas a partir del principio de máxima entropía. En primer lugar, consideremos una variable $X \geq 0$

tal que

$$\langle X \rangle_{x_0} = x_0. \tag{12.38}$$

Usando (12.36) con un prior $P(X|\varnothing) \propto \Theta(X)$ tenemos

$$P(X = x|x_0) = \frac{\Theta(x)}{Z(\lambda)} \exp(-\lambda x) \tag{12.39}$$

con

$$Z(\lambda) = \int_0^\infty \mathrm{d}x \Theta(x) \exp(-\lambda x) = \frac{1}{\lambda} \tag{12.40}$$

y

$$x_0 = -\frac{\partial}{\partial \lambda} \ln Z(\lambda) = \frac{\partial}{\partial \lambda} \ln \lambda = \frac{1}{\lambda}, \tag{12.41}$$

así que tenemos

$$P(X = x|x_0) = \frac{\Theta(x)}{x_0} \exp(-x/x_0) \tag{12.42}$$

que es la distribución exponencial $X \sim \mathrm{Exp}(1/x_0)$. Si además de la restricción en (12.38) incorporamos una restricción nueva sobre la expectación del logaritmo de X, de forma que tenemos

$$\langle X \rangle_{x_0, L_0} = x_0, \tag{12.43a}$$

$$\langle \ln X \rangle_{x_0, L_0} = L_0, \tag{12.43b}$$

la solución (12.36) queda como

$$P(X = x|x_0) = \frac{\Theta(x)}{Z(\lambda_1, \lambda_2)} \exp(-\lambda_1 x - \lambda_2 \ln x), \tag{12.44}$$

con

$$\begin{aligned} Z(\lambda_1, \lambda_2) &= \int_0^\infty \mathrm{d}x \exp(-\lambda_1 x - \lambda_2 \ln x) \\ &= \int_0^\infty \mathrm{d}x \exp(-\lambda_1 x) x^{-\lambda_2} \\ &= \lambda_1^{\lambda_2 - 1} \Gamma(1 - \lambda_2) \end{aligned} \tag{12.45}$$

si $\lambda_1 > 0$ y $\lambda_2 < 1$. Definiendo $k := 1 - \lambda_2$ y $\theta := 1/\lambda_1$ vemos que (12.44) es la distribución gamma,

$$P(X = x|k, \theta) = \Theta(x) \frac{\exp(-x/\theta) x^{k-1}}{\Gamma(k) \theta^k}, \tag{12.46}$$

esto es, $X \sim \mathrm{Gamma}(k, \theta)$. Similarmente, si usamos como restricciones

$$\langle X \rangle_{x_0, (\Delta x)^2} = x_0, \tag{12.47a}$$

$$\langle (X - x_0)^2 \rangle_{x_0, (\Delta x)^2} = (\Delta x)^2, \tag{12.47b}$$

entonces se tiene

$$\begin{aligned} P(X = x|\lambda_1, \lambda_2) &= \frac{1}{Z(\lambda_1, \lambda_2)} \exp(-\lambda_1 x - \lambda_2 (x - x_0)^2) \\ &= \frac{1}{Z(\lambda_1, \lambda_2)} \exp\Big(-\lambda_2 \Big[x^2 - 2x_0 x + x_0^2 + 2\big(\frac{\lambda_1}{2\lambda_2}\big) x \Big] \Big) \\ &= \frac{1}{\eta(\lambda_1, \lambda_2)} \exp\left(-\lambda_2 \Big(x - \Big[x_0 - \frac{\lambda_1}{2\lambda_2} \Big] \Big)^2 \right) \end{aligned} \tag{12.48}$$

donde hemos definido una nueva constante de normalización

$$\eta(\lambda_1, \lambda_2) = Z(\lambda_1, \lambda_2) \exp\left(\lambda_1 x_0 - \frac{\lambda_1^2}{4\lambda_2}\right). \tag{12.49}$$

Esta constante puede calcularse de inmediato usando la integral gaussiana en (12),

$$\eta(\lambda_1, \lambda_2) = \int_{-\infty}^{\infty} \mathrm{d}x \exp\left(-\lambda_2\left(x - \left[x_0 - \frac{\lambda_1}{2}\right]\right)^2\right) = \sqrt{\frac{\pi}{\lambda_2}} \tag{12.50}$$

si $\lambda_2 > 0$, y vemos que no depende de λ_1. Con (12.49) y (12.50) podemos aplicar (12.37) para determinar los multiplicadores λ_1 y λ_2. En primer lugar, se tiene que

$$x_0 = -\frac{\partial}{\partial\lambda_1} \ln Z(\lambda_1, \lambda_2) = -\frac{\partial}{\partial\lambda_1} \ln \eta(\lambda_1, \lambda_2) + x_0 - \frac{\lambda_1}{2\lambda_2} \tag{12.51}$$

por lo tanto $\lambda_1 = 0$, y vemos de inmediato en (12.49) que $Z(\lambda_2) = \eta(\lambda_2)$, con lo que obtenemos

$$(\Delta x)^2 = -\frac{\partial}{\partial\lambda_2} \ln Z(\lambda_2) = \frac{1}{2}\frac{\partial}{\partial\lambda_2} \ln \lambda_2 = \frac{1}{2\lambda_2} \tag{12.52}$$

así que finalmente

$$\lambda_2 = \frac{1}{2(\Delta x)^2}$$

y podemos escribir nuestra distribución como

$$P(X = x | x_0, (\Delta x)^2) = \frac{1}{\sqrt{2\pi}|\Delta x|} \exp\left(-\frac{1}{2}\left(\frac{x - x_0}{|\Delta x|}\right)^2\right), \tag{12.53}$$

que es la distribución normal.

El siguiente nivel de generalización es pasar de un número m finito de restricciones, con multiplicadores $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ a un continuo de restricciones parametrizadas por una coordenada $t \in [0, 1]$, donde el vector de multiplicadores de Lagrange es promovido a una función multiplicadora de Lagrange, esto es,

$$\lambda_j \to \lambda(t_j) \quad \text{con } t \in [0, 1].$$

Este paso al continuo podemos conseguirlo considerando el límite de la suma en el argumento de la exponencial en (12.36) como

$$\begin{aligned}
\lim_{m\to\infty} \sum_{j=1}^{m} \lambda_j f_j(X) &= \lim_{m\to\infty} \int_0^1 \mathrm{d}t \delta(t - t_j) \sum_{j=1}^{m} \lambda(t_j) f(X; t_j) \\
&= \int_0^1 \mathrm{d}t \underbrace{\left[\lim_{m\to\infty} \sum_{j=1}^{m} \delta(t - t_j)\right]}_{=\mu(t)} \lambda(t) f(X; t) \\
&= \int_0^1 \mathrm{d}t \mu(t) \lambda(t) f(X; t).
\end{aligned} \tag{12.54}$$

Sin pérdida de generalidad podemos redefinir $\lambda(t)$ como $\mu(t)\lambda(t)$ y por lo tanto tenemos que la generalización de (12.36) es

$$P(X = x|\mathcal{R}, I_0) = \frac{P(X = x|I_0)}{Z[\lambda]} \exp\left(- \int_0^1 \mathrm{d}t\lambda(t)f(x;t) \right). \tag{12.55}$$

donde ahora $Z[\lambda]$ es el *funcional de partición*,

$$Z[\lambda] = \int_{-\infty}^{\infty} \mathrm{d}x P(X = x|I_0) \exp\left(- \int_0^1 \mathrm{d}t\lambda(t)f(x;t) \right). \tag{12.56}$$

La interpretación del resultado en (12.55) es claramente la distribución de máxima entropía relativa sujeta al continuo de restricciones

$$\left\langle f(X;t) \right\rangle_{\mathcal{R}} = F(t) \qquad \text{para todo } t \in [0,1] \tag{12.57}$$

que es precisamente la condición que permite fijar la función multiplicadora de Lagrange. Esta será ahora obtenida como solución de la ecuación funcional

$$-\frac{\delta}{\delta\lambda(t)} \ln Z[\lambda] = F(t) \qquad \text{para todo } t \in [0,1]. \tag{12.58}$$

Es importante notar que los límites del parámetro $t \in [0,1]$ son arbitrarios, y en efecto nada importante cambia si simplemente los reemplazamos por otros, digamos $t \in [a,b]$ o incluso $t \in \mathbb{R}$.

Ejemplo 12.3.1. *Sabemos que la densidad de probabilidad de la variable $z = x^2$ está dada por $p(z)$. ¿Cuál es la densidad de probabilidad menos sesgada de la variable x?*
Solución. *Comenzamos escribiendo la restricción sobre la densidad de probabilidad como una expectación conocida,*

$$P(X^2 = z|\mathcal{R}, I_0) = \left\langle \delta(X^2 - z) \right\rangle_{\mathcal{R},I_0} = p(z) \qquad \textit{para todo } z \geq 0, \tag{12.59}$$

y aplicamos la solución de máxima entropía (12.55) con función multiplicadora de Lagrange $\lambda(z)$,

$$\begin{aligned} P(x|\mathcal{R}, I_0) &= \frac{P(x|I_0)}{Z[\lambda]} \exp\left(- \int_0^{\infty} \mathrm{d}z\lambda(z)\delta(x^2 - z) \right) \\ &= \frac{P(x|I_0)}{Z[\lambda]} \exp(-\lambda(x^2)). \end{aligned} \tag{12.60}$$

Ahora necesitamos determinar la función $\lambda(x^2)$ imponiendo la restricción en (12.59) y para eso tenemos

$$\begin{aligned} p(z) &= \left\langle \delta(X^2 - z) \right\rangle_{\mathcal{R},I_0} \\ &= \frac{\exp(-\lambda(z))}{Z[\lambda]} \int_0^{\infty} \mathrm{d}x P(x|I_0)\delta(x^2 - z) \\ &= \frac{\exp(-\lambda(z))}{Z[\lambda]} p_0(z) \end{aligned} \tag{12.61}$$

lo que nos entrega

$$P(x|\mathcal{R}, I_0) = P(x|I_0)\left[\frac{p(x^2)}{p_0(x^2)}\right]. \tag{12.62}$$

Finalmente, la densidad previa de $Z = X^2$ *puede calcularse como*

$$p_0(z) = \int_0^\infty \mathrm{d}x P(x|I_0)\delta(x^2 - z) = \frac{P(X = \sqrt{z}|I_0)}{2\sqrt{z}} \tag{12.63}$$

con lo que reemplazando en (12.62), obtenemos

$$P(x|\mathcal{R}, I_0) = \cancel{P(x|I_0)}\left[\frac{2x\, p(x^2)}{\cancel{P(x|I_0)}}\right] = 2x\, p(x^2) \tag{12.64}$$

que es exactamente la solución que obtendríamos transformando de acuerdo a (7.85), ya que $2x$ *es la derivada de la función* $x \mapsto x^2$*. Alternativamente, podemos verificar el resultado simplemente operando con expectaciones y la delta de Dirac,*

$$\begin{aligned} 2x\, p(x^2) &= 2x\left\langle \delta(X^2 - x^2)\right\rangle_{\mathcal{R},I_0} \\ &= \left\langle 2x\, \delta(X^2 - x^2)\right\rangle_{\mathcal{R},I_0} \\ &= \left\langle \cancel{2x}\left[\frac{\delta(X - x)}{\cancel{|2x|}}\right]\right\rangle_{\mathcal{R},I_0} \\ &= \left\langle \delta(X - x)\right\rangle_{\mathcal{R},I_0} \\ &= P(x|\mathcal{R}, I_0). \end{aligned} \tag{12.65}$$

12.4 — Identidades diferenciales para modelos de máxima entropía

Recordemos la definición de familia exponencial en (12.36), que es nuestra solución general al problema de máxima entropía con m restricciones,

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\lambda}, I_0) = \frac{P(\boldsymbol{X} = \boldsymbol{x}|I_0)}{Z(\boldsymbol{\lambda})} \exp\left(-\sum_{j=1}^{m} \lambda_j f_j(\boldsymbol{x})\right).$$

Desarrollamos su teorema de fluctuación-disipación para $\omega = \omega(\boldsymbol{X}, \boldsymbol{\lambda})$ como

$$\begin{aligned} \frac{\partial}{\partial\lambda_k}\left\langle\omega\right\rangle_{\boldsymbol{\lambda},I_0} &= \left\langle\frac{\partial\omega}{\partial\lambda_k}\right\rangle_{\boldsymbol{\lambda},I_0} + \left\langle\omega\frac{\partial}{\partial\lambda_k}\ln P(\boldsymbol{X}|\boldsymbol{\lambda}, I_0)\right\rangle_{\boldsymbol{\lambda},I_0} \\ &= \left\langle\frac{\partial\omega}{\partial\lambda_k}\right\rangle_{\boldsymbol{\lambda},I_0} - \left\langle\omega\left(f_k(\boldsymbol{X}) + \frac{\partial}{\partial\lambda_k}\ln Z(\boldsymbol{\lambda})\right)\right\rangle_{\boldsymbol{\lambda},I_0} \\ &= \left\langle\frac{\partial\omega}{\partial\lambda_k}\right\rangle_{\boldsymbol{\lambda},I_0} - \left\langle\omega\Big(\underbrace{f_k(\boldsymbol{X}) - \left\langle f_k\right\rangle_{\boldsymbol{\lambda},I_0}}_{=\delta f_k}\Big)\right\rangle_{\boldsymbol{\lambda},I_0}, \end{aligned} \tag{12.66}$$

para luego usar la propiedad

$$\left\langle\omega\delta f\right\rangle_I = \left\langle\delta\omega\delta f\right\rangle_I \tag{12.67}$$

con lo que finalmente tenemos

$$\frac{\partial}{\partial \lambda_k}\langle \omega \rangle_{\boldsymbol{\lambda}, I_0} = \left\langle \frac{\partial \omega}{\partial \lambda_k} \right\rangle_{\boldsymbol{\lambda}, I_0} - \left\langle \delta\omega \delta f_k \right\rangle_{\boldsymbol{\lambda}, I_0} . \tag{12.68}$$

El caso particular con ω igual a una de las funciones $\{f_i\}$ permite relacionar derivadas de estas funciones con la matriz de covarianza,

$$\frac{\partial}{\partial \lambda_j}\langle f_i \rangle_{\boldsymbol{\lambda}, I_0} = -\left\langle \delta f_i \delta f_j \right\rangle_{\boldsymbol{\lambda}, I_0} . \tag{12.69}$$

Ahora desarrollemos el teorema de variables conjugadas para $\omega = \omega(\boldsymbol{X})$. Tenemos

$$\begin{aligned} \left\langle \frac{\partial \omega}{\partial X_k} \right\rangle_{\boldsymbol{\lambda}, I_0} &= -\left\langle \omega \frac{\partial}{\partial X_k} \ln P(\boldsymbol{X}|\boldsymbol{\lambda}, I_0) \right\rangle_{\boldsymbol{\lambda}, I_0} \\ &= \sum_{j=1}^{m} \lambda_j \left\langle \omega \frac{\partial f_j}{\partial X_k} \right\rangle_{\boldsymbol{\lambda}, I_0} - \left\langle \omega \frac{\partial}{\partial X_k} \ln P(\boldsymbol{X}|I_0) \right\rangle_{\boldsymbol{\lambda}, I_0}, \end{aligned} \tag{12.70}$$

que podemos escribir en forma vectorial como

$$\left\langle \nabla \cdot \boldsymbol{\omega} \right\rangle_{\boldsymbol{\lambda}, I_0} + \left\langle \boldsymbol{\omega} \cdot \nabla \ln P(\boldsymbol{X}|\boldsymbol{\lambda}, I_0) \right\rangle_{\boldsymbol{\lambda}, I_0} = \sum_{j=1}^{m} \lambda_j \left\langle \boldsymbol{\omega} \cdot \nabla f_j \right\rangle_{\boldsymbol{\lambda}, I_0}. \tag{12.71}$$

Esta identidad permite, en principio, escribir un sistema lineal de ecuaciones para los multiplicadores $\{\lambda_j\}$ escogiendo adecuadamente las funciones de prueba $\{\omega_j\}$, el cual podría ser más sencillo que resolver (12.37) directamente.

12.5 — Entropía y priors no informativos

Veremos una aplicación del principio de máxima entropía para determinar *priors* no informativos de los parámetros de un modelo.

Supongamos un modelo $P(x|\theta, M) = p(x;\theta)$. Queremos determinar $P(\theta|M)$, y para esto escribimos la probabilidad $P(x|\theta, M)$ como una expectación,

$$\begin{aligned} p(x;\theta) &= P(x|\theta, M) \\ &= \frac{P(x, \theta|M)}{P(\theta|M)} \\ &= \frac{\langle \delta(X - x)\delta(\Theta - \theta) \rangle_M}{\langle \delta(\Theta - \theta) \rangle_M} \end{aligned} \tag{12.72}$$

para todo θ y todo x, la cual puede ser reescrita como

$$\left\langle \delta(\Theta - \theta) \Big[\delta(X - x) - p(x;\theta) \Big] \right\rangle_M = 0 \qquad \text{para todo } \theta, x. \tag{12.73}$$

Aplicando nuestra solución en (12.55) con una función multiplicadora de Lagrange $\lambda(\theta, x)$ tenemos

$$\begin{aligned} P(\theta, x|M) &= \frac{P(\theta, x|\varnothing)}{Z[\lambda]} \exp\Big(- \int d\theta' dx' \lambda(\theta', x') \delta(\theta - \theta') \Big[\delta(x - x') - p(x'; \theta')\Big]\Big) \\ &= \frac{P(\theta, x|\varnothing)}{Z[\lambda]} \exp\Big(- \lambda(\theta, x) + \int dx' \lambda(\theta, x') p(x'; \theta)\Big) \\ &= \frac{P(\theta, x|\varnothing)}{Z[\lambda]} \exp\Big(- \lambda(\theta, x) + L(\theta)\Big) \end{aligned} \tag{12.74}$$

donde hemos definido

$$L(\theta) := \int dx' \lambda(\theta, x') p(x'; \theta) = \int dx' \lambda(\theta, x') P(x'|\theta, M) = \big\langle \lambda \big\rangle_{\theta, M}. \tag{12.75}$$

Si ahora usamos la regla de marginalización para eliminar x, y factorizando

$$P(\theta, x|\varnothing) = P(\theta|\varnothing) P(x|\theta, \varnothing),$$

tenemos que la distribución marginal para θ es

$$\begin{aligned} P(\theta|M) &= \frac{P(\theta|\varnothing)}{Z[\lambda]} \exp(L(\theta)) \int dx' P(x'|\theta, \varnothing) \exp(-\lambda(\theta, x')) \\ &= \frac{P(\theta|\varnothing)}{Z[\lambda]} \exp(L(\theta)) \mu(\theta) \end{aligned} \tag{12.76}$$

con

$$\mu(\theta) := \int dx' P(x'|\theta, \varnothing) \exp(-\lambda(\theta, x')). \tag{12.77}$$

Ahora podemos escribir directamente la distribución condicional $P(x|\theta, M)$ en términos de $\lambda(\theta, x)$ y $\mu(\theta)$ como

$$P(x|\theta, M) = \frac{P(\theta, x|M)}{P(\theta|M)} = \frac{P(x|\theta, \varnothing) \exp(-\lambda(\theta, x))}{\mu(\theta)}, \tag{12.78}$$

con lo que $\lambda(\theta, x)$ está dado por

$$\lambda(\theta, x) = -\ln \left[\frac{P(x|\theta, M)}{P(x|\theta, \varnothing)}\right] - \ln \mu(\theta). \tag{12.79}$$

Tomando expectación a ambos lados en el estado de conocimiento (θ, M) tenemos

$$\begin{aligned} L(\theta) = \big\langle \lambda \big\rangle_{\theta, M} &= \left\langle -\ln \left[\frac{P(x|\theta, M)}{P(x|\theta, \varnothing)}\right] - \ln \mu(\theta) \right\rangle_{\theta, M} \\ &= \mathcal{S}_{x|\theta}(\theta) - \ln \mu(\theta) \end{aligned} \tag{12.80}$$

donde $\mathcal{S}_{x|\theta}$ es la entropía (relativa) condicional de la variable x dado θ. De aquí se sigue que

$$\exp(L(\theta)) \mu(\theta) = \exp\big(\mathcal{S}_{x|\theta}(\theta)\big), \tag{12.81}$$

y reemplazando en (12.76) finalmente llegamos a la definición del *prior entrópico*

$$\boxed{P(\theta|M) = \frac{1}{Z} P(\theta|\varnothing) \exp\big(\mathcal{S}_{x|\theta}(\theta)\big),} \tag{12.82}$$

con

$$Z = \int d\theta P(\theta|\varnothing) \exp\left(\mathcal{S}_{x|\theta}(\theta)\right). \tag{12.83}$$

La lectura inmediata de (12.82) es que, si $P(\theta|\varnothing)$ es plano, el parámetro más probable θ^* es tal que maximiza la entropía condicional de x dado θ.

Ejemplo 12.5.1. *Para $X \sim Exp(\lambda)$ y priors planos $P(\lambda, x|\varnothing)$ y $P(\lambda|\varnothing)$, determinar una distribución $P(\lambda|M)$ que incorpore el hecho de que λ es el parámetro de un modelo exponencial.*

Solución. *La entropía de X dado λ es simplemente*

$$\mathcal{S}_{x|\lambda}(\lambda) = \Big\langle -\ln\left(\lambda \exp(-\lambda X)\right)\Big\rangle_\lambda = 1 - \ln\lambda \tag{12.84}$$

ya que $\langle X \rangle_\lambda = 1/\lambda$, y se tiene

$$\exp(\mathcal{S}_{x|\lambda}(\lambda)) \propto \frac{1}{\lambda},$$

con lo que reemplazando en (12.82) tenemos simplemente

$$P(\lambda|M) \propto \frac{1}{\lambda}, \tag{12.85}$$

que coincide con el prior de Jeffreys.

► Para ver detalles sobre la idea de *priors* entrópicos con tratamientos ligeramente distintos, ver los artículos de Caticha y Preuss (2004) y de Abe (2014).

PROBLEMAS

Problema 12.1. *Sea un dado de 6 caras, que en un promedio de muchos lanzamientos obtiene $\overline{n} = 2.1$. ¿Cuál es la probabilidad de obtener un 6?*

Problema 12.2. *El promedio de notas finales de un curso es de $\overline{x} = 4.8$. ¿Cuál es la probabilidad de que un alumno haya reprobado (nota menor o igual que 4.0)? No descarte usar herramientas numéricas para llegar a su resultado.*

Problema 12.3. *En una búsqueda de un naufragio se estima que el navío perdido está dentro de un círculo de radio $R = 30$ km. Se estima que la distancia desde el centro del círculo al navío naufragado a lo largo de un cierto eje (llamémoslo $\hat{x}$), es de aproximadamente $x_0 = 22$ km.*

(a) Plantee el modelo $P(r, \theta|\lambda)$, con λ un multiplicador de Lagrange.

(b) Obtenga la función partición y encuentre el valor numérico del multiplicador de Lagrange λ.

(c) ¿Cuál es la probabilidad de que el navío esté en la coordenada $(20\hat{x} + 10\hat{y})$ *km?*

Problema 12.4. *Un rociador dispara gotas de agua verticalmente, con una velocidad inicial* v_0 *desconocida (y de hecho variable en el tiempo, es decir, distinta para cada lanzamiento). Sin embargo, se ha conseguido medir el promedio de la altura máxima que alcanzan las gotas, dado por* $\overline{h}$*. Recordando que la altura máxima para un lanzamiento está dada por* $h(v_0) = v_0^2/2g$*, y considerando que* v_0 *nunca supera un máximo de* $v_{max} = \sqrt{2gH}$*,*

(a) Cuál es el mejor modelo $P(v_0|\overline{h}, I)$*?*

(b) Cuál es el valor esperado de v_0*? Está de acuerdo con su intuición?*

(c) Determine la incerteza sobre v_0*, dada por* $\left\langle (\delta v_0)^2 \right\rangle_{\overline{h},I}$*. Cómo varía ésta con* $\overline{h}$*? Comente sobre por qué ésto debe ser así.*

Problema 12.5. *Un colectivo puede llevar máximo 4 pasajeros. Si la probabilidad de que un colectivo en cierto recorrido lleve más de 2 pasajeros es 0.6, ¿Cuál es la probabilidad de que un colectivo de ese recorrido pase completamente ocupado?*

***Hint**: Recuerde que hay*

$$\binom{4}{n} = \frac{4!}{n!(4-n)!}$$

ordenamientos posibles para n pasajeros ubicándose en los 4 asientos del colectivo.

CAPÍTULO **13**

Procesos Estocásticos

> Roads go ever ever on,
> Under cloud and under star.
> Yet feet that wandering have gone
> Turn at last to home afar.
>
> J. R. R. Tolkien, The Lord of the Rings

En términos simples, un *proceso estocástico* es un modelo para cantidades que cambian con el tiempo. Más precisamente, llamaremos proceso estocástico a una secuencia de variables desconocidas $\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ donde $\boldsymbol{X}_t$ será interpretado como el *estado de un sistema* al tiempo t. Esto significa que consideraremos el tiempo como discreto, y de esta forma, podemos asociar una distribución de probabilidad *instantánea* $P(\boldsymbol{X}_t = \boldsymbol{x}|I)$ a la variable $\boldsymbol{X}_t$, la cual nos entrega la probabilidad de estar en el estado $\boldsymbol{x}$ al tiempo t en el estado de conocimiento I.

Para una notación más compacta, representaremos la distribución instantánea de probabilidad como

$$\rho_t(\boldsymbol{x}) := P(\boldsymbol{X}_t = \boldsymbol{x}|I). \tag{13.1}$$

Nuestro objetivo es determinar la evolución de la distribución de probabilidad $\rho_t(\boldsymbol{X})$, es decir, cómo cambia ésta en el tiempo, suponiendo conocida la distribución inicial $\rho_0(\boldsymbol{X})$. Formalmente, dicha evolución siempre puede obtenerse usando la regla de marginalización para hacer aparecer el estado inicial $\boldsymbol{X}_0$, esto es,

$$\begin{aligned} P(\boldsymbol{X}_t = \boldsymbol{x}'|I) &= \sum_{\boldsymbol{x}_0} P(\boldsymbol{X}_t = \boldsymbol{x}', \boldsymbol{X}_0 = \boldsymbol{x}_0|I) \\ &= \sum_{\boldsymbol{x}_0} P(\boldsymbol{X}_0 = \boldsymbol{x}_0|I) P(\boldsymbol{X}_t = \boldsymbol{x}'|\boldsymbol{X}_0 = \boldsymbol{x}_0, I) \end{aligned} \tag{13.2}$$

que podemos escribir usando la notación compacta como

$$\rho_t(\boldsymbol{x}') = \sum_{\boldsymbol{x}_0} \rho_0(\boldsymbol{x}_0) K_t(\boldsymbol{x}_0 \to \boldsymbol{x}'). \tag{13.3}$$

Este ya es un resultado importante: nos dice que **siempre** existirá una transformación lineal que conecta dos tiempos, sin importar cuán alejados, transformación cuyo *kernel* K_t está dado por

$$K_t(\boldsymbol{x}_0 \to \boldsymbol{x}') := P(\boldsymbol{X}_t = \boldsymbol{x}' | \boldsymbol{X}_0 = \boldsymbol{x}_0, I). \tag{13.4}$$

13.1 — Cadenas de Markov

De acuerdo a (13.3), la evolución de cualquier proceso estocástico está determinada por la probabilidad de transición desde el estado inicial al estado final al tiempo t, para cualquier valor de t. Sin embargo, muchas veces la situación es mucho más fácil, y sólo se requieren probabilidades de transición entre tiempos cortos. Probablemente la clase más importante de procesos estocásticos son las llamadas *cadenas de Markov* o modelos markovianos, donde justamente existen este tipo de simplificaciones adicionales que hacen más tratable el problema de obtener K_t.

En una cadena de Markov el sistema no tiene memoria de los instantes anteriores al actual, por lo que el estado actual por sí solo define las probabilidades de transitar a uno de los posibles estados siguientes. Ejemplos de cadenas de Markov son los modelos de texto predictivo como los existentes en los teléfonos celulares actuales: en ellos, cuando uno tipea una palabra, automáticamente aparecen sugerencias de las posibles alternativas para la siguiente palabra. Comenzando desde una frase prefijada como «El mercado» es posible continuar autocompletando palabras hasta formar una frase, un poco sin sentido pero gramaticalmente correcta, como «El mercado de la construcción de un nuevo sistema de seguridad para la tarde de hoy». Claramente, «de la construcción» es una continuación muy frecuente de «mercado» en los textos que sirvieron para construir la cadena de Markov, al igual que «de seguridad» es una continuación muy frecuente de «sistema».

Otro ejemplo de cadena de Markov es una *caminata al azar*: un caminante con posición inicial $X_0 = 0$ puede saltar con desplazamiento ya sea $\Delta X = -1$ o $\Delta X = 1$ en el siguiente paso con igual probabilidad. La posición luego de un paso depende únicamente de la posición actual y del desplazamiento aleatorio escogido, es decir, se tendrá

$$X_{t+1} = X_t + (\Delta X)_t, \tag{13.5}$$

donde el desplazamiento $(\Delta X)_t$ al paso t es aleatorio y modelado por alguna distribución.

Una definición más formal de una cadena de Markov es la siguiente.

Definición 13.1 — Cadena de Markov
Un proceso estocástico constituye una cadena de Markov si se cumple

$$P(\boldsymbol{X}_{k+1} = \boldsymbol{x}_{k+1}|\boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_k = \boldsymbol{x}_k, I) = P(\boldsymbol{X}_{k+1} = \boldsymbol{x}_{k+1}|\boldsymbol{X}_k = \boldsymbol{x}_k, I), \quad (13.6)$$

es decir, la probabilidad de visitar un estado al instante $k+1$ sólo depende del estado k, y no de instantes anteriores.

Si esto se cumple, la probabilidad de una *trayectoria* particular

$$(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$$

puede ser separada como

$$\begin{aligned} P(\boldsymbol{X}_0 = \boldsymbol{x}_0, \ldots, \boldsymbol{X}_N = \boldsymbol{x}_N|I) &= P(\boldsymbol{X}_0 = \boldsymbol{x}_0|I) \prod_{i=0}^{N-1} P(\boldsymbol{X}_{i+1} = \boldsymbol{x}_{i+1}|\boldsymbol{X}_i = \boldsymbol{x}_i, I) \\ &= \rho_0(\boldsymbol{x}_0) \prod_{i=0}^{N-1} M_i(\boldsymbol{x}_i \to \boldsymbol{x}_{i+1}) \end{aligned} \quad (13.7)$$

esto es, en la probabilidad del estado inicial ρ_0 y un producto de *probabilidades de transición* $M_i(\boldsymbol{a} \to \boldsymbol{b})$ definidas como

$$M_t(\boldsymbol{x} \to \boldsymbol{x}') := P(\boldsymbol{X}_{t+1} = \boldsymbol{x}'|\boldsymbol{X}_t = \boldsymbol{x}, I)$$

correspondiente a la probabilidad de encontrar al sistema en el estado $\boldsymbol{x}'$ al tiempo $t+1$ dado que se encuentra en el estado $\boldsymbol{x}$ al tiempo t.

Al construir la probabilidad conjunta de $\boldsymbol{X}_{t+1}$ y $\boldsymbol{X}_t$ dado I y usar la regla de marginalización, al igual que hicimos para obtener (13.3), podemos concluir que

$$\begin{aligned} P(\boldsymbol{X}_{t+1} = \boldsymbol{x}'|I) &= \sum_{\boldsymbol{x}} P(\boldsymbol{X}_{t+1} = \boldsymbol{x}', \boldsymbol{X}_t = \boldsymbol{x}|I) \\ &= \sum_{\boldsymbol{x}} P(\boldsymbol{X}_{t+1} = \boldsymbol{x}'|\boldsymbol{X}_t = \boldsymbol{x}, I) P(\boldsymbol{X}_t = \boldsymbol{x}|I), \end{aligned} \quad (13.8)$$

es decir, en nuestra notación abreviada,

$$\rho_{t+1}(\boldsymbol{x}') = \sum_{\boldsymbol{x}} M_t(\boldsymbol{x} \to \boldsymbol{x}')\rho_t(\boldsymbol{x}). \quad (13.9)$$

Esta igualdad es conocida en la literatura como la ecuación de Chapman-Kolmogorov, y corresponde a un caso particular de (13.3) conectando tiempos contiguos t y $t+1$. Si aplicamos (13.9) iterativamente podemos ver que

$$\begin{aligned} \rho_t(\boldsymbol{x}') &= \int d\boldsymbol{x}\rho_0(\boldsymbol{x}) \left[\int d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{t-1} M_0(\boldsymbol{x} \to \boldsymbol{x}_1) \ldots M_{t-1}(\boldsymbol{x}_{t-1} \to \boldsymbol{x}') \right] \\ &= \int d\boldsymbol{x}\rho_0(\boldsymbol{x}) K_t(\boldsymbol{x} \to \boldsymbol{x}') \end{aligned} \quad (13.10)$$

donde vemos que el *kernel* definido en (13.4) es la convolución de las probabilidades de transición

$$K_t(\boldsymbol{x} \to \boldsymbol{x}') = \int d\boldsymbol{x}_1 \dots d\boldsymbol{x}_{t-1} M_0(\boldsymbol{x} \to \boldsymbol{x}_1) \dots M_{t-1}(\boldsymbol{x}_{t-1} \to \boldsymbol{x}'). \tag{13.11}$$

En una situación estacionaria, es decir, cuando la distribución instantánea $\rho_t(\boldsymbol{x})$ no depende de t y la podemos escribir como $\rho^*(\boldsymbol{x})$, la ecuación de Chapman-Kolmogorov se reduce a

$$\rho^*(\boldsymbol{x}') = \sum_{\boldsymbol{x}} M_t(\boldsymbol{x} \to \boldsymbol{x}')\rho^*(\boldsymbol{x}), \tag{13.12}$$

que podemos escribir de manera más simétrica como

$$\rho^*(\boldsymbol{x}') \sum_{\boldsymbol{x}} M_t(\boldsymbol{x}' \to \boldsymbol{x}) = \sum_{\boldsymbol{x}} M_t(\boldsymbol{x} \to \boldsymbol{x}')\rho^*(\boldsymbol{x}), \tag{13.13}$$

ya que se cumple que

$$\sum_{\boldsymbol{b}} M_t(\boldsymbol{a} \to \boldsymbol{b}) = \sum_{\boldsymbol{b}} P(\boldsymbol{X}_{t+1} = \boldsymbol{b} | \boldsymbol{X}_t = \boldsymbol{a}, I) = 1 \tag{13.14}$$

para todo estado inicial $\boldsymbol{a}$. La condición (13.13) es una condición necesaria para que ρ^* sea una distribución estacionaria dada una dinámica descrita por M_t. Una condición suficiente —pero no necesaria— es la llamada condición de *balance detallado*, que impone

$$\rho^*(\boldsymbol{x}')M_t(\boldsymbol{x}' \to \boldsymbol{x}) = \rho^*(\boldsymbol{x})M_t(\boldsymbol{x} \to \boldsymbol{x}'). \tag{13.15}$$

13.1.1 La ecuación maestra

Ahora veamos la forma *diferencial* de la ecuación de Chapman-Kolmogorov. Si tomamos (13.9) y formamos la diferencia $\rho_{t+1}(\boldsymbol{x}) - \rho_t(\boldsymbol{x})$ en el lado izquierdo, podemos escribir

$$\begin{aligned} \rho_{t+1}(\boldsymbol{x}) - \rho_t(\boldsymbol{x}) &= -\rho_t(\boldsymbol{x}) + \sum_{\boldsymbol{x}'} M_t(\boldsymbol{x}' \to \boldsymbol{x})\rho_t(\boldsymbol{x}') \\ &= -\rho_t(\boldsymbol{x}) \sum_{\boldsymbol{x}'} M_t(\boldsymbol{x} \to \boldsymbol{x}') + \sum_{\boldsymbol{x}'} M_t(\boldsymbol{x}' \to \boldsymbol{x})\rho_t(\boldsymbol{x}'). \end{aligned} \tag{13.16}$$

Cancelando entre ambas sumas del lado derecho el término $\boldsymbol{x}' = \boldsymbol{x}$, reemplazando $t+1$ por $t+\Delta t$ y dividiendo por Δt podemos escribir, sin pérdida de generalidad,

$$\begin{aligned} \frac{\rho_{t+\Delta t}(\boldsymbol{x}) - \rho_t(\boldsymbol{x})}{\Delta t} = -\rho_t(\boldsymbol{x}) \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} \left[\frac{M_t(\boldsymbol{x} \to \boldsymbol{x}')}{\Delta t} \right] \\ + \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} \left[\frac{M_t(\boldsymbol{x}' \to \boldsymbol{x})}{\Delta t} \right] \rho_t(\boldsymbol{x}'), \end{aligned} \tag{13.17}$$

En el límite de tiempo continuo, es decir $\Delta t \to 0$, (13.17) puede escribirse como

$$\frac{\partial \rho(x;t)}{\partial t} = -\rho(x;t) \sum_{x' \neq x} W_t(x \to x') + \sum_{x' \neq x} W_t(x' \to x)\rho(x';t). \tag{13.18}$$

que es la famosa *ecuación maestra*. Aquí hemos definido la tasa de transición $W_t(\boldsymbol{a} \to \boldsymbol{b})$ como

$$W_t(\boldsymbol{a} \to \boldsymbol{b}) := \lim_{\Delta t \to 0} \frac{M_t(\boldsymbol{a} \to \boldsymbol{b})}{\Delta t}.$$

▶ Para ver más sobre la ecuación maestra se recomienda el libro de Zwanzig (2001) y el de van Kampen (2007).

13.1.2 La ecuación de Fokker-Planck

Consideremos la ecuación maestra para una variable continua X, en la forma

$$\frac{\partial \rho(x;t)}{\partial t} = -\rho(x;t) \int dx' W_t(x \to x') + \int dx' W_t(x' \to x)\rho(x';t), \tag{13.19}$$

y supongamos ahora que la tasa de transición $W_t(x \to x')$ está concentrada en valores de x' muy cercanos a x. Esto es esperable ya que representa transiciones que ocurren en un intervalo de tiempo $\Delta t \to 0$. Escribamos ahora las tasas de transición como

$$W_t(x \to x') = \omega(x; -\varepsilon), \tag{13.20a}$$

$$W_t(x' \to x) = \omega(x - \varepsilon; \varepsilon) \tag{13.20b}$$

con $\varepsilon := x - x'$ y donde el primer argumento de ω indica la posición inicial mientras que el segundo indica el desplazamiento desde el punto inicial hasta el punto final. Hemos supuesto además que W_t no depende explícitamente del tiempo t. Reescribimos entonces (13.19) como

$$\frac{\partial \rho(x;t)}{\partial t} = \int d\varepsilon \Big(-\rho(x;t)\omega(x;-\varepsilon) + \omega(x-\varepsilon;\varepsilon)\rho(x-\varepsilon;t) \Big), \tag{13.21}$$

y como ε es muy pequeño, procedemos a expandir hasta segundo orden en ε el segundo término de la sumatoria, tomándolo como función de $x - \varepsilon$,

$$\rho(x-\varepsilon;t)\omega(x-\varepsilon;\varepsilon) \approx \rho(x;t)\omega(x;\varepsilon) - \varepsilon \frac{\partial}{\partial x}\big(\rho(x;t)\omega(x;\varepsilon)\big) + \frac{1}{2}\varepsilon^2 \frac{\partial^2}{\partial x^2}\big(\rho(x;t)\omega(x;\varepsilon)\big). \tag{13.22}$$

Si la tasa de transición $\omega(x;\varepsilon)$ es simétrica, se tiene $\omega(x;\varepsilon) = \omega(x;-\varepsilon)$ y entonces (13.21) se reduce a

$$\begin{aligned} \frac{\partial \rho(x;t)}{\partial t} &= \int d\varepsilon \Big(-\varepsilon \frac{\partial}{\partial x}\big(\rho(x;t)\omega(x;\varepsilon)\big) + \frac{1}{2}\varepsilon^2 \frac{\partial^2}{\partial x^2}\big(\rho(x;t)\omega(x;\varepsilon)\big) \Big) \\ &= -\frac{\partial}{\partial x}\Big(p(x;t) \int d\varepsilon \varepsilon \omega(x;\varepsilon) \Big) + \frac{\partial^2}{\partial x^2}\Big(p(x;t) \frac{1}{2} \int d\varepsilon \varepsilon^2 \omega(x;\varepsilon) \Big). \end{aligned} \tag{13.23}$$

Definiendo

$$\mu(x) := \int d\varepsilon \varepsilon \omega(x;\varepsilon), \tag{13.24a}$$

$$D(x) := \frac{1}{2}\int d\varepsilon \varepsilon^2 \omega(x;\varepsilon) \tag{13.24b}$$

podemos escribir finalmente la *ecuación de Fokker-Planck*,

$$\boxed{\frac{\partial \rho(x;t)}{\partial t} = -\frac{\partial}{\partial x}\big(\mu(x)p(x;t)\big) + \frac{\partial^2}{\partial x^2}\big(D(x)p(x;t)\big).} \tag{13.25}$$

Veamos con un poco más de detalle el significado de los coeficientes $\mu(x)$ y $D(x)$. En primer lugar, regresando a la notación completa la definición de $\mu(x)$,

$$\begin{aligned}\mu(x) &= \int dx'(x'-x)W_t(x\to x') \\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int dx'(x'-x)P(X_{t+\Delta t}=x'|X_t=x,I),\end{aligned} \tag{13.26}$$

esto es,

$$\boxed{\mu(x) = \lim_{\Delta t\to 0}\left\langle \frac{X_{t+\Delta t}-X_t}{\Delta t}\right\rangle_{X_t=x,I}.} \tag{13.27}$$

A la función $\mu(x)$ se le denomina el coeficiente de advección o de deriva (en inglés *drift coefficient*), y (13.27) nos dice que es la expectación condicional de la velocidad media en el intervalo Δt dada la posición al tiempo t. Por otro lado, se tiene para $D(x)$

$$\begin{aligned}D(x) &= \frac{1}{2}\int dx'(x'-x)^2 W_t(x\to x') \\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int dx'\frac{1}{2}(x'-x)^2 P(X_{t+\Delta t}=x'|X_t=x,I),\end{aligned} \tag{13.28}$$

que puede escribirse como

$$\boxed{D(x) = \frac{1}{2}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\Big\langle (X_{t+\Delta t}-X_t)^2\Big\rangle_{X_t=x,I},} \tag{13.29}$$

y se le llama el *coeficiente de difusión*. Notemos que el caso con $D(x) = D$ constante y finito, que exploraremos en la siguiente sección, implica que

$$\Big\langle (X_{t+\Delta t}-X_t)^2\Big\rangle_I = 2D\Delta t, \tag{13.30}$$

ya que en ese caso la expectación condicional en (13.29) no depende de X_t.

► Para todo el detalle sobre las múltiples derivaciones y técnicas de solución de la ecuación de Fokker-Planck, la referencia obligatoria es el libro de Risken (1996).

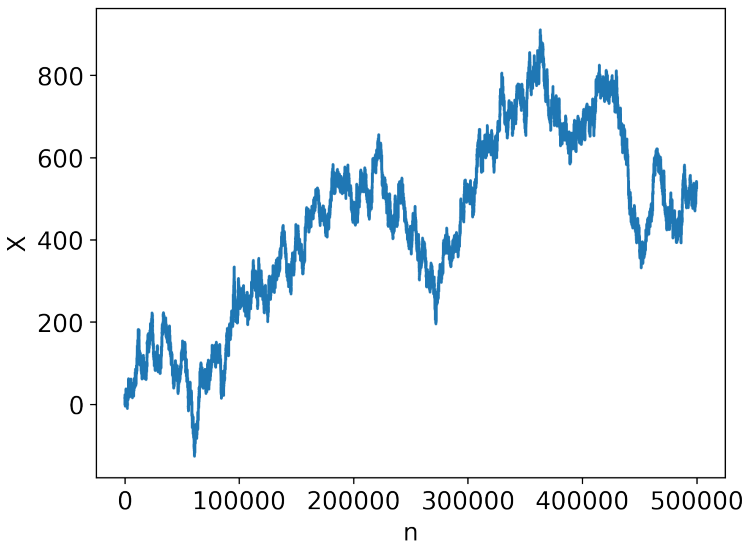


Figura 13.1: Movimiento browniano en una dimensión.

13.2 — Caminatas al azar

Un objeto de estudio por excelencia donde se emplea el formalismo de procesos estocásticos es el de las caminatas al azar (*random walks* en inglés), donde un «sistema» descrito por un conjunto de coordenadas $\boldsymbol{X}$ describe un movimiento errático, producto del efecto de pequeños incrementos en direcciones aleatorias. Típicamente se consideran dos condiciones para clasificar un proceso estocástico como una caminata al azar, (a) los desplazamientos no tienen una dirección privilegiada y (b) no existe correlación entre la dirección de un paso y la del paso siguiente. El primer tipo de caminatas al azar es el históricamente llamado movimiento browniano[(1)], que describiremos a continuación.

13.2.1 Movimiento browniano

Denominaremos movimiento browniano a una caminata al azar con desplazamientos «libres» en el espacio, esto es, sin un tamaño fijo y sin una dirección privilegiada, tal que es descrito por el caso particular de la ecuación de Fokker-Planck (13.25) con $\mu(x) = 0$ y $D(x) = D$ una constante, denominada la *ecuación de difusión*,

$$\frac{\partial \rho(x;t)}{\partial t} = D\frac{\partial^2 \rho(x;t)}{\partial x^2}. \tag{13.31}$$

Un ejemplo de movimiento browniano en una dimensión se muestra en la Figura 13.1. Para resolver la ecuación de difusión, es común introducir la transformada (3.227a) de la distribución $\rho(x;t)$ sobre la variable x,

$$\tilde{\rho}(k;t) := \int_{-\infty}^{\infty} \mathrm{d}x \exp(\mathrm{i}kx)\rho(x;t), \tag{13.32}$$

con lo que se tiene

$$\rho(x;t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx)\tilde{\rho}(k;t). \tag{13.33}$$

[(1)] En honor al biólogo y botánico Robert Brown, quien describió el movimiento errático de pequeñas partículas de polen suspendidas en agua.

Reemplazando (13.33) en la ecuación de difusión tenemos

$$\begin{aligned}\frac{\partial}{\partial t}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx)\tilde{\rho}(k;t) &= D\frac{\partial^2}{\partial x^2}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx)\tilde{\rho}(k;t) \\ &= -D\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx)\exp(-\mathrm{i}kx)k^2\tilde{\rho}(k;t)\end{aligned} \tag{13.34}$$

con lo que ahora sólo se necesita resolver la ecuación diferencial de primer orden

$$\frac{\partial\tilde{\rho}(k;t)}{\partial t} = -Dk^2\tilde{\rho}(k;t), \tag{13.35}$$

cuya solución se obtiene directamente como

$$\tilde{\rho}(k;t) = \tilde{\rho}(k;0)\exp(-Dk^2t). \tag{13.36}$$

De acuerdo al **Teorema de convolución**, de inmediato sabemos que $\rho(x;t)$ será la convolución entre la probabilidad inicial $\rho(x;0)$ y un nuevo *kernel*, dado esencialmente por la transformada inversa de $\exp(-Dk^2t)$. Para verlo explícitamente, desarrollamos

$$\begin{aligned}\rho(x;t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx)\tilde{\rho}(k;0)\exp(-Dk^2t) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}kx)\left[\int_{-\infty}^{\infty} \mathrm{d}x' \exp(\mathrm{i}kx')\rho(x';0)\right]\exp(-Dk^2t) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x'\rho(x';0)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}k(x'-x)-Dk^2t)\right] \\ &= \int_{-\infty}^{\infty} \mathrm{d}x'\rho(x';0)G(x'-x;t),\end{aligned} \tag{13.37}$$

donde hemos definido $G(x'-x;t)$ como

$$\begin{aligned}G(x'-x;t) &:= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}k(x'-x)-Dk^2t) \\ &= \frac{1}{\sqrt{4\pi Dt}}\exp\Big(-\frac{(x'-x)^2}{4Dt}\Big).\end{aligned} \tag{13.38}$$

Finalmente tenemos que la solución general de la ecuación de difusión (13.31) es

$$\rho(x;t) = \frac{1}{\sqrt{4\pi Dt}}\int_{-\infty}^{\infty} \mathrm{d}x'\rho(x';0)\exp\Big(-\frac{(x'-x)^2}{4Dt}\Big). \tag{13.39}$$

En el caso en que el proceso de difusión comienza desde un valor inicial $X_0 = x_0$ conocido exactamente, tenemos $\rho(x;0) = \delta(x-x_0)$ y entonces podemos escribir

$$\boxed{P(X_t = x|X_0 = x_0, I) = \frac{1}{\sqrt{4\pi Dt}}\exp\Big(-\frac{(x-x_0)^2}{4Dt}\Big),} \tag{13.40}$$

es decir, $X_t \sim \mathcal{N}(x_0, 2Dt)$. Más aún, usando (13.40) podemos entender la convolución en (13.39) como la aplicación de la ecuación de Chapman-Kolmogorov

—que de hecho es simplemente la regla de marginalización— para eliminar la información del valor inicial x_0, ya que reescribimos (13.39) como

$$P(X_t = x|I) = \int dx_0 P(X_0 = x_0|I)P(X_t = x|X_0 = x_0, I). \tag{13.41}$$

Como la distribución (13.40) es normal, se sigue de inmediato que

$$\langle X_t \rangle_{X_0=x_0,I} = x_0, \tag{13.42a}$$

$$\left\langle (X_t - X_0)^2 \right\rangle_{X_0=x_0,I} = 2Dt, \tag{13.42b}$$

y si hacemos que t sea muy cercano a cero, podemos escribir

$$\frac{1}{2}\lim_{t\to 0}\frac{1}{t}\left\langle (X_t - X_0)^2 \right\rangle_{X_0=x_0,I} = D, \tag{13.43}$$

en completo acuerdo con (13.29). Además, confirmamos que $\mu(x) = 0$ ya que

$$\mu(x) = \lim_{t\to 0}\frac{1}{t}\langle X_t - X_0 \rangle_{X_0=x_0,I} = \lim_{t\to 0}\frac{1}{t}\Big(\langle X_t \rangle_{X_0=x_0,I} - x_0\Big) = 0. \tag{13.44}$$

Una perspectiva alternativa —pero completamente equivalente del punto de vista formal— consiste en tratar el movimiento browniano como la suma de incrementos sucesivos $(\Delta X)_i$ de manera que

$$\begin{aligned} X_n &= X_{n-1} + (\Delta X)_{n-1} \\ &= X_{n-2} + (\Delta X)_{n-2} + (\Delta X)_{n-1} \\ &\vdots \\ &= X_0 + \sum_{i=0}^{n-1} (\Delta X)_i. \end{aligned} \tag{13.45}$$

Notemos que aquí n no es un tiempo continuo como en el caso de la ecuación de difusión sino es el número de pasos de tiempo Δt, es decir,

$$t = n\Delta t$$

con $\Delta t \to 0$. Si los incrementos ΔX son a su vez suficientemente pequeños y tienen media cero, entonces

$$\langle (\Delta X)_i \rangle_I = 0, \tag{13.46}$$

$$\langle (\Delta X)_i^2 \rangle_I = \sigma^2. \tag{13.47}$$

y se sigue que la distribución de probabilidad para $(\Delta X)_n$ debe ser normal,

$$P((\Delta X)_i = \Delta x|I) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(\Delta x)^2}{2\sigma^2}\Big) \qquad \text{para todo } i. \tag{13.48}$$

Usando la definición del coeficiente de difusión, podemos determinar el valor de σ^2 como

$$\begin{aligned} D &= \frac{1}{2}\lim_{\Delta\to 0}\frac{1}{\Delta t}\langle (X_{t+\Delta t} - X_t)^2 \rangle_{X_t=x,I} \\ \text{(usando } t = n\Delta t\text{)} \quad &= \lim_{\Delta t\to 0}\frac{1}{2\Delta t}\langle (\Delta X)_n^2 \rangle_I \\ &= \lim_{\Delta t\to 0}\Big(\frac{\sigma^2}{2\Delta t}\Big), \end{aligned} \tag{13.49}$$

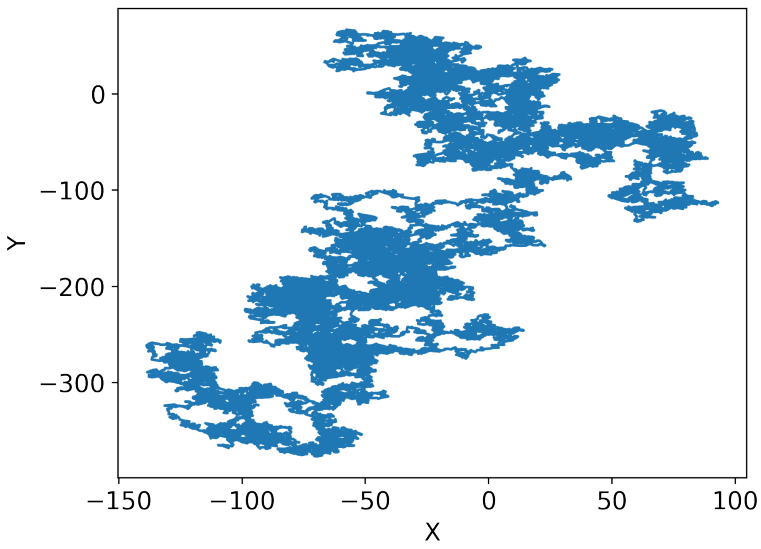


Figura 13.2: Movimiento browniano en dos dimensiones.

y ya que D es finito, debemos imponer que $\sigma^2 \to 0$ cuando $\Delta t \to 0$, de forma que

$$\sigma^2 = 2D\Delta t. \tag{13.50}$$

Ahora, como $X_n - X_0$ es una suma de variables normales, es a su vez una variable normal con

$$\langle X_n - X_0 \rangle_I = n\langle \Delta X \rangle_I = 0, \tag{13.51}$$

$$\langle (X_n - X_0)^2 \rangle_I = n\sigma^2 = 2D\Delta t\, n = 2Dt, \tag{13.52}$$

y entonces se obtiene

$$P(X_t = x | X_0 = x_0, I) = \frac{1}{\sqrt{4\pi Dt}} \exp\Big(- \frac{(x - x_0)^2}{4Dt} \Big), \tag{13.53}$$

que coincide con (13.40). En $d > 1$ dimensiones, simplemente reemplazamos (13.45) por

$$\boldsymbol{X}_n = \boldsymbol{X}_0 + \sum_{i=0}^{n-1} (\Delta \boldsymbol{X})_i, \tag{13.54a}$$

con

$$(\Delta \boldsymbol{X})_i = \boldsymbol{X}_i - \boldsymbol{X}_{i-1}. \tag{13.54b}$$

Dado que las d componentes son independientes entre sí, las propiedades de los desplazamientos $(\Delta \boldsymbol{X})_i$ son

$$\langle (\Delta \boldsymbol{X})_i \rangle_I = \boldsymbol{0}, \tag{13.55}$$

$$\Big\langle (\Delta \boldsymbol{X}_i)^2 \Big\rangle_I = d\sigma^2 = 2d\, D\Delta t. \tag{13.56}$$

Un ejemplo de movimiento browniano en dos dimensiones se muestra en la Figura 13.2.

Tenemos entonces que la distribución de $\boldsymbol{X}_t$ es el producto de las distribuciones por componente,

$$\begin{aligned} P(\boldsymbol{X}_t = \boldsymbol{x}|\boldsymbol{X}_0 = \boldsymbol{x}_0, I) &= \prod_{\alpha=1}^{d} \Big\{ \frac{1}{\sqrt{4\pi D t}} \exp\Big(- \frac{(\boldsymbol{e}_\alpha \cdot [\boldsymbol{x} - \boldsymbol{x}_0])^2}{4Dt} \Big) \Big\} \\ &= \frac{1}{(\sqrt{4\pi D t})^d} \exp\Big(- \frac{1}{4Dt} \sum_\alpha (\boldsymbol{e}_\alpha \cdot [\boldsymbol{x} - \boldsymbol{x}_0])^2 \Big) \\ &= \frac{1}{(\sqrt{4\pi D t})^d} \exp\Big(- \frac{(\boldsymbol{x} - \boldsymbol{x}_0)^2}{4Dt} \Big), \end{aligned} \tag{13.57}$$

y entonces podemos escribir el resultado como la distribución normal multivariable

$$P(\boldsymbol{X}_t = \boldsymbol{x}|\boldsymbol{X}_0 = \boldsymbol{x}_0, I) = \frac{1}{(\sqrt{4\pi D t})^d} \exp\Big(- \frac{(\boldsymbol{x} - \boldsymbol{x}_0)^2}{4Dt} \Big), \tag{13.58}$$

que tiene media $\boldsymbol{\mu} = \boldsymbol{x}_0$ y matriz de covarianza

$$\Sigma_{ij} = 2Dt\delta(i,j), \tag{13.59}$$

proporcional a la identidad. Podemos obtener el desplazamiento cuadrático medio simplemente derivando el logaritmo de la constante de normalización respecto a t,

$$\frac{\partial}{\partial t} \ln(\sqrt{4\pi D t})^d = \frac{d}{2t} = \frac{1}{4Dt^2} \Big\langle (\boldsymbol{X}_t - \boldsymbol{X}_0)^2 \Big\rangle_I, \tag{13.60}$$

esto es,

$$\big\langle (\boldsymbol{X}_t - \boldsymbol{X}_0)^2 \big\rangle_I = 2dDt, \tag{13.61}$$

que por supuesto coincide con la suma de n desplazamientos cuadráticos individuales,

$$\sum_{i=0}^{n-1} \Big\langle (\Delta \boldsymbol{X}_i)^2 \Big\rangle_I = 2d\, Dn\Delta t = 2dDt. \tag{13.62}$$

13.2.2 Caminata al azar en una red periódica

Sea $\boldsymbol{r}_0$ la posición inicial de un caminante en una red periódica, como se ve en la **Figura 13.3**, y sea $\boldsymbol{r}_i$ la posición luego del paso i-ésimo. Consideraremos el caso general para d dimensiones. Al igual que en el caso del movimiento browniano, al paso n para una realización particular de la caminata la posición $\boldsymbol{r}_n$ del caminante estará dada por

$$\boldsymbol{r}_n = \boldsymbol{r}_0 + \sum_{i=1}^{n} (\Delta \boldsymbol{r})_i, \tag{13.63}$$

donde el vector $\Delta \boldsymbol{r}_i$ es el desplazamiento asociado al paso i-ésimo, el cual ahora es una variable discreta, tomando uno de m posibles vectores que unen los sitios de la red. Esto es, se tiene que $\Delta \boldsymbol{r} \in \{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_m\}$.

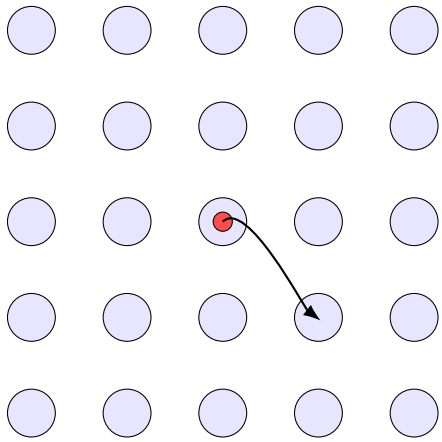


Figura 13.3: Una red periódica sobre la cual se mueve un caminante.

Nos interesa la probabilidad $P(\boldsymbol{r}_n = \boldsymbol{r}|I)$ de que el caminante se encuentre en el sitio $\boldsymbol{r}$ luego del paso n-ésimo, que calculamos como

$$\begin{aligned} P(\boldsymbol{r}_n = \boldsymbol{r}|I) &= \langle \delta(\boldsymbol{r}_n - \boldsymbol{r}) \rangle_I \\ &= \sum_{\{\Delta \boldsymbol{r}\}} P(\Delta \boldsymbol{r}_1, \ldots, \Delta \boldsymbol{r}_n|I) \delta\Big(\boldsymbol{r}_0 + \sum_{i=1}^{n} (\Delta \boldsymbol{r})_i - \boldsymbol{r}\Big) \end{aligned} \tag{13.64}$$

donde

$$\sum_{\{\Delta \boldsymbol{r}\}} f(\Delta \boldsymbol{r}_1, \ldots, \Delta \boldsymbol{r}_n) = \sum_{\Delta \boldsymbol{r}_1} \cdots \sum_{\Delta \boldsymbol{r}_n} f(\Delta \boldsymbol{r}_1, \ldots, \Delta \boldsymbol{r}_n)$$

para cualquier función f de los desplazamientos. Para poder avanzar en el cálculo, introduciremos a continuación la *función de estructura* de una red.

Definición 13.2 — Factor de estructura
Definimos el factor de estructura $\lambda(\boldsymbol{k})$ de una red periódica como

$$\begin{aligned} \lambda(\boldsymbol{k}) &:= \sum_{\Delta \boldsymbol{r}} P(\Delta \boldsymbol{r}|I) \exp(\mathrm{i}\boldsymbol{k} \cdot \Delta \boldsymbol{r}) \\ &= \sum_{j=1}^{m} P(\Delta \boldsymbol{r} = \boldsymbol{R}_j|I) \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{R}_j) \end{aligned} \tag{13.65}$$

Dado que los desplazamientos en los distintos pasos son independientes entre sí, podemos separar

$$P(\Delta \boldsymbol{r}_1, \ldots, \Delta \boldsymbol{r}_n|I) = \prod_{i=1}^{n} P(\Delta \boldsymbol{r}_i|I), \tag{13.66}$$

y usando la representación (3.47) de la delta de Dirac, tenemos

$$\begin{aligned} P(\boldsymbol{r}_n = \boldsymbol{r}|I) = \sum_{\{\Delta \boldsymbol{r}\}} \prod_{i=1}^{n} P(\Delta \boldsymbol{r}_i|I) \\ \times \left[\frac{1}{(2\pi)^d} \int \mathrm{d}\boldsymbol{k} \exp\Big(-\mathrm{i}\boldsymbol{k} \cdot \big(\boldsymbol{r} - \boldsymbol{r}_0 - \sum_{j=1}^{n} (\Delta \boldsymbol{r})_j\big)\Big) \right]. \end{aligned} \tag{13.67}$$

Separando la exponencial en sus factores, escribimos

$$\begin{aligned} P(\boldsymbol{r}_n = \boldsymbol{r}|I) &= \frac{1}{(2\pi)^d}\int \mathrm{d}\boldsymbol{k}\exp(-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}_0))\sum_{\{\Delta\boldsymbol{r}\}}\prod_{i=1}^{n}P(\Delta\boldsymbol{r}_i|I)\exp\Big(\mathrm{i}\boldsymbol{k}\cdot\sum_{j=1}^{n}(\Delta\boldsymbol{r})_j\Big) \\ &= \frac{1}{(2\pi)^d}\int \mathrm{d}\boldsymbol{k}\exp(-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}_0))\sum_{\{\Delta\boldsymbol{r}\}}\prod_{i=1}^{n}P(\Delta\boldsymbol{r}_i|I)\exp(\mathrm{i}\boldsymbol{k}\cdot\Delta\boldsymbol{r}_i) \\ &= \frac{1}{(2\pi)^d}\int \mathrm{d}\boldsymbol{k}\exp(-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}_0))\Big(\sum_{\Delta\boldsymbol{r}}P(\Delta\boldsymbol{r}|I)\exp(\mathrm{i}\boldsymbol{k}\cdot\Delta\boldsymbol{r})\Big)^n, \end{aligned} \tag{13.68}$$

por lo tanto, usando la definición de $\lambda(\boldsymbol{k})$ finalmente llegamos a

$$\boxed{P(\boldsymbol{r}_n = \boldsymbol{r}|I) = \frac{1}{(2\pi)^d}\int \mathrm{d}\boldsymbol{k}\exp\big(-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}_0)\big)\lambda(\boldsymbol{k})^n.} \tag{13.69}$$

Notemos que esto no es más que la solución del problema de suma de variables por medio de funciones características (sección 8.4), ya que $\lambda(\boldsymbol{k})$ es precisamente la función característica asociada a los desplazamientos $\Delta\boldsymbol{r}$ y buscamos la distribución de la suma de dichos desplazamientos.

Podemos verificar esta solución al considerar, para una dimensión, desplazamientos $\Delta x \sim \mathcal{N}(0,\sigma^2)$ lo cual produce un factor de estructura

$$\begin{aligned} \lambda(k) &= \sum_{\Delta x}P(\Delta x|I)\exp(\mathrm{i}k\Delta x) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\mathrm{d}x\exp\Big(-\frac{(\Delta x)^2}{2\sigma^2}\Big)\exp(\mathrm{i}k\Delta x) \\ \text{(usando } u = k\Delta x\text{)} \quad &= \frac{1}{\sqrt{2\pi}\sigma k}\int_{-\infty}^{\infty}\mathrm{d}u\exp\Big(-\frac{u^2}{2k^2\sigma^2}\Big)\exp(\mathrm{i}u) \\ &= \exp\Big(-\frac{k^2\sigma^2}{2}\Big), \end{aligned} \tag{13.70}$$

el cual al reemplazarlo en (13.69) produce

$$\begin{aligned} P(X_n = x|I) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\exp\Big(-\mathrm{i}k(x-x_0)-\frac{nk^2\sigma^2}{2}\Big) \\ &= \frac{1}{\sqrt{2\pi n}\sigma}\exp\Big(-\frac{(x-x_0)^2}{2n\sigma^2}\Big), \end{aligned} \tag{13.71}$$

que coincide con (13.53) si sustituimos $n\sigma^2 = 2Dt$ de acuerdo a (13.52). Muchas veces para evaluar (13.69) es más fácil considerar $P(\boldsymbol{r}_n = \boldsymbol{r}|I)$ como una secuencia y construir su función generadora,

$$F(z;\boldsymbol{r}) := \sum_{n=0}^{\infty}P(\boldsymbol{r}_n = \boldsymbol{r}|I)z^n, \tag{13.72}$$

la cual se reduce a

$$\begin{aligned} F(z;\boldsymbol{r}) &= \frac{1}{(2\pi)^d}\int \mathrm{d}\boldsymbol{k}\exp\big(-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}_0)\big)\sum_{n=0}^{\infty}z^n\lambda(\boldsymbol{k})^n \\ \text{usando (14)} \quad &= \frac{1}{(2\pi)^d}\int \mathrm{d}\boldsymbol{k}\left[\frac{\exp\big(-\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}_0)\big)}{1-z\lambda(\boldsymbol{k})}\right], \end{aligned} \tag{13.73}$$

donde hemos usado la serie geométrica en la segunda línea.

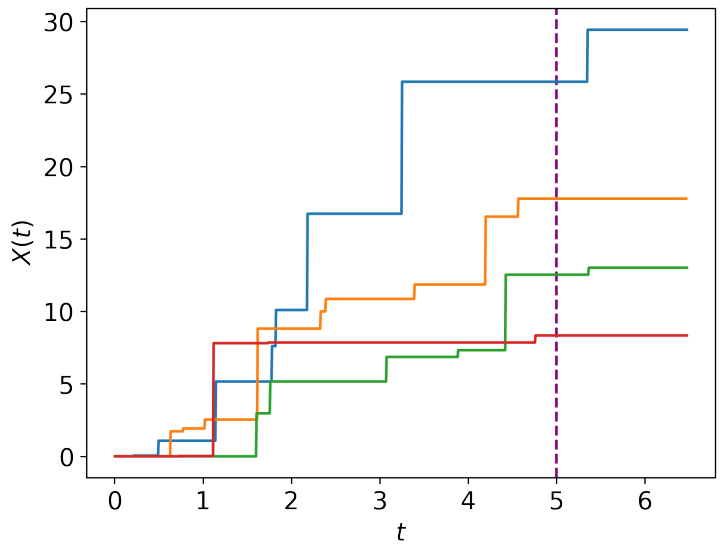


Figura 13.4: Distintas realizaciones de una caminata al azar de tiempo continuo. Al tiempo $t = 5$ el caminante llega a distintas posiciones y completa un número variable de saltos.

13.2.3 Caminatas al azar con tiempo continuo

En una caminata al azar con tiempo continuo se levanta la suposición de que los saltos se realizan a intervalos de tiempo fijos Δt, ahora considerando estos intervalos como variables con distribución $P(\Delta t|I)$ pero siempre manteniendo la suposición de que *el caminante se encuentra inmóvil entre saltos*, como se ve en la **Figura 13.4**. El tiempo total T_n transcurrido luego de n pasos ya no será $T_n = n\Delta t$ sino

$$T_n := \sum_{i=1}^{n} (\Delta t)_i, \tag{13.74}$$

mientras que, considerando por simplicidad el caso en una dimensión, la posición al paso n seguirá siendo

$$X_n := x_0 + \sum_{i=1}^{n} (\Delta X)_i. \tag{13.75}$$

Ahora podemos preguntarnos por la probabilidad de que el caminante se encuentre en la coordenada x al tiempo $t \leq 0$, que escribiremos

$$P(X(t) = x|I).$$

Como llegar a la coordenada x puede suceder antes del primer paso (si $x = x_0$), o entre el primer y el segundo paso, o entre el segundo y el tercer paso y así sucesivamente, definiendo la proposición

$$A_n(x,t) := (X_n = x) \wedge (T_n < t \leq T_{n+1}) \tag{13.76}$$

tendremos que

$$P(X(t) = x|I) = P(A_1(x,t) \vee A_2(x,t) \vee A_3(x,t) \vee \ldots |I). \tag{13.77}$$

Notando que las proposiciones A_n son mutuamente excluyentes, y suponiendo que X_n es independiente de T_n, podemos escribir

$$\begin{aligned} P(X(t) = x|I) &= \sum_{n=0}^{\infty} P(A_n(x,t)|I) \\ &= \sum_{n=0}^{\infty} P(X_n = x|I) P(T_n \leq t \leq T_{n+1}|I). \end{aligned} \tag{13.78}$$

Nuevamente nuestro problema involucra sumas de variables, ahora incorporando T como una variable que se acumula en incrementos Δt. Definiendo la función $\psi(s)$ como la transformada de Laplace[2] de la distribución $P(\Delta t = t|I)$,

$$\psi(s) := \left\langle \exp(-s\Delta t) \right\rangle_I = \int_0^{\infty} \mathrm{d}t \exp(-st) P(\Delta t = t|I) \tag{13.79}$$

vemos que para n intervalos de tiempo se cumple

$$\left\langle \exp(-sT_n) \right\rangle_I = \left\langle \exp\Big(-s \sum_{i=1}^{n} (\Delta t)_i \Big) \right\rangle_I = \psi(s)^n. \tag{13.80}$$

Para resolver $P(X(t) = t|I)$ dadas las distribuciones $P(\Delta X|I)$ y $P(\Delta t|I)$ expresamos la transformada en X y en T como

$$\begin{aligned} \tilde{p}(k,s) &:= \Big\langle \exp(\mathrm{i}k(X - x_0) - sT) \Big\rangle_I \\ &= \int_{-\infty}^{\infty} \mathrm{d}x \int_0^{\infty} \mathrm{d}t P(X(t) = x|I) \exp(\mathrm{i}k(x - x_0) - st) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x \int_0^{\infty} \mathrm{d}t \sum_{n=0}^{\infty} P(X_n = x|I) P(T_n \leq t \leq T_{n+1}|I) \exp(\mathrm{i}k(x - x_0) - st) \\ &= \sum_{n=0}^{\infty} \lambda(k)^n \int_0^{\infty} \mathrm{d}t P(T_n \leq t \leq T_{n+1}|I) \exp(-st). \end{aligned} \tag{13.81}$$

Aquí usamos la forma explícita de $P(T_n \leq t \leq T_{n+1}|I)$ como expectación de la función rectangular para evaluar

$$\begin{aligned} &\int_0^{\infty} \mathrm{d}t P(T_n \leq t \leq T_{n+1}|I) \exp(-st) \\ &\quad = \int_0^{\infty} \mathrm{d}t \left\langle \mathrm{rect}(t; T_n, T_{n+1}) \right\rangle_I \exp(-st) \\ &\quad = \left\langle \int_0^{\infty} \mathrm{d}t \, \mathrm{rect}(t; T_n, T_{n+1}) \exp(-st) \right\rangle_I \\ &\quad = \left\langle \frac{1}{s} \Big(\exp(-sT_n) - \exp(-sT_{n+1}) \Big) \right\rangle_I \\ &\quad = \frac{1}{s} \Big(\psi(s)^n - \psi(s)^{n+1} \Big) \\ &\quad = \frac{\psi(s)^n}{s} \Big(1 - \psi(s) \Big), \end{aligned} \tag{13.82}$$

[2] Esta función $\psi(s)$ corresponde a la función generadora de momentos de la distribución de Δt.

por lo que reemplazando en (13.81) finalmente tenemos

$$\begin{aligned}\tilde{p}(k,s) &= \sum_{n=0}^{\infty} \lambda(k)^n \left[\frac{\psi(n)^n}{s}\Big(1-\psi(s)\Big)\right] \\ \text{usando (14)} \quad &= \frac{1-\psi(s)}{s}\frac{1}{1-\lambda(k)\psi(s)}.\end{aligned} \tag{13.83}$$

A este resultado se le conoce como la *fórmula de Montroll-Weiss*,

$$\Big\langle \exp(\mathrm{i}k(X-x_0)-sT)\Big\rangle_I = \frac{1-\psi(s)}{s}\frac{1}{1-\lambda(k)\psi(s)}. \tag{13.84}$$

Una vez calculada $\tilde{p}(k,s)$ como función de k y de s, podemos recuperar la información de la posición X mediante

$$\begin{aligned}\Big\langle \delta(X-x)\exp(-sT)\Big\rangle_I &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}k \exp(-\mathrm{i}k(x-x_0))\tilde{p}(k,s) \\ &= \frac{1-\psi(s)}{2\pi s}\int_{-\infty}^{\infty} \mathrm{d}k \frac{\exp(-\mathrm{i}k(x-x_0))}{1-\lambda(k)\psi(s)}.\end{aligned} \tag{13.85}$$

▶ Para más detalles sobre el formalismo de caminatas al azar, ver el ya clásico libro de Hughes (1995) y el más moderno de Klafter y Sokolov (2011).

Problemas

Problema 13.1. *Calcule, usando (13.73), la probabilidad de estar en el origen al paso n para una caminata al azar en una red unidimensional de paso a, con*

$$P(\Delta X = a|I) = P(\Delta X = -a|I) = \tfrac{1}{2}.$$

Problema 13.2. *Resuelva el problema de la caminata al azar en la red unidimensional de paso $a = 1$, directamente obteniendo la distribución de X_n donde*

$$X_{i+1} = X_i + (\Delta X)_i$$

tal que $(\Delta X)_i \in \{-1,1\}$ y $P(\Delta X = 1|I) = P(\Delta X = -1|I) = \frac{1}{2}$. ¿Cuál es la probabilidad de estar en el origen al paso n? Compare su solución con la del Problema 13.1.

CAPÍTULO 14

Simulación Monte Carlo

> Part of the inhumanity of the computer is that, once it is competently programmed and working smoothly, it is completely honest.
>
> Isaac Asimov

Aunque hemos visto que las herramientas de la probabilidad y expectación nos entregan *en principio* respuestas claras, no siempre es sencillo o incluso viable obtenerlas en la práctica, muchas veces porque necesitamos evaluar expresiones sin una forma analítica cerrada. Un ejemplo claro son las integrales necesarias para evaluar distribuciones posteriores o distribuciones predictivas. En estos casos bien podemos recurrir a aproximaciones de dichas soluciones, como por ejemplo el uso de la aproximación de Laplace, o a códigos computacionales que implementan métodos numéricos de solución.

Una clase muy importante dentro de estos métodos numéricos la constituyen los llamados métodos de simulación Monte Carlo, cuyo nombre hace referencia al casino de Monte Carlo, en el principado de Mónaco. Estos métodos fueron desarrollados en los inicios de la computación científica, motivados por la necesidad de resolver problemas de naturaleza militar, —en particular el estudio de reacciones nucleares tanto de fisión como de fusión que llevarían a la bomba atómica y a la bomba de hidrógeno— cuyo secreto hacía necesario un nombre código. Uno de sus desarrolladores originales fue Nicholas Metropolis, en honor a quien se ha bautizado al algoritmo de Metropolis que veremos en la **Sección 14.4**.

Los métodos Monte Carlo se basan en la generación computacional de *números pseudoaleatorios* para evaluar expectaciones por medio de la ley de los grandes números. Por ejemplo, supongamos que deseamos calcular la integral multidimensional

$$Z = \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{x} f(\boldsymbol{x}) \tag{14.1}$$

y que podemos factorizar $f(\boldsymbol{x})$ como

$$f(\boldsymbol{x}) = p(\boldsymbol{x})A(\boldsymbol{x}),$$

donde $p(\boldsymbol{x}) \geq 0$ y es posible generar números pseudoaleatorios $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ distribuidos según $P(\boldsymbol{X} = \boldsymbol{x}|I) = p(\boldsymbol{x})$. Entonces podemos reinterpretar Z como la expectación de A y usar la ley de los grandes números en la forma

$$Z = \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{x} P(\boldsymbol{X} = \boldsymbol{x}|I) A(\boldsymbol{x}) \approx \frac{1}{n} \sum_{i=1}^{n} A(\boldsymbol{x}_i) \tag{14.2}$$

para n suficientemente grande, como lo vimos en el **Ejemplo 9.3.1**.

▶ Para más información sobre los métodos de Monte Carlo en general, ver el libro de Gould, Tobochnik y Christian (2007) y el de Landau y Binder (2015).

14.1 — Números pseudoaleatorios

Cuando decimos números pseudoaleatorios nos referimos a que, a pesar de ser generados por un dispositivo determinista como es nuestro computador, para todos los efectos es muy difícil —en la práctica imposible— predecir sus valores. Un computador digital genera una cadena o secuencia de enteros pseudoaleatorios $(r_0, r_1, r_2, \ldots)$ comenzando a partir de una semilla r_0 que debe ser proporcionada, y tal que la secuencia completa es una función determinista de dicha semilla, esto es, la misma semilla siempre producirá la misma secuencia. Un ejemplo sencillo de generador de números enteros pseudoaleatorios es una familia de algoritmos denominados generadores lineales congruentes, donde la secuencia está dada por una relación de recurrencia de la forma

$$r_{n+1} = (a\, r_n + c) \bmod m \tag{14.3}$$

con m es un entero muy grande, y donde a y c son enteros no negativos menores que m, de forma que $0 \leq r_i < m$.
Por ejemplo, el lenguaje de programación C usa

$$m = 2^{31}, \tag{14.4a}$$
$$a = 1103515245, \tag{14.4b}$$
$$c = 12345. \tag{14.4c}$$

Un buen generador de números pseudoaleatorios está diseñado para tener una distribución uniforme de frecuencias de sus valores posibles. En el caso de (14.3) con los valores en (14.4) esto mayormente se cumple. Usando estos números pseudoaleatorios enteros es posible generar números pseudoaleatorios uniformes $X \sim \mathrm{U}(0,1)$ simplemente a través de

$$X_n := \frac{r_n}{m}, \tag{14.5}$$

y para tener una precisión adecuada (32 *bits* en este caso) en el valor X_n generado es importante que el valor de m sea suficientemente grande.

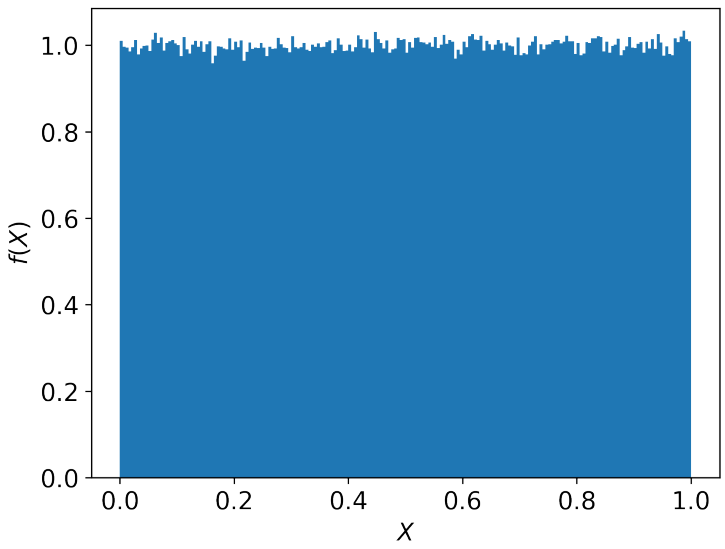


Figura 14.1: Histograma de un millón de números pseudoaleatorios uniformes en $[0, 1)$.

La **Figura 14.1** muestra un histograma de números pseudoaleatorios uniformes, obtenidos mediante un generador lineal congruente de acuerdo a (**14.3**) y (**14.4**).

▶ Para conocer los detalles computacionales acerca de la generación de números pseudoaleatorios, incluyendo las técnicas más modernas se recomienda el libro *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Press, Teukolsky, Vetterling y Flannery 2007).

14.2 — Generación de eventos con probabilidades dadas

Supongamos que queremos simular el lanzamiento de una moneda, o de un dado, o una caminata al azar, entre muchos otros ejemplos. Para ello necesitaremos que nuestro programa computacional decida entre distintas opciones $A_1, A_2, \ldots, A_n$ donde $P(A_i|I) = p_i$. ¿Cómo conseguimos esto?

En primer lugar, veamos el caso donde existen sólo dos opciones, A y $\neg A$, con $P(A|I) = p \in [0, 1]$. En términos simples, el evento sólo se activa si al «sacar» el número de $\mathrm{U}(0, 1)$, éste resulta ser menor que p, de forma que si $p \to 0$ más difícil es realizarlo, y por el contrario cuando $p \to 1$ prácticamente todos los números de $\mathrm{U}(0, 1)$ serán menores que p, y el evento se activará prácticamente siempre. Este procedimiento asegura que el evento ocurre con frecuencia igual a p.

Un fragmento de código sencillo en Python para generar una secuencia de n eventos de este tipo es el siguiente, donde `random()` es la función en Python que devuelve un número pseudoaleatorio uniforme en $[0, 1)$.

Programa 14.1 — Generación de eventos con probabilidad p
La función `EscogerEvento` recibe un número `p` entre 0 y 1, para luego retornar 1 con probabilidad p y 0 con probabilidad $1 - p$.

```
from random import random

def EscogerEvento(p):
    r = random()
    if r < p: return 1
    else: return 0
```

Ahora generalicemos para el caso con n opciones $A_1, A_2, \ldots, A_n$ donde $P(A_i|I) = p_i$ tal que $\sum_{i=1}^{n} p_i = 1$. Usando la misma regla anterior, en este caso el algoritmo va haciendo comparaciones sucesivas hasta que una de ellas tiene éxito.

```
from random import random

def EscogerEvento(p):
    r = random()
    if r < p[0]: return 0
    elif r - p[0] < p[1]: return 1
    elif r - p[0] - p[1] < p[2]: return 2
    ...
    elif r - sum(p[i] for i in range(n-3)) < p[n-2]: return (n-2)
    else: return (n-1)
```

> **Importante:** Notemos que en el código anterior los elementos del arreglo `p` comienzan desde cero, y luego p_1 corresponde a `p[0]`, p_2 a `p[1]`, y así sucesivamente.

De fallar la primera comparación (que activaría el evento A_1) se desplaza el punto inicial de la comparación, que originalmente sería cero, para que ahora sea $r - p_1$, de forma que éste se compara con p_2. Si esta nueva comparación falla, se compara $r - p_1 - p_3$ con p_3, y así sucesivamente hasta que en último caso ocurre el evento A_n. Este procedimiento asegura que los eventos se producen con las frecuencias correspondientes a las probabilidades dadas. Una manera más elegante de escribir el mismo procedimiento utiliza un ciclo `for`, para así extender el algoritmo a n arbitrariamente grande (dentro de las capacidades de tiempo de cómputo, por supuesto).

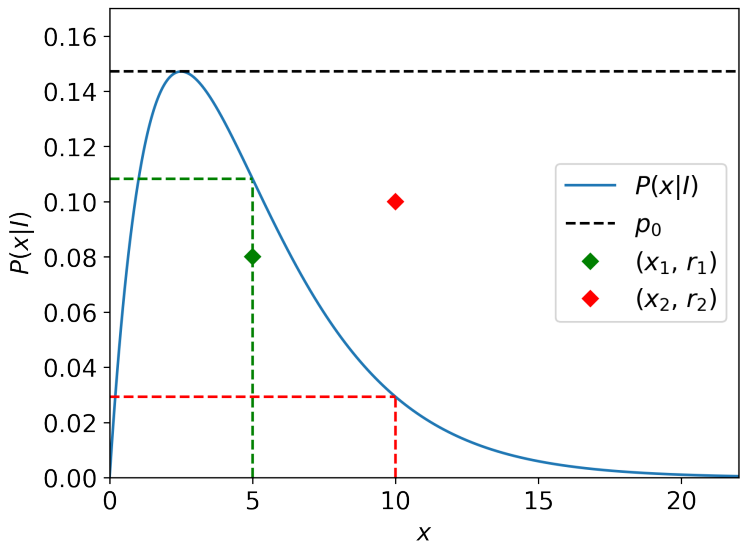


Figura 14.2: Método de la aceptación y rechazo. En este ejemplo, la muestra $x = 5$ (verde) es aceptada, mientras que $x = 10$ (rojo) resulta rechazada.

Programa 14.2 — Generación de eventos con probabilidades dadas

```
from random import random

def EscogerEvento(p):
    r = random()
    n = len(p)
    for k in range(n):
        if r < p[k]: return k
        else: r = r - p[k]
```

14.3 — Método de la aceptación y rechazo

El método de la aceptación y rechazo es un método para generar números pseudoaleatorios para una variable discreta o continua $X \sim I$ dada la distribución de probabilidad $P(X = x|I) = p(x)$, y que se basa en el procedimiento descrito en el programa (14.1).

Si denominamos p_0 al valor máximo de $p(x)$, entonces podemos formular el método de aceptación y rechazo como sigue. En primer lugar, escogemos desde una distribución uniforme un valor de x, el cual será aceptado o rechazado según el siguiente criterio: si un número $r \sim \mathrm{U}(0,1)$ es tal que $r \cdot p_0 < p(x)$, entonces x es aceptado, de otra forma x es rechazado y se intenta un nuevo valor de x.

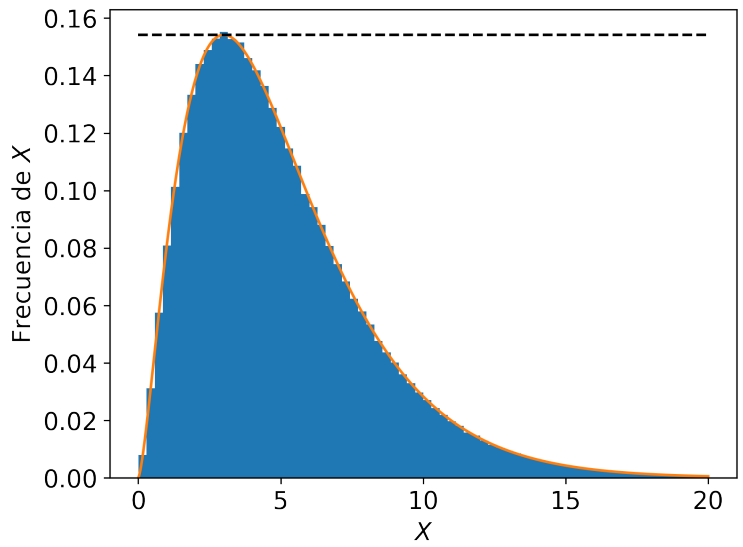


Figura 14.3: Aplicación del método de la aceptación y rechazo. El histograma en azul fue construido usando 1 millón de valores entre 0 y 20 aceptados de acuerdo a una distribución gamma con $k = 5/2$ y θ=2, la cual se muestra en la curva naranja. La línea negra punteada representa el valor p_0, correspondiente al máximo de la densidad de probabilidad.

Programa 14.3 — Método de aceptación y rechazo

```
from random import random

def GeneraMuestras(N, a, b, p, p0):
    muestras = list()
    while len(muestras) < N:
          x = (b-a)*random() + a
          r = random()
          if r*p0 < p(x): muestras.append(x)
    return muestras
```

Este procedimiento, que se detalla en el Programa (14.3) y gráficamente en la Figura 14.2, asegura que las muestras aceptadas son producidas con una frecuencia que coincide con $p(x)$.

14.4 — Algoritmo Metropolis

Si quisiéramos extender la idea del método de aceptación y rechazo a más dimensiones, en principio esto es inmediato: simplemente generamos propuestas $\boldsymbol{x}$ tal que

$$r \cdot p_0 < P(\boldsymbol{X} = \boldsymbol{x}|I),$$

sin embargo el problema es uno de eficiencia numérica. Dado que, si estamos en n dimensiones, debemos generar muestras en un hipercubo de volumen L^n —donde $L = b - a$ era el largo del intervalo para el caso en una dimensión— y entre ellas seleccionar las que «caen» bajo la distribución (como la muestra en verde de la Figura 14.2), y la fracción de muestras aceptadas

f_n va como

$$f_n \approx \left(\frac{A_p}{Lp_0}\right)^n, \tag{14.6}$$

con $A_p < Lp_0$, se sigue que para n muy grande la fracción f_n va rápidamente a cero. Por ejemplo, si $A_p \approx \frac{1}{3}Lp_0$, entonces para $n = 10$ tendríamos

$$f_n \approx 1.7 \cdot 10^{-5},$$

esto es, para conseguir un punto aceptado deberíamos intentar más de 59 mil veces. Para $n = 100$ ¡sólo 1 puntos por cada 10^{47} intentos serían aceptados!

A este fenómeno se le denomina la *maldición de la dimensionalidad*: la dificultad de muestrear un espacio aumenta exponencialmente con el número de dimensiones. Una solución a este problema se conoce como el algoritmo Metropolis, presentado por primera vez en el artículo "Equation of state calculations by fast computing machines" (1953), y que se basa en la construcción de una cadena de Markov tal que en su estado estacionario las frecuencias de los estados $\boldsymbol{x}$ visitados por ésta se aproximan a la probabilidad

$$p(\boldsymbol{x}) := P(\boldsymbol{X} = \boldsymbol{x}|I)$$

deseada.
El algoritmo Metropolis consiste en la propuesta sucesiva de estados

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + (\Delta\boldsymbol{x})_i \tag{14.7}$$

a partir de un estado inicial o «semilla» $\boldsymbol{x}_0$, propuestas que son aceptadas con la siguiente probabilidad.

Definición 14.1 — Tasa de aceptación de Metropolis

$$A_M(\boldsymbol{x} \to \boldsymbol{x}') := \text{mín}\left(1, \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x})}\right). \tag{14.8}$$

Esta tasa de aceptación implica que una movida es aceptada de inmediato si $p(\boldsymbol{x}') > p(\boldsymbol{x})$, es decir cuando el sistema salta hacia un estado más probable que el actual, mientras que es *aceptada condicionalmente* en caso contrario, de acuerdo al valor de la razón $p(\boldsymbol{x}')/p(\boldsymbol{x})$. Una característica muy afortunada del método es que, como puede expresarse únicamente en términos del cuociente $p(\boldsymbol{x}')/p(\boldsymbol{x})$, la distribución $p(\boldsymbol{x})$ no necesita estar correctamente normalizada, lo cual se agradece en el caso de muchas dimensiones, donde calcular la constante de normalización implicaría un problema tan complejo como el que tratamos de resolver.

El algoritmo Metropolis en su versión original supone que los desplazamientos $\Delta\boldsymbol{x}$ son elegidos de manera uniforme, típicamente con cada componente $\hat{\boldsymbol{e}}_i \cdot \Delta\boldsymbol{x}$ generada a partir de una distribución uniforme $\text{U}(a_i, b_i)$. Una descripción en pseudocódigo del algoritmo Metropolis se da a continuación.

Programa 14.4 — Algoritmo Metropolis (pseudocódigo)

```
1: x ← x_0
2: repetir
3:     Proponer un cambio Δx desde una distribución uniforme
4:     x' ← x + Δx
5:     si p(x') > p(x) entonces
6:         se acepta la propuesta
7:     de otra forma si r < p(x')/p(x) con r ∼ U(0,1) entonces
8:         se acepta la propuesta
9:     de otra forma
10:        se rechaza la propuesta
11:    fin si
12:    si la propuesta fue aceptada entonces
13:        x ← x'
14:    fin si
15:    Guardar la muestra x
16: hasta tener suficientes muestras
```

> **Importante:** En cada paso se guarda el estado actual, sea aceptada o rechazada la propuesta. En el caso de una propuesta rechazada, el valor guardado será idéntico al anterior estado guardado.

Recuadro 14.1 — Las programadoras del algoritmo Metropolis
La primera implementación computacional del algoritmo Metropolis, en el computador MANIAC I del laboratorio de Los Alamos, fue programada por Augusta Teller (de soltera Schütz-Harkányi), mientras que la implementación más elaborada fue completamente reescrita por Arianna Rosenbluth (de soltera Wright).

Una implementación abstracta del algoritmo Metropolis se muestra en el Programa (**14.5**).

Programa 14.5 — Algoritmo Metropolis
La función `Metropolis` recibe tres argumentos: el estado semilla `x0`, el número `N` de pasos a simular, y la función `P` que asigna probabilidades a cada estado. También necesita la función `Cambio`, que toma un estado y devuelve otro modificado ligeramente.

```
def Metropolis(x0, N, P):
    x = x0
    p = P(x)
    for paso in range(N):
        newx = Cambio(x)
        newp = P(newx)
        if newp > p: aceptar = True
        elif random() < (newp/p): aceptar = True
        else: aceptar = False
        if aceptar:
           x = newx
           p = newp
        yield x
```

¿Por qué funciona el algoritmo Metropolis?

Recordemos la probabilidad de transición para una cadena de Markov,

$$M_t(\boldsymbol{x} \to \boldsymbol{x}') = P(\boldsymbol{X}_{i+1} = \boldsymbol{x}' | \boldsymbol{X}_i = \boldsymbol{x}, I).$$

En nuestro caso, se deben pasar dos etapas para que ocurra la transición: en primer lugar, la transición de $\boldsymbol{x}$ a $\boldsymbol{x}'$ debe ser **propuesta**, y luego debe también ser **aceptada**. Si llamamos $G(\boldsymbol{x}'; \boldsymbol{x})$ a la probabilidad de proponer un salto hacia $\boldsymbol{x}'$ cuando se está en $\boldsymbol{x}$, se tendrá

$$M_t(\boldsymbol{x} \to \boldsymbol{x}') = G(\boldsymbol{x}'; \boldsymbol{x}) A(\boldsymbol{x} \to \boldsymbol{x}') \tag{14.9}$$

que se reduce a

$$M_t(\boldsymbol{x} \to \boldsymbol{x}') = G_0 \cdot \text{mín}\left(1, \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x})}\right), \tag{14.10}$$

donde $G_0 = G(\boldsymbol{x}'; \boldsymbol{x})$ representa la probabilidad (uniforme) de propuestas. Entonces, la condición de distribución estacionaria para p es, de acuerdo a (13.13),

$$p(\boldsymbol{x}) \sum_{\boldsymbol{x}'} \cancel{G_0} \cdot \text{mín}\left(1, \frac{p(\boldsymbol{x}')}{p(\boldsymbol{x})}\right) = \sum_{\boldsymbol{x}'} \cancel{G_0} \cdot \text{mín}\left(1, \frac{p(\boldsymbol{x})}{p(\boldsymbol{x}')}\right) p(\boldsymbol{x}'). \tag{14.11}$$

Distribuyendo los factores p al interior de la operación $\text{mín}(1, q)$, tenemos

$$\sum_{\boldsymbol{x}'} \text{mín}\left(p(\boldsymbol{x}), p(\boldsymbol{x}')\right) = \sum_{\boldsymbol{x}'} \text{mín}\left(p(\boldsymbol{x}'), p(\boldsymbol{x})\right), \tag{14.12}$$

igualdad que es claramente cierta ya que $\text{mín}(a, b) = \text{mín}(b, a)$ para todo a, b. Luego la tasa de transición de Metropolis asegura que, de existir una distribución estacionaria, ésta debe corresponder a $p(\boldsymbol{x})$.

Una generalización: Metropolis-Hastings

En la variante del algoritmo Metropolis introducida por Hastings (1970), conocida como el algoritmo Metropolis-Hastings, se proponen cambios

$$\Delta x = x' - x$$

a partir de x ya no de manera equiprobable sino que de acuerdo a una distribución de probabilidad

$$P((\Delta X)_i = z | X_i = x, I) := G(z + x; x), \tag{14.13}$$

En este caso, la forma correcta de la probabilidad de aceptación para cumplir con (14.9) es la siguiente.

Definición 14.2 — Tasa de aceptación Metropolis-Hastings

$$A_{MH}(x \to x') := \text{mín}\left(1, \frac{p(x') \cdot G(x; x')}{p(x) \cdot G(x'; x)}\right). \tag{14.14}$$

Es fácil ver por qué, siguiendo la misma idea que en el caso de Metropolis. Escribimos la condición de distribución estacionaria como

$$p(x) \sum_{x'} G(x'; x) \,\text{mín}\left(1, \frac{p(x')G(x; x')}{p(x)G(x'; x)}\right) = \sum_{x'} G(x; x') \,\text{mín}\left(1, \frac{p(x)G(x'; x)}{p(x')G(x; x')}\right) p(x'). \tag{14.15}$$

Distribuyendo los factores G y p al interior de la operación $\text{mín}(1, q)$ se tiene

$$\sum_{x'} \text{mín}\left(p(x)G(x'; x), p(x')G(x; x')\right) = \sum_{x'} \text{mín}\left(p(x')G(x; x'), p(x)G(x'; x)\right), \tag{14.16}$$

igualdad que nuevamente es cierta ya que $\text{mín}(a, b) = \text{mín}(b, a)$ para todo a, b. Por supuesto, el caso con G constante recupera el algoritmo Metropolis original.

14.5 — Algoritmo de Gibbs

Supongamos que se desea muestrear una distribución conjunta $P(X, Y|I)$ donde se conoce $P(X|Y, I)$ y $P(Y|X, I)$ por separado y estas distribuciones condicionales son fáciles de muestrear. Entonces, para la propuesta

$$(x_i, y_i) \to (x_i, y_{i+1}),$$

donde modificamos y manteniendo x, escribimos la probabilidad de aceptación de Metropolis-Hastings como

$$A((x_i, y_i) \to (x_i, y_{i+1})) = \text{mín}\left(1, \frac{P(x_i, y_{i+1}|I)G(x_i, y_i; x_i, y_{i+1})}{P(x_i, y_i|I)G(x_i, y_{i+1}; x_i, y_i)}\right)$$
$$= \text{mín}\left(1, \frac{P(y_{i+1}|x_i, I)G(x_i, y_i; x_i, y_{i+1})}{P(y_i|x_i, I)G(x_i, y_{i+1}; x_i, y_i)}\right), \quad (14.17)$$

donde hemos cancelado $P(x_i|I)$. Aquí podemos ver inmediatamente que, si elegimos $G(x, y'; x, y)$ para cambios en y manteniendo x de acuerdo a

$$G(x, y'; x, y) = P(y'|x, I), \quad (14.18)$$

entonces $A((x_i, y_i) \to (x_i, y_{i+1})) = \text{mín}(1, 1) = 1$, es decir, *tendremos una tasa de aceptación del 100 %*. Esto es una ventaja muy importante del punto de vista de la eficiencia ya que, a diferencia de la implementación usual de Metropolis-Hastings, no se desperdicia ninguna propuesta. De la misma manera, para una propuesta

$$(x_i, y_i) \to (x_{i+1}, y_i),$$

donde modificamos x manteniendo y, nos conviene usar la probabilidad de propuesta $G(x', y; x, y)$ dada por

$$G(x', y; x, y) = P(x'|y, I), \quad (14.19)$$

En resumen, podemos escribir el algoritmo de Gibbs mediante el siguiente pseudocódigo.

Programa 14.6 — Algoritmo de Gibbs (pseudocódigo)

1: $(x, y) \leftarrow (x_0, y_0)$
2: $i \leftarrow 0$
3: **repetir**
4: Elegir y_{i+1} desde la distribución $P(y|x_i, I)$.
5: Elegir x_{i+1} desde la distribución $P(x|y_{i+1}, I)$.
6: $i \leftarrow i + 1$
7: **hasta** tener suficientes muestras

Notemos que como siempre se eligen las variables x e y de distribuciones condicionales de una variable dadas las demás, es perfectamente posible hacerlo mediante el método de aceptación y rechazo sin caer en la *maldición de la dimensionalidad*.

A continuación se muestra, en el Programa (**14.7**), la implementación abstracta del algoritmo de Gibbs. En este código la función `Elegir` toma una distribución como argumento y devuelve un número pseudoaleatorio de acuerdo a esa distribución, y es esta función la que puede ser implementada usando el método de la aceptación y rechazo.

Programa 14.7 — Algoritmo de Gibbs
La función `Gibbs` recibe como argumentos las semillas `x0`, `y0` y las distribuciones condicionales `pxy` y `pyx`, correspondientes a $P(X|Y, I)$ y a $P(Y|X, I)$, respectivamente.

```
def Gibbs(x0, y0, pxy, pyx):
    x = x0
    y = y0
    i = 0
    while i < N:
          y = Elegir(lambda z: pyx(z, x))
          x = Elegir(lambda z: pxy(z, y))
          i = i + 1
          yield (x, y)
```

► Sobre la aplicación de los algoritmos Metropolis, Metropolis-Hastings y Gibbs la referencia obligada es el libro *Bayesian Data Analysis* (Gelman, Carlin, Stern, Dunson, Vehtari y Rubin 2013), además del libro de Gamerman y Lopes (2006).

Bibliografía

Abe, Sumiyoshi (2014). "Conditional maximum-entropy method for selecting prior distributions in Bayesian statistics". *EPL* **108**, pág. 40008.

Arfken, George B. y Hans J. Weber (2005). *Mathematical methods for physicists*. Elsevier Academic Press.

Bailer-Jones, Coryn A. L. (2017). *Practical Bayesian Inference*. Cambridge University Press.

Bayes, Thomas (1763). "An essay towards solving a problem in the doctrine of chances". *Philosophical transactions of the Royal Society of London* 53, págs. 370-418.

Bunge, Mario (2013). *Intuición y Razón*. Penguin Random House Grupo Editorial Argentina.

Caticha, Ariel y Roland Preuss (2004). "Maximum entropy and Bayesian data analysis: Entropic prior distributions". *Physical Review E* **70**, pág. 46127.

Cover, Thomas M. y Joy A. Thomas (2006). *Elements of Information Theory*. John Wiley y Sons.

Cox, Richard T. (1946). "Probability, frequency and reasonable expectation". *Am. J. Phys.* **14**, págs. 1-13.

Davis, Sergio y Gonzalo Gutiérrez (2012). "Conjugate variables in continuous maximum-entropy inference". *Phys. Rev. E* **86**, pág. 051136.

Davis, Sergio y Gonzalo Gutiérrez (2016). "Applications of the divergence theorem in Bayesian inference and MaxEnt". *AIP Conf. Proc.* **1757**, pág. 20002.

Doyle, Arthur Conan (1960). *The Complete Sherlock Holmes*. Doubleday.

Fisher, Ronald A. (1956). *Statistical methods and scientific inference.* Hafner Publishing Co.

Gamerman, Dani y Hedibert F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Taylor y Francis.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari y Donald B. Rubin (2013). *Bayesian Data Analysis*. CRC Press.

Gould, Harvey, Jan Tobochnik y Wolfgang Christian (2007). *An introduction to computer simulation methods*. Pearson Addison Wesley.

Hastings, Wilfred K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* **57**, pág. 97.

Howson, Colin y Peter Urbach (2006). *Scientific reasoning - the Bayesian approach*. Open Court Publishing Company.

Hughes, Barry D. (1995). *Random Walks and Random Environments*. Clarendon Press, Oxford.

Hustad, Kristian G. (2021). URL: https://github.com/hplgit/doconce/blob/master/doc/src/pgf_tikz/fig/maze.tikz.

Irwin, William y Henry Jacoby (2008). *House and Philosophy: Everybody Lies*. Wiley.

Jaynes, Edwin T. (1957). "Information Theory and Statistical Mechanics". *Physical Review* **106**, págs. 620-630.

Jaynes, Edwin T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, Harold (1998). *The Theory of Probability*. Oxford University Press.

Kadane, Joseph B. (2009). "Bayesian thought in early modern detective stories: Monsieur Lecoq, C. Auguste Dupin and Sherlock Holmes". *Statistical Science* **24**, págs. 238-243.

Klafter, Joseph e Igor M. Sokolov (2011). *First steps in random walks: from tools to applications*. Oxford University Press.

Kolmogorov, Andrey (1933). *Foundations of the theory of probability*. Chelsea Publishing Company, NY, USA.

Konishi, Sadanori y Genshiro Kitagawa (2008). *Information Criteria and Statistical Modeling*. Springer.

Landau, David P. y Kurt Binder (2015). *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press.

Laplace, Pierre-Simon de (1820). *Théorie analytique des probabilités*. Vol. 7. Courcier.

MacKay, David J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller y Edward Teller (1953). "Equation of state calculations by fast computing machines". *Journal of Chemical Physics* **21**, pág. 1087.

Popper, Karl (1959). *The logic of scientific discovery*. Routledge.

Press, William H., Saul A. Teukolsky, William T. Vetterling y Brian P. Flannery (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press.

Riley, Ken F., Michael P. Hobson y Stephen J. Bence (2006). *Mathematical methods for Physics and Engineering*. Cambridge University Press.

Risken, Hannes (1996). *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer.

Rosenkrantz, Roger D. (1977). *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*. Springer Netherlands.

Selvin, Steve (1975). "On the Monty Hall problem". *The American Statistician* **29**, pág. 134.

Shannon, Claude (1948). "A mathematical theory of communication". *Bell System Technical Journal* **27**, págs. 379-423.

Sivia, Devinder S. y John Skilling (2006). *Data Analysis: A Bayesian Tutorial*. Oxford University Press.

Stone, James V. (2013). *Bayes' rule: A Tutorial introduction to Bayesian analysis*. Sebtel Press.

van Kampen, Nicolaas G. (2007). *Stochastic Processes in Physics and Chemistry*. North Holland.

von der Linden, Wolfgang, Volker Dose y Udo von Toussaint (2014). *Bayesian Probability Theory: Applications in the Physical Sciences*. Cambridge University Press.

Whittle, Peter (2000). *Probability via expectation*. Springer Science & Business Media.

Wilczek, Frank (2008). *The lightness of being: Mass, ether, and the unification of forces*. Basic Books (AZ).

Zwanzig, Robert (2001). *Nonequilibrium Statistical Mechanics*. Oxford University Press.

Algunas definiciones útiles

DISTRIBUCIÓN DE POISSON

Una variable entera $k \geq 0$ sigue una distribución de Poisson si

$$P(k|\lambda) = \frac{\exp(-\lambda)\lambda^k}{k!}. \tag{1}$$

La notación estándar es $k \sim \text{Pois}(\lambda)$.

DISTRIBUCIÓN GAMMA

Una variable real no negativa $X > 0$ sigue una distribución gamma si

$$P(X = x|k, \theta) = \frac{\exp(-x/\theta)x^{k-1}}{\Gamma(k)\theta^k}. \tag{2}$$

La notación estándar es $X \sim \text{Gamma}(k, \theta)$.

DISTRIBUCIÓN EXPONENCIAL

Una variable real no negativa $X > 0$ sigue una distribución exponencial si

$$P(X = x|\lambda) = \lambda \exp(-\lambda x). \tag{3}$$

La notación estándar es $X \sim \text{Exp}(\lambda)$.

DISTRIBUCIÓN LOGNORMAL

Una variable real no negativa $X > 0$ sigue una distribución lognormal si

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma\, x} \exp\Big(- \frac{(\ln x - \mu)^2}{2\sigma^2} \Big). \tag{4}$$

La notación estándar es $X \sim \text{LogNorm}(\mu, \sigma^2)$.

DISTRIBUCIÓN UNIFORME

Una variable real $X \in [a, b]$ sigue una distribución uniforme si

$$P(X = x|a, b) = \text{rect}(x; a, b). \tag{5}$$

La notación estándar es $X \sim \text{U}(a, b)$.

DISTRIBUCIÓN NORMAL

Una variable real $X \in \mathbb{R}$ sigue una distribución normal si

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big). \tag{6}$$

La notación estándar es $X \sim \mathcal{N}(\mu, \sigma^2)$.

DISTRIBUCIÓN BETA

Una variable real $X \in [0, 1]$ sigue una distribución beta si

$$P(X = x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}. \tag{7}$$

La notación estándar es $X \sim \text{Beta}(\alpha, \beta)$.

DISTRIBUCIÓN BINOMIAL

Una variable entera $0 \leq k \leq n$ sigue una distribución binomial si

$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{8}$$

La notación estándar es $k \sim \text{Bin}(n, p)$.

FUERZA BRUTA

Método de solución de un problema que consiste en barrer las posibles soluciones una a una hasta dar con la correcta.

FUNCIÓN CONVEXA

Una función con derivada monótonamente creciente, es decir, segunda derivada positiva en todos lados.

INTEGRACIÓN POR PARTES

Método de integración que utiliza la identidad

$$\int_a^b \mathrm{d}v\, u = (uv)\Big|_a^b - \int_a^b \mathrm{d}u\, v. \tag{9}$$

INTEGRAL TIPO GAMMA INVERSA

$$\int_0^\infty \mathrm{d}\sigma \exp\Big(- \frac{A}{2\sigma^2} \Big) \sigma^{-n} = 2^{(n-3)/2} \Gamma\Big(\frac{n-1}{2} \Big) A^{\frac{1-n}{2}}. \tag{10}$$

INTEGRAL GAUSSIANA MULTIDIMENSIONAL

Para una matriz invertible $\mathbb{A}$ de $n \times n$ se tiene

$$\int \mathrm{d}\boldsymbol{x} \exp\left(-\frac{1}{2} \boldsymbol{x}^\intercal \mathbb{A} \boldsymbol{x} \right) = \sqrt{\frac{(2\pi)^n}{\det \mathbb{A}}}. \tag{11}$$

INTEGRAL GAUSSIANA

Para todo $A \geq 0$ se tiene

$$\int_{-\infty}^{\infty} \mathrm{d}x \exp\left(-Ax^2\right) = \sqrt{\frac{\pi}{A}}. \tag{12}$$

PUNTO DE ENSILLADURA

Un extremo de una función multidimensional que no es un mínimo ni un máximo, sino que la curvatura depende de la dirección en que uno mire.

RAZONAMIENTO BAYESIANO

Razonamiento que extiende la lógica más allá de verdadero y falso, hacia un continuo de probabilidades entre verdadero y falso.

REGLA DE LEIBNIZ

Para dos funciones $f(x)$ y $g(x)$,

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(f(x)g(x)\big) = g(x)\frac{\mathrm{d}f(x)}{\mathrm{d}x} + f(x)\frac{\mathrm{d}g(x)}{\mathrm{d}x}. \tag{13}$$

REGLA MNEMOTÉCNICA

Regla visual, de lenguaje o simbólica que nos sirve como pista para ayudarnos a recordar una pieza de información.

SEMÁNTICA

Un significado particular asignado a un símbolo o a una palabra.

SERIE GEOMÉTRICA

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z} \quad \text{para } |z| \leq 1. \tag{14}$$

SUMA DE RIEMANN

$$\lim_{n\to\infty} \sum_{i=1}^{n} \Delta_n \cdot f(x_i) = \int_a^b \mathrm{d}x f(x), \tag{15}$$

con $x_i = a + (i-1)\Delta_n$, y $\Delta_n := \dfrac{b-a}{n-1}$.

TEOREMA DE LA DIVERGENCIA

Para una región de integración Ω con borde $\partial\Omega$ y un campo vectorial $\boldsymbol{\omega}(\boldsymbol{x})$ arbitrario se cumple

$$\int_{\Omega} \mathrm{d}\boldsymbol{x}\ \nabla \cdot \boldsymbol{\omega} = \int_{\partial\Omega} \mathrm{d}s\ \boldsymbol{n} \cdot \boldsymbol{\omega}, \tag{16}$$

donde $\boldsymbol{n}$ es el vector unitario normal a $\partial\Omega$.

TEOREMA DEL BINOMIO

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k \, b^{n-k}. \tag{17}$$

Índice alfabético

D

E

F

G

H

I

J

K

L

M