0pt0.4pt

Departamento de Estadística
Pontificia Universidad Católica de Chile

PROBABILIDADES

Versión Preliminar 2002.2

Ricardo Aravena
Guido del Pino
Fernando Quintana

Agosto, 2002

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1. Modelos Probabilísticos y Determinísticos

La *Teoría de Probabilidad* es una rama de las Matemáticas que permite estudiar todo tipo de fenómenos en que aparecen conceptos como indeterminismo, incertidumbre, impredecible, heterogeneidad, variabilidad, errores de medición, imprecisión y azar. En esta sección desarrollamos algunas de estas ideas para motivar el estudio de dicha teoría.

1.1.1. Determinismo y leyes físicas

La imposibilidad práctica de conocer los valores de todas las variables que influyen sobre el comportamiento de un sistema hace que los *modelos determinísticos* tengan un ámbito de aplicación limitado. En estos modelos, el cumplimiento de ciertas condiciones garantiza la ocurrencia de un hecho dado. El paradigma clásico es la Mecánica de Newton, donde se puede predecir exactamente la trayectoria de un objeto, una vez especificadas la posición inicial, la velocidad inicial y todas las fuerzas que actúan sobre él. Desde un punto de vista filosófico, la idea es que *si* tuviéramos toda la información y contáramos con un modelo adecuado, podríamos determinar completamente todos los acontecimientos relacionados. Aún dentro del ámbito de la Física, tal idea está en abierta contradicción con las teorías más modernas, como la Mecánica Cuántica.

Si somos tan afortunados como para disponer de un modelo teórico perfecto que vincula los valores de ciertas variables con los de otras, su aplicación se ve entrabada por la imposibilidad de conocer estos valores con absoluta *precisión*, es decir, cuando hay *errores de medición*. Los *modelos probabilísticos* constituyen una alternativa atractiva a los modelos determinísticos en situaciones de este tipo.

Por otra parte, muchas de las leyes que rigen los fenómenos físicos y químicos han sido descubiertos experimentalmente. Este es el caso de la ley de Boyle: $PV = \kappa T$, que relaciona la presión P, el volumen V, y la temperatura T de un gas. Los errores de medición hacen que las fórmulas matemáticas no se verifiquen de manera exacta con datos experimentales. Cómo ajustar modelos teóricos a datos experimentales o cómo rechazar teorías a partir de estos datos es un problema importante que se ataca utilizando *métodos estadísticos*, para los cuales la Teoría de Probabilidad

sirve de base. Cabe hacer notar, además, que leyes experimentales como la ley de Boyle rigen sólo aproximadamente y para ciertos rangos de valores de las variables.

1.1.2. Probabilidad y leyes científicas

La Teoría de Probabilidad proporciona no sólo un marco conveniente para estudiar el ajuste de modelos matemáticos a datos que contienen errores de medición, sino también una base para desarrollar modelos teóricos en ciertas ciencias. Tal es el caso de las leyes de la Termodinámica, donde se vinculan la presión y la temperatura de un gas con la energía cinética total de un enorme número de moléculas, cuyo movimiento individual es obviamente impredecible. En otras palabras, coexiste una gran incertidumbre a nivel microscópico con una virtual certeza a nivel macroscópico del gas. Algo análogo ocurre con las poblaciones humanas, donde la libertad del individuo es compatible con un comportamiento bastante predecible a nivel agregado. La herramienta teórica que permite fundamentar estas aseveraciones es la *Ley de los Grandes Números*, popularmente conocida como la *Ley de los Promedios*. Ella establece que, bajo ciertas condiciones, se puede predecir con exactitud el valor del promedio, aún cuando los valores individuales sean por completo inciertos. Hemos privilegiado la discusión de ejemplos físicos, donde históricamente los modelos determinísticos han tenido bastante éxito. En las ciencias biológicas y sociales los modelos determinísticos son sólo interpretables en términos de un *comportamiento promedio*. Interesa definir qué significa promedio en este contexto y deducir el comportamiento promedio a partir de supuestos más simples de naturaleza probabilística.

1.1.3. Heterogeneidad y variabilidad

Una dificultad para emplear modelos determinísticos es la presencia de *heterogeneidad* o *variabilidad*. A modo de ejemplo, consideremos las siguientes situaciones: (a) la composición de un lote de mineral varía entre un lote y otro; (b) los tubos catódicos en mil televisores de un fabricante determinado no tendrán exactamente las mismas especificaciones; (c) la vida útil de mil equipos presentará gran variabilidad; (d) el número de automóviles que pasan por una intersección no será el mismo en dos intervalos de cinco minutos tomados aproximadamente a la misma hora.

Las leyes básicas se refieren generalmente a medios homogéneos, por ejemplo, gases y líquidos ideales. La heterogeneidad complica notablemente la formulación matemática y rara vez se dispone de información precisa que permita tomarla en cuenta en el modelo.

1.1.4. Fenómenos caóticos y probabilidad

Los *fenómenos caóticos* son aquellos en que una peque nísima perturbación de las condiciones iniciales de un sistema genera grandes cambios en el estado final del mismo. El matemático Henri Poincaré estudió este tipo de fenómenos a principios de siglo y utilizó el carácter impredecible de estos fenómenos como un modelo físico para la probabilidad. El lanzamiento de una moneda, el lanzamiento de un dado, o el hacer girar la ruleta, son ejemplos familiares en que el resultado se puede interpretar como el estado final de un sistema cuya evolución es caótica.

1.1.5. Vocabulario asociado a la probabilidad

El término *probabilístico*, se usa vagamente como contraposición a *determinístico* y se le asocia implícitamente con palabras como *incierto, impredecible, variable* y *presencia de error*. Hay *incerteza* en la respuesta a preguntas tan diversas como ¿Quién es el culpable de un crimen?, ¿Tendré cáncer?, ¿Acertará el disparo en el blanco?, ¿Será hombre o mujer mi futuro hijo?, ¿Aprobaré el examen?. La *variabilidad* aparece en los lotes de material, en la diversidad genética, en el clima y en las posturas políticas. Como ya se ha mencionado, los *errores de medición* aparecen en variables físicas, pero también en exámenes de laboratorio y en la determinación del nivel socioeconómico.

A menudo la incerteza se refiere a la respuesta correcta a ciertas preguntas, o a la verdad o falsedad de ciertas proposiciones. Cuando la pregunta admite *Sí* o *No* por respuesta, ella tiene asociada la proposición *la respuesta es afirmativa* y el suceso que ocurre si y sólo si la respuesta correcta es positiva. Si la pregunta es ¿Se quemará la ampolleta la próxima vez que se la encienda? el suceso asociado A ocurre cuando la ampolleta se quema. Del mismo modo, el suceso asociado a la pregunta ¿Saldrá un as al lanzar un dado? es B : *Sale un as*. Las probabilidades de estos sucesos se escriben en modo subjuntivo, e.g. $P(A)$ es la probabilidad que se *queme* la ampolleta y $P(B)$ es la probabilidad que *salga* un as.

Muchos modelos determinísticos son causales, por lo que el término *probabilístico* se asocia con ausencia de *causa*. En otras palabras, atribuimos un hecho al *azar*, palabra que el lector habrá encontrado en relación a los *juegos de azar*. En estos juegos, suelen intervenir procedimientos mecánicos de carácter caótico, como lanzar un dado, tirar una moneda al aire, barajar un naipe, hacer girar la ruleta, o elegir una bolita en juegos como el KINO, el LOTO y otros. El resultado de tal juego es claramente incierto, impredecible, y variable.

1.2. Interpretaciones del Concepto de Probabilidad

1.2.1. Interpretación frecuentista

La *interpretación frecuentista o empírica* de la probabilidad se aplica directamente en aquellos casos donde es posible repetir físicamente un experimento muchas veces y bajo condiciones controladas. Cuando para cada repetición del experimento se determina un número real, como el valor de cierta variable cuantitativa, se sabe empíricamente que, bajo ciertas condiciones, los promedios exhiben una gran estabilidad a medida que el número de repeticiones aumenta. Este es un ejemplo de *regularidad estadística* y se conoce popularmente bajo el nombre de *Ley de los promedios* (Ley de los grandes números en teoría de probabilidad). Ya hemos mencionado que esta idea sirve de base para la teoría estadística de la termodinámica. Ella es también clave para los métodos de simulación computacional, conocidos como métodos de Monte Carlo.

Para fijar ideas, consideremos un ejemplo más pedestre, pero sencillo de llevar a cabo, – instamos al lector a hacerlo – que consiste en lanzar repetidamente un dado. Para el i-ésimo lanzamiento anotamos el resultado x_i que muestra el dado – un número entre 1 y 6 – y calculamos secuencial-

mente para $n = 1, 2, 3, \ldots$ el promedio t_n de los primeros n números obtenidos, i.e.

$$t_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

El gráfico de t_n versus n, donde se unen los puntos consecutivos por segmentos rectos, presenta inicialmente gran inestabilidad, pero para valores grandes de n él se asemeja a una curva suave que tiende asintóticamente a una recta horizontal, a una altura aproximada de 3.50.

Una modificación del ejemplo anterior es determinar en cada lanzamiento si ocurre un seis o no. Definiendo $x_i = 1$ en caso positivo y $x_i = 0$ en caso negativo, t_n coincide con la proporción p_n de veces que sale un seis en los primeros n lanzamientos del dado. La Ley de los Grandes Números implica que p_n tiene un valor límite, el cual coincide con la probabilidad que salga seis al lanzar *un* dado. Si el dado es equilibrado el gráfico de p_n versus n presentará una asíntota horizontal, a una altura de aproximadamente 0.167. La interpretación frecuentista *define* esta probabilidad como ese valor límite. Como la Ley de los Grandes Números es un resultado matemático, que depende de ciertos axiomas pero no de una interpretación particular, no queda claro si tenemos derecho a utilizar a priori la existencia del límite sin caer en un argumento circular.

Por el momento denominemos *suceso* a algo cuya ocurrencia o no queda determinada por el resultado de un experimento repetible. Sea Ω el conjunto de *resultados posibles* en cualquiera de estas repeticiones– que se denominará posteriormente *espacio muestral* – y sea $A \subseteq \Omega$ el conjunto de *resultados favorables* al suceso de interés, es decir, aquellos para los que éste tiene lugar. Los elementos del conjunto complementario $\Omega \setminus A$ son, entonces, los resultados desfavorables. La probabilidad del suceso depende exclusivamente del conjunto A y no de la descripción en palabras del suceso de interés. Tiene sentido así denotarla por $P(A)$ y, de hecho, podemos identificar al suceso con A si lo deseamos.

Para entender mejor la interpretación de $P(A)$ e introducir la notación, consideremos un millón de lanzamientos de un dado. Los resultados posibles son 1, 2, 3, 4, 5 y 6. La segunda columna de la siguiente tabla muestra el número de veces que ocurrió cada número (los valores son ficticios para simplificar la aritmética). La tercera columna muestra las proporciones empíricas, las que debieran parecerse bastante a los valores límites, o sea, a las probabilidades correspondientes, dado el elevado número de repeticiones del experimento (lanzamientos del dado). El símbolo $\surd$ en las restantes columnas indica los resultados favorables para diversos sucesos, indicando las dos últimas filas cuantas veces ocurrió cada uno y la proporción respectiva.

ω	$N_n(\omega)$ (en miles)	$p_n(\omega)$	$\omega \geq 5$	ω es par	$\omega \leq 4$	ω entre 3 y 4
1	200	.20			$\surd$	
2	180	.18		$\surd$	$\surd$	
3	170	.17			$\surd$	$\surd$
4	160	.16		$\surd$	$\surd$	$\surd$
5	150	.15	$\surd$			
6	140	.14	$\surd$	$\surd$		
Número (en miles)			290	480	710	330
$P_n(A)$			.29	.48	.71	.33

Denotemos por ω a un resultado posible del experimento y supongamos que éste se repite n veces. Denotemos por $N_n(\omega)$ al número de veces que ocurre el resultado ω, por $N_n(A)$ al número de veces que ocurre el suceso representado por el subconjunto A, y por $p_n(\omega) = \frac{N_n(\omega)}{n}$ y $P_n(A) = \frac{N_n(A)}{n}$ a las proporciones respectivas. Si los límites existen, las probabilidades que ocurra el resultado ω y el suceso representado por A se definen por

$p(\omega) \stackrel{\text{def}}{=} \lim_{n\to\infty} p_n(\omega)$	probabilidad que ocurra ω.
$P(A) \stackrel{\text{def}}{=} \lim_{n\to\infty} P_n(A)$	probabilidad que ocurra A.

La función P que asigna a cada suceso A su probabilidad se denomina *distribución de probabilidad*. La función $p(\cdot)$ se denomina *función de probabilidad* y se expresa normalmente como una tabla o como una fórmula matemática. Cuando los elementos de ω son los valores de una variable aleatoria X, la función P se denomina también distribución de probabilidad de X y se suele denotar por P_X.

Claramente $N_n(A) = \sum_{\omega\in A} N_n(\omega)$. Dividiendo por el número de repeticiones y tomando el límite cuando $n \to \infty$, la definición frecuentista de la probabilidad implica que ella es *no negativa* y satisface, además, las importantes igualdades

$$P(\Omega) = 1. \tag{1.2.1}$$

$$P(A) = \sum_{\omega\in A} p(\omega). \tag{1.2.2}$$

La igualdad (1.2.2) se traduce en la siguiente regla, *válida por ahora sólo para la interpretación frecuentista*:

> *Para un espacio muestral finito la probabilidad que un suceso ocurra es la suma de las probabilidades de los resultados favorables.*

Las propiedades

$$P(\emptyset) = 0, \tag{1.2.3}$$

$$\sum_{\omega\in\Omega} p(\omega) = 1, \tag{1.2.4}$$

pueden obtenerse de la misma forma, pero también se desprenden lógicamente a partir de (1.2.1) y (1.2.2). Por (1.2.2), la función de probabilidad permite calcular las probabilidades de todos los sucesos asociados con el experimento. La afirmación recíproca es trivialmente cierta. Por lo tanto:

> *La función de probabilidad determina la distribución de probabilidad y viceversa.*

Si $\omega_1, \omega_2, \ldots, \omega_n$ es una enumeración de los elementos de Ω, se acostumbra escribir p_i en vez de $p(\omega_i)$ para simplificar la notación, con lo cual (1.2.2) queda

$$P(A) = \sum_{\{i/\omega_i\in A\}} p_i, \tag{1.2.5}$$

$$\sum_{i=1}^{n} p_i = 1. \tag{1.2.6}$$

Un concepto probabilístico clave es el de sucesos *mutuamente excluyentes*, es decir, que la ocurrencia de uno de ellos torna imposible que algún otro ocurra. Para una familia de sucesos esta condición equivale a que *a lo más uno de ellos puede ocurrir.* Un resultado ω no puede ser favorable a múltiples sucesos de esta familia, lo que significa que los conjuntos que los representan son disjuntos. Recíprocamente, si los conjuntos de resultados favorables a dos sucesos son disjuntos, los sucesos son mutuamente excluyentes. Se deduce de (1.2.2) la propiedad aditiva:

$$P(A \cup B) = P(A) + P(B) \text{ si } A \text{ y } B \text{ son disjuntos,} \tag{1.2.7}$$

lo que se generaliza a una unión disjunta de un número finito de sucesos. Cuando convenga, escribiremos una unión de conjuntos disjuntos reemplazando el símbolo $\cup$ por el suma ($+$ o $\sum$), i.e. $A + B$ en vez de $A \cup B$ y $\sum_{i=1}^{n} A_i$ en vez de $\bigcup_{i=1}^{n} A_i$. Con esta convención notacional, la propiedad de aditividad puede escribirse de manera sugerente como

$$P\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i). \tag{1.2.8}$$

1.2.2. Interpretación subjetiva

La interpretación frecuentista es muy especulativa cuando la probabilidad se refiere a situaciones de carácter irrepetible, como conseguir empleo durante el próximo mes, que un familiar se case con su pareja actual, que el candidato XYZ gane las próximas elecciones presidenciales, etc. Cuando la probabilidad se aplica en casos como estos, ella se interpreta más bien como *grado de certeza.* Dado que este último varía entre un sujeto y otro, aunque se refiera a un mismo hecho, esta interpretación se denomina *subjetiva*. Para precisar el valor de esta probabilidad subjetiva para un sujeto determinado, podemos plantearle otros sucesos de probabilidad conocida y establecer comparaciones.

En la vida cotidiana es raro poder contestar preguntas importantes sin algún margen de duda, lo que genera expresiones como *estoy casi seguro que*, *me inclino a pensar que,* y otras por el estilo, de carácter cualitativo. La interpretación de la probabilidad como grado de certeza, expresado en una escala continua entre 0 y 1, representa una cuantificación de las expresiones anteriores.

El enfoque subjetivo de la teoría de probabilidad ha sido el tema de muchas investigaciones matemáticas y filosóficas, sobre las cuales no podemos extendernos. Lo que haremos es traducir los resultados intuitivos de la interpretación frecuentista en un sistema de axiomas, los cuales son aplicables a cualquier interpretación de la probabilidad, incluyendo la subjetiva. Cabe hacer notar, sin embargo, que los axiomas se pueden obtener a partir de consideraciones sobre el comportamiento racional frente a la toma de decisiones con incertidumbre.

De esta manera, suponemos que las propiedades (1.2.1), (1.2.2), (1.2.3) y (1.2.4) siguen siendo aplicables. El espacio muestral se interpreta como un listado de todas las alternativas posibles (escenario en la terminología de los economistas). Por ejemplo, si se quiere apostar en una carrera de caballos, pese a no estar seguro de cuál va a ganar, se debe asignar una probabilidad de ganar a cada uno. Si se piensa en invertir dinero en comprar dolares, para venderlos un mes después, se tiene incerteza sobre el futuro valor del dolar, que puede identificarse con ω. También se puede simplificar el problema y atribuir una probabilidad a que el alza del dolar exceda una cota determinada.

A nivel intuitivo, si la ocurrencia del suceso A implica la del suceso B, debiéramos tener un mayor grado de certeza en B que en A. Identificando los sucesos con conjuntos, la afirmación anterior corresponde a

$$A \subseteq B \Rightarrow P(A) \leq P(B).$$

Esta propiedad de monotonicidad parece un supuesto mínimo cuando lo que interesa es decidir cuál de dos sucesos es más probable. En la interpretación frecuentista las probabilidades se comportan como proporciones y satisfacen un supuesto de aditividad $P(C \cup D) = P(C) + P(D)$. El orden relativo entre dos probabilidades se preserva si aplicamos una transformación g estrictamente creciente definida sobre $[0, 1]$. En otras palabras,

$$P(A) \leq P(B) \Rightarrow g(P(A)) \leq g(P(B)).$$

Si denotamos por p a la probabilidad de un suceso, una transformación útil y que tiene importancia histórica es $g(p) = \frac{p}{1-p}$. En inglés se le llama a $g(p)$ *odds* y no existe una traducción universalmente aceptada. En algunos libros se usa el término *momios*. Aunque sea un anglicismo usaremos la palabra *chances*, dado su uso en apuestas, como carreras de caballos o concursos de belleza. Así, si uno cree algo *3 a 2*, lo que significa es

$$\frac{p}{1-p} = \frac{3}{2},$$

lo que implica $p = \frac{3}{3+2} = 0{,}6$. Una apuesta *1 a 1* corresponde a $p = 0{,}5$. A diferencia de la probabilidad, las chances no son aditivas.

1.2.3. Equiprobabilidad y la regla de Laplace

1.2.3.1. Regla de Laplace

Hay situaciones muy especiales en las que se puede argumentar que todos los resultados posibles son *equiprobables*, es decir, tienen la misma probabilidad. Este argumento se justifica habitualmente apelando a un argumento de simetría y puede interpretarse de manera frecuentista o subjetiva. Por ejemplo, las características geométricas y físicas de una moneda permiten sospechar que el supuesto de equiprobabilidad se cumple aproximadamente. En términos frecuentistas, esto significa que para un número grande de lanzamientos, la proporción de caras sea muy parecida a la de sellos, aunque los números de caras y de sellos sean muy distintos. Desde el punto de vista subjetivo, la equiprobabilidad indica que nos es indiferente apostar a que sale sello o que sale cara, lo que puede sustentarse tanto en consideraciones físicas como en la experiencia empírica previa que tengamos. Por cierto, no podemos esperar que la equiprobabilidad se cumpla exactamente con monedas reales, sino de manera aproximada. Matemáticamente hablando, una *moneda ideal* o *moneda equilibrada* arroja resultados equiprobables por definición. Algo parecido pasa con un dado ideal, donde las 6 caras son equiprobables.

Desde el punto de vista frecuentista la equiprobabilidad significa que la función de probabilidad es constante. Por (1.2.4) su valor es el recíproco de la cardinalidad del espacio muestral y aplicando (1.2.2) se llega a la famosa regla, atribuida a Laplace:

Cuando los resultados posibles son equiprobables, la probabilidad de un suceso es el número de casos favorables dividido por el número de casos posibles,

donde la palabra casos se usa como sinónimo de resultado. En libros antiguos de Algebra, esta regla suele aparecer como definición de probabilidad. Esto es muy peligroso, ya que se puede fácilmente caer en contradicción con la interpretación frecuentista.

1.2.3.2. Algunos modelos equiprobables

Un modelo físico para la generación de resultados equiprobables es el de una urna de N fichas, de las cuales se extrae una al azar. Cada ficha tiene probabilidad $\frac{1}{N}$ de ser elegida. Si se extraen al azar y de manera independiente n fichas de la urna, el resultado es representable por un arreglo o muestra ordenada $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. El elemento y_i es la ficha o cualquier identificador. Sin pérdida de generalidad, podemos enumerar las fichas de la urna de 1 hasta N y usar este número como y_i. Los arreglos de largo n son equiprobables, tanto si el muestreo se hace *sin reposición* (se restituye a la urna la ficha seleccionada) o *con reposición* (cuando se la restituye). El número de tales arreglos es $N(N-1)\times\cdots\times(N-n+1)$ y N^n, respectivamente. Cuando el suceso de interés se refiere sólo al número de fichas de cada color en la muestra, el orden en que aparecen los colores es irrelevante. Sin embargo, *la equiprobabilidad de las muestras no ordenadas sólo ocurre para muestreo sin reposición*. En este caso, cada una de estas muestras corresponde a $n!$ arreglos $\boldsymbol{y}$, por lo que su probabilidad es

$$\frac{n!}{N(N-1)\times\cdots\times(N-n+1)} = \frac{1}{\binom{N}{n}}.$$

Este resultado implica que hay $\binom{N}{n}$ muestras no ordenadas. Como cada una se puede representar por un subconjunto de tama no n de otro de tama no N, $\binom{N}{n}$ es también el número de estos subconjuntos.

Otro modelo común es el de n lanzamientos de un dado equilibrado de N caras. Probabilísticamente, él equivale a una muestra con reposición de n fichas de una urna que contiene a N fichas. Físicamente, el dado se puede lograr para muy pocos valores de n. Una moneda es equivalente a un dado de dos caras.

Ejemplo 1.2.1 Calcular la probabilidad de obtener una suma de k al lanzar dos dados, donde $1 < k < 6$.

Como los 36 pares (x, y) son equiprobables, basta con contar aquellos que son favorables. Pero un tal par satisface $x + y = k$, de modo que toma la forma $(x, k - x)$. Por ejemplo, para $k = 4$, los resultados favorables son $(1, 3), (2, 2)$ y $(3, 1)$ y la probabilidad es $\frac{4}{36} = \frac{1}{9}$. El resultado general es

$$P(\text{Suma} = k) = \frac{k-1}{36}.$$

Ejemplo 1.2.2 Calcular la probabilidad de obtener una suma de 6 al lanzar tres dados. Sea x_i el resultado del i-ésimo dado y sea $\boldsymbol{x} = (x_1, x_2, x_3)$. El espacio muestral consta de $6^3 = 216$ resultados equiprobables. Sea $B = \{\boldsymbol{x}/x_1 + x_2 + x_3 = 6\}$ y sea A_j el

suceso *sale j en el primer dado y la suma es 6*. Si $x_1 = j$, $\boldsymbol{x}$ es favorable si y sólo si $x_2 + x_3 = 6 - j$. Entonces, $B = B_1 + B_2 + B_3 + B_4$. Pero $\text{card}\,(B_i) = 5 - i$, de modo que $\text{card}\,(B) = 4 + 3 + 2 + 1 = 10$. La probabilidad buscada es $P(B) = \frac{10}{216}$.

Finalmente, muchos problemas interesantes involucran permutaciones. Barajar un naipe de k cartas significa elegir al azar uno de los $k!$ órdenes posibles, lo que equivale a elegir al azar una permutación. Probabilísticamente, esto equivale a obtener una muestra ordenada de tama no k de una urna con k fichas.

Ejemplo 1.2.3 Se baraja al azar un naipe de 4 cartas, asignándole a cada una las letras a, b, c y d. A continuación mostramos un listado exhaustivo de las 24 permutaciones de estas letras.

bcde	bced	bdce	bdec	becd	bedc	6
cbde	cbed	cdbe	cdeb	cebd	cedb	6
dbce	dbec	dcbe	dceb	adebc	decb	6
ebcd	ebdc	ecbd	ecdb	edbc	edcb	6

La probabilidad de cualquier suceso se obtiene contando casos favorables y dividiendo por 24.

- Por inspección, hay 6 resultados en que la primera letra es b, de modo que la probabilidad que esto ocurra es $\frac{6}{24} = \frac{1}{4}$. Análogamente, la probabilidad que la segunda letra sea c es también $\frac{1}{4}$. La probabilidad que alguno de estos dos sucesos ocurra no es $\frac{1}{2}$, debido a que estos sucesos no son mutuamente excluyentes. En efecto, abcd y abdc son los dos casos en que ambos sucesos ocurren. Por lo tanto, hay $6 + 6 - 2 = 10$ resultados favorables y la probabilidad buscada es $\frac{10}{24}$.
- La probabilidad que la letra b aparezca antes de la c es $\frac{12}{24} = \frac{1}{2}$, lo que es evidente por simetría.
- La probabilidad que ninguna de las letras caiga en su ubicación natural, i.e. 1 para b, 2 para c, 3 para d y 4 para e, es $\frac{9}{24}$, lo que se obtiene marcando estos casos en el listado y contando cuantos hay.
- Se deja al lector con paciencia repetir esto para las 120 permutaciones de 5 elementos, que aparecen en la siguiente tabla.

abcde	abced	abdce	abdec	abecd	abedc	acbde	acbed	acdbe	acdeb
acebd	acedb	adbce	adbec	adcbe	adceb	adebc	adecb	aebcd	aebdc
aecbd	aecdb	aedbc	aedcb	bacde	baced	badce	badec	baecd	baedc
bcade	bcaed	bcdae	bcdea	bcead	bceda	bdace	bdaec	bdcae	bdcea
bdeac	bdeca	beacd	beadc	becad	becda	bedac	bedca	cabde	cabed
cadbe	cadeb	caebd	caedb	cbade	cbaed	cbdae	cbdea	cbead	cbeda
cdabe	cdaeb	cdbae	cdbea	cdeab	cdeba	ceabd	ceadb	cebad	cebda
cedab	cedba	dabce	dabec	dacbe	daceb	daebc	daecb	dbace	dbaec
dbcae	dbcea	dbeac	dbeca	dcabe	dcaeb	dcbae	dcbea	dceab	dceba
deabc	deacb	debac	debca	decab	decba	eabcd	eabdc	eacbd	eacdb
eadbc	eadcb	ebacd	ebadc	ebcad	ebcda	ebdac	ebdca	ecabd	ecadb
ecbad	ecbda	ecdab	ecdba	edabc	edacb	edbac	edbca	edcab	edcba
12	12	12	12	12	12	12	12	12	12

Contrario a la intuición de la mayoría, el número obtenido es $\frac{44}{120}$, que es levemente inferior a $\frac{9}{24}$.

1.3. Modelo de Probabilidad Finito

1.3.1. Caso general

La función de probabilidad y la distribución de probabilidad ya fueron ya definidas en el contexto frecuentista, obteniéndose la serie de relaciones (1.2.1)–(1.2.8).

Definición 1.3.1 Sea Ω un espacio muestral finito. Sea $p(\cdot)$ una función no negativa con dominio Ω que satisface las condiciones:

$$\sum_{\omega\in\Omega} p(\omega) = 1, \quad p(\omega) \geq 0. \tag{1.3.1}$$

La distribución de probabilidad generada por $p(\cdot)$ es la función $P(\cdot)$ que asigna a todo $A \subset \Omega$ el valor

$$P(A) \quad = \quad \sum_{\omega\in A} p(\omega). \tag{1.3.2}$$

Por (1.3.2), la distribución de probabilidad $P(\cdot)$ determina $p(\cdot)$, pues

$$p(\omega) = P(\{\omega\}), \qquad \omega \in \Omega. \tag{1.3.3}$$

Todas las ecuaciones (1.2.1)–(1.2.8) rigen por definición o como consecuencia lógica. En particular, la probabilidad de un suceso es la suma de las probabilidades de los resultados favorables.

Ejemplo 1.3.1 Por ejemplo, el Teorema del Binomio indica que

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

Por lo tanto, si $p > 0$, $q > 0$, $p + q = 1$, la función

$$p(k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, 1, \ldots, n,$$

se puede usar como la función de probabilidad de cierta distribución. De hecho, ella corresponde a la famosa distribución Binomial, que estudiaremos posteriormente.

Cuando la función de probabilidad se entrega como un listado de resultados y de las probabilidades correspondientes, basta marcar los resultados favorables, por ejemplo, con $\surd$, y sumar las probabilidades respectivas para obtener la probabilidad de un suceso. Cuanto se desea automatizar el procedimiento computacionalmente, conviene generar una columna (o fila) adicional, en que los $\surd$ se reemplazan por 1 y los blancos por 0, la que se interpreta como la función indicatriz del suceso.

Ejemplo 1.3.2 Un equipo tiene dos componentes (a) y (b) e interesa si ellas están operativas o no. Definamos

$x_1 = 1$ si (a) funciona, $x_1 = 0$ en caso contrario.
$x_2 = 1$ si (b) funciona, $x_2 = 0$ en caso contrario.

El estado del equipo está determinado por el par (x_1, x_2), al cual podemos considera como el resultado del experimento. Los resultados se pueden enumerar como indica la tabla. Las probabilidades asignadas en la última columna son positivas y suman 1, de modo que tal asignación es válida.

Resultado	x_1	x_2	Probabilidad
ω_1	1	1	0,6
ω_2	1	0	0,2
ω_3	0	1	0,1
ω_4	0	0	0,1

Consideremos ahora los sucesos:

S_1 : (a) está operativa.
S_2 : (b) está operativa.
S_3 : Exactamente una componente está operativa.
S_4 : Al menos una componente está operativa.

La tabla siguiente muestra cómo representar estos sucesos usando las variables como subconjuntos.

S_1 :	$x_1 = 1$	$A_1 = \{(1,0),(1,1)\}$	$B_1 = \{\omega_1, \omega_2\}$
S_2 :	$x_2 = 1$	$A_2 = \{(0,1),(1,1)\}$	$B_2 = \{\omega_1, \omega_3\}$
S_3 :	$x_1 + x_2 = 1$	$A_3 = \{(0,1),(1,0)\}$	$B_3 = \{\omega_2, \omega_3\}$
S_4 :	$x_1 + x_2 \geq 1$	$A_4 = \{(0,1),(1,0),(1,1)\}$	$B_4 = \{\omega_1, \omega_2, \omega_3\}$

La tabla siguiente muestra cómo representar esta misma información de una manera más cómoda:

Resultado	x_1	x_2	S_1 : $x_1 = 1$	S_2 : $x_2 = 1$	S_3 : $x_1 + x_2 = 1$	S_4 : $x_1 + x_2 \geq 1$
ω_1	1	1	$\surd$	$\surd$		$\surd$
ω_2	1	0	$\surd$		$\surd$	$\surd$
ω_3	0	1		$\surd$	$\surd$	$\surd$
ω_4	0	0				

Reemplazando en la columna correspondiente a S_i el símbolo $\surd$ por 1 y un blanco por 0, se obtiene una nueva columna. Multiplicándola término a término por la columna de probabilidades se obtiene la probabilidad de S_i. La siguiente tabla ilustra este procedimiento.

Resultado										
x_1	x_2	p_i	y_1	$p_i y_1$	y_2	$p_i y_2$	y_3	$p_i y_3$	y_4	$p_i y_4$
1	1	0,6	1	0,6	1	0,6	0	0	1	0,6
1	0	0,2	1	0,2	0	0	1	0,2	1	0,2
0	1	0,1	0	0	1	0,1	1	0,1	1	0,1
0	0	0,1	0	0	0	0	0	0	0	0
		1,0	$P(S_1) =$	0,8	$P(S_2) =$	0,7	$P(S_3) =$	0,3	$P(S_4) =$	0,9

1.3.2. Resultados favorables equiprobables

Cuando *para un suceso dado* todos los resultados *favorables* son equiprobables, es decir,

$$p(x) = p(x') \quad \forall x, x' \in A, \tag{1.3.4}$$

la probabilidad $P(A)$ se puede encontrar mediante la fórmula

$$P(A) = \text{ card } (A) \times p(x_0), \quad \text{donde } x_0 \in A. \tag{1.3.5}$$

En otras palabras, la probabilidad del suceso se obtiene multiplicando la probabilidad de un resultado favorable (cualquiera de ellos da lo mismo) por el número de resultados favorables. El cálculo de $\text{card } A$ requiere habitualmente de las herramientas de teoría combinatoria.

Ejemplo 1.3.3 Se lanzan 5 monedas idénticas, pero no necesariamente equilibradas. Nos interesa la probabilidad de obtener exactamente dos caras entre las cinco monedas. Sea $0 < p < 1$ la probabilidad que una moneda determinada salga cara, y $q = 1 - p$ la probabilidad que salga sello. Sea $y_i = 1$ si la i-ésima moneda es cara, e $y_i = 0$ en caso contrario. Un resultado cualquiera del experimento se puede escribir como $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)$, pero estos 32 resultados no son equiprobables, a menos que $p = 0{,}5$. En un capítulo posterior veremos que bajo un supuesto de independencia entre los lanzamientos se deduce que la función de probabilidad es

$$p^{\sum_{i=1}^{5} y_i} \, q^{5 - \sum_{i=1}^{5} y_i}.$$

Cada resultado favorable tiene probabilidad $p^2(1-p)^3$, de modo que son equiprobables. Para desarrollar nuestra intuición, escribamos dos resultados favorables, por ejemplo, $(1, 0, 0, 1, 0)$ y $(0, 1, 1, 0, 0)$. Cada resultado favorable queda determinado por la posición de los unos (o de los ceros). Como hay 10 maneras de elegir 2 elementos de un conjunto de 5, la probabilidad buscada es $10p^2q^3$.

Ejemplo 1.3.4 Sea $\boldsymbol{x} = (x_1, \cdots x_n) \in \Omega = \{0, 1\}^n$, y sea A_i el suceso $x_i = 1$. Entonces, $S_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ es el número de sucesos A_i que ocurre. Impongamos, además, un supuesto de simetría, que dice que la probabilidad de cada resultado $\boldsymbol{x}$ no depende del orden de sus argumentos. Esto implica que la probabilidad de $\boldsymbol{x}$ depende

sólo del valor s de $S_n(\boldsymbol{x})$, o sea, es igual a $g(s)$ para cierta función g. Bajo este supuesto, todos los casos favorables para el suceso $S_n = s$ son equiprobables. Ellos están en correspondencia uno a uno con los $\binom{n}{s}$ conjuntos $\{i/x_i = 1\}$ de cardinalidad s. Por (1.3.4),

$$P(S_n = s) = \binom{n}{s} g(s).$$

El valor de $g(s)$ se puede calcular tomando cualquier resultado favorable, por ejemplo, una sucesión de s unos seguida de $n - s$ ceros.

Ejemplo 1.3.5 Una situación práctica que queda cubierta por el resultado anterior es el de una población de tama no N, m de cuyos integrantes poseen un atributo dado, por ejemplo, ser mujer, tener un ingreso superior a un monto dado, haber padecido cierta enfermedad, etc. Definiendo $x_i = 1$ si la i-ésima persona en la muestra posee el atributo y $x_i = 0$ en caso contrario, $S_n(\boldsymbol{x}) = s$ es el número de personas en la muestra que poseen el atributo.

Se deja al lector verificar que

$$g(s) = \frac{M^s (N-M)^{n-s}}{N^n} \qquad \text{para muestreo con reposición,}$$

y

$$g(s) = \frac{M^{[s]} (N-M)^{[n-s]}}{N^{[n]}} \qquad \text{para muestreo sin reposición,}$$

donde $a^{[r]} = a \times (a-1) \times \cdots \times (a-r+1)$. De aquí se obtiene

$$P(S_n = s) = \binom{n}{s} \left(\frac{M}{N}\right)^s \left(1 - \frac{M}{N}\right)^{n-s}$$

para muestreo con reposición, y

$$\begin{aligned} P(S_n = s) &= \binom{n}{s} \frac{M^{[s]} (N-M)^{[n-s]}}{N^{[n]}} \\ &= \frac{n! M! (N_M)! (N-n)!}{s!(n-s)!(M-s)!(N-M-n+s)! N!} \\ &= \frac{\binom{M}{s}\binom{N-M}{n-s}}{\binom{N}{n}} \end{aligned}$$

para muestreo sin reposición.

1.3.3. Simulación del caso finito a partir del caso equiprobable

Si sabemos generar N resultados equiprobables, es posible generar resultados aleatorios para cualquier espacio muestral finito, bajo la condición que las probabilidades de los resultados tengan la forma $\frac{r}{N}$. Si las probabilidades están dadas por fracciones, basta elegir N como el máximo común

denominador, o un múltiplo de éste. Si ellas están expresadas de modo decimal, con r cifras, se puede tomar $N = 10^s$ con $s \geq r$.

Consideremos nuevamente la urna y agreguemos un nuevo ingrediente al modelo. Suponemos que existe una conjunto en correspondencia biunívoca con el conjunto de fichas de la urna. Llamamos a este conjunto *población* y a sus elementos *individuos*. Extraer una ficha al azar de la urna equivale a seleccionar un individuo de la población al azar. Supongamos, además, que hay una variable definida para los individuos, como edad, peso, número de cargas familiares, renta, candidato preferido, pasta de dientes favorita, etc. El número de valores distintos está acotado por N pero puede ser muy inferior. Finalmente, establecemos una correspondencia biunívoca entre el conjunto de valores de la variable y un conjunto de colores, que se aplicarán a las fichas. Por ejemplo, si un grupo consta de 60 personas con ingreso alto y 140 personas de ingreso bajo, elegir una persona al azar equivale a extraer al azar una ficha, de una urna con 60 fichas blancas y 140 negras. Denotaremos por m al número de colores.

Un consecuencia inmediata de la equiprobabilidad, que tiene importantes aplicaciones, es:

> *La probabilidad que la ficha extraida sea de un color determinado coincide con la proporción de fichas de ese color en la urna.*

Para demostrar este hecho, introducimos algo de notación. El espacio muestral natural es el conjunto Ω de las N fichas en la urna. Denotemos por t a un color (valor de la variable) y por x al color de *la ficha extraida* (el valor que toma la variable para aquel elemento de la población asignado a la ficha extraida). Sea $\Omega(t)$ el conjunto de fichas de ese color en la urna, y $N(t)$ su número. Si se realizan muchas extracciones con reposición, la proporción de fichas de color x se aproxima a la probabilidad $p(x)$ que el color de la ficha extraida sea x. Pero $p(x)$ es la probabilidad que la ficha seleccionada pertenezca a $\Omega(x)$. Por equiprobabilidad se obtiene

$$p(x) = \frac{\text{card }(\Omega(x))}{\text{card }(\Omega)} = \frac{N(x)}{N},$$

lo que justifica la afirmación anterior.

Ejemplo 1.3.6 Se desea *simular*, a partir de una urna con mil fichas, un dado de 6 caras con probabilidades dadas en la segunda columna de la siguiente tabla:

1	0.3	1–300
2	0.2	301–500
3	0.15	501–650
4	0.10	651–750
5	0.14	751–890
6	0.11	891–1000

Como las probabilidades tienen dos decimales bastaría con 100 fichas, pero 1000 es múltiplo de 100, de modo que lo pedido es factible. Enumerando las fichas de 1 a 1000, podemos tomar $\Omega = \{1, 2, \ldots, 1000\}$ y subdividirlo en 6 conjuntos $\Omega(x)$ de cardinalidad $1000p(x)$, donde $p(x)$ es la probabilidad de la cara con el número x. La tercera columna de la tabla muestra una de las muchas subdivisiones posibles.

1.3.3.1. Extracciones sucesivas de una urna

Cuando sólo interesa el color de las fichas, lo natural es tomar como resultado al arreglo ordenado $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, donde x_i es el color de la ficha obtenida en la i–ésima extracción (no confundir con la i-ésima ficha en la urna). Si m es el número de colores, hay m^n arreglos $\boldsymbol{x}$ cuando el muestreo es con reposición.

El caso de extracciones sucesivas al azar y con reposición nos da un modelo físico concreto para entender la repetición de experimentos en la interpretación frecuentista. La ausencia de asociación entre las distintas extracciones se denomina *independencia* o *independencia estadística* y se tratará en el próximo capítulo. Los lanzamientos repetidos de un dado o una moneda es otro modelo simple de repeticiones independientes de un experimento. Si X_i representa al resultado incierto de la i-ésima extracción, tenemos una sucesión de variables aleatorias independientes, cada una de las cuales tiene a $p(x)$ como función de probabilidad.

Con $m = 2$ y $m = 6$ podemos simular n lanzamientos de una moneda o un dado no equilibrados. En el caso equilibrado basta poner un mismo número de fichas de cada color en la urna (una ficha de cada color basta).

Ejemplo 1.3.7 Una urna contiene 2 fichas blancas y una negra. Se extraen, en forma consecutiva, dos fichas de esta urna. Interesa listar los resultados cuando el muestreo es con o sin reposición. Enumeremos las fichas, de modo que las dos primeras sean blancas y la última sea negra.

Si el muestreo es con reposición, de los cuatro espacios muestrales

$$\Omega_1 = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$$
$$\Omega_2 = \{(b,b),(b,n),(n,b),(n,n)\}$$
$$\Omega_3 = \{11,22,33,12,13,23\}$$
$$\Omega_4 = \{bb,bn,nn\}$$

sólo Ω_1 tiene elementos equiprobables, de modo que la probabilidad de cada resultado es $\frac{1}{9}$. Si el muestreo es sin reposición, de los cuatro espacios muestrales

$$\Omega_5 = \{(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)\}$$
$$\Omega_6 = \{(b,n),(n,b),(b,b)\}$$
$$\Omega_7 = \{12,13,23\}$$
$$\Omega_8 = \{bn,bb\}$$

Ω_5 y Ω_7 tienen elementos equiprobables, con probabilidades $\frac{1}{6}$ y $\frac{1}{3}$, respectivamente. A partir de estos espacios con elementos equiprobables se puede deducir las probabilidades de los resultados en otros espacios. Por ejemplo, a 11 y 12 en Ω_3 se le asocian los subconjuntos $\{(1,1)\}$ y $\{(1,2),(2,1)\}$ en Ω_1, lo que implica que sus probabilidades respectivas son $\frac{1}{9}$ y $\frac{2}{9}$. La equiprobabilidad de los elementos de Ω_7 proviene del hecho que todos ellos están asociados con subconjuntos de cardinalidad 2 en Ω_1.

1.4. Preámbulos para la Formulación Axiomática

1.4.1. Experimentos y resultados

En la interpretación frecuentista se considera un experimento repetible, que tiene ciertos resultados posibles, y se identifica un suceso con el conjunto de resultados para los cuales él ocurre, es decir, los resultados favorables. La incerteza que tenemos sobre un suceso determinado deriva de la incerteza sobre el resultado del experimento. Es importante distinguir entre el resultado obtenido en una *realización del experimento*, que es único, y un resultado potencial. Antes de realizar el experimento, se tiene un conjunto de resultados potenciales y existe incerteza sobre cual será el resultado que se obtenga. Una vez realizado el experimento, el resultado se conoce y la incerteza desaparece.

En la formulación general de la teoría de probabilidad, que incluye la interpretación subjetiva, la palabra *experimento* se utiliza en un sentido muy amplio. Si bien en algunas ocasiones se realiza efectivamente un experimento de laboratorio y se miden los valores de diversas variables, esta es la excepción más bien que la regla. Situaciones tales como elegir al azar una persona de una población y hacerla llenar un cuestionario, o lanzar dados o monedas, o incluso anotar los tiempos de llegada de los automóviles a una intersección durante un cierto período, serían difícilmente denominados experimentos en el lenguaje usual. En situaciones donde las probabilidades son interpretables subjetivamente, como la probabilidad que un empleado recién contratado tenga un buen desempe no en su trabajo, no es fácil visualizar cuál puede ser el experimento correspondiente.

Matemáticamente, el *experimento* es un concepto no definido, es decir, se elude definirlo para evitarse problemas y ampliar el campo de aplicación de la teoría. Para eludir la definición, la estrategia consiste en centrar la atención en la colección Ω de resultados potenciales, a la que se denomina *espacio muestral*, por razones históricas que discutiremos más adelante. Podemos interpretar al experimento como un mecanismo abstracto o caja negra que genera resultados inciertos. Esta incerteza se transfiere a todo suceso cuya ocurrencia dependa del resultado del experimento. El conjunto de resultados favorables representa matemáticamente al suceso. Por analogía con el caso frecuentista, debiéramos esperar que la probabilidad de un suceso coincida con la suma de las probabilidades de los resultados (casos) favorables.

Toda situación admite múltiples descripciones y el resultado de un experimento no es la excepción. Esto implica que el espacio muestral Ω admite diversas especificaciones. El punto de partida de la teoría moderna de la probabilidad, creada por el matemático ruso Kolmogorov en 1933, es considerar a Ω como especificado externamente, es decir, la teoría no indica en absoluto cómo elegirlo. No obstante esto, la especificación de los resultados, y por tanto de Ω, es esencial para la aplicación de modelos probabilísticos a situaciones reales.

La elección de lo que consideraremos resultado debe evitar que dos resultados distintos correspondan al mismo acontecimiento. A su vez, el listado de resultados potenciales debe ser exhaustivo, de modo que se cubran todas las eventualidades. Una manera más sintética de expresar esto es que *exactamente un resultado ocurra en cualquier realización del experimento.*

Ejemplo 1.4.1 En el caso de un dado es posible describir su trayectoria, su posición final sobre la mesa, la cara que queda hacia arriba, el número que está escrito en tal cara, etc. Cualquiera de estas cosas puede considerarse como resultado del experimento.

Ejemplo 1.4.2 Se lanzan dos monedas al aire. Si distinguimos las monedas (por ejemplo pintándolas de distintos colores), es natural distinguir 4 resultados: (cara, cara), (cara, sello), (sello, cara), y (sello, sello)). Si no se distinguen, lo único que sabemos es el número de caras, lo que da tres resultados posibles. Sin embargo, en probabilidad los resultados de un experimento no requieren ser *observables*, lo que contrasta con el uso habitual en los experimentos reales. Más adelante veremos numerosos ejemplos en que los elementos del espacio muestral más conveniente son no observables.

1.4.2. Sucesos y subconjuntos

Dada una familia $\mathcal{A}$ de *sucesos de interés* y un espacio muestral Ω, la idea es identificar a cada suceso $A \in \mathcal{A}$ con el subconjunto de Ω formado por los resultados favorables. La dificultad surge cuando no es claro si cierto resultado $\omega \in \Omega$ es favorable o no, pues el subconjunto queda indefinido. Diremos que Ω está *adaptado* a $\mathcal{A}$ cuando la dificultad mencionada no se presenta para ningún par (ω, A), con $\omega \in \Omega$, $A \in \mathcal{A}$. En otras palabras, Ω está *adaptado* a $\mathcal{A}$ si para cualquier resultado que se produzca, y dado un único suceso de interés A, siempre existe un espacio muestral Ω adaptado a él, o sea, a $\{A\}$. Basta tomar $\Omega = \{\omega_1, \omega_2\}$ e identificar ω_1 con la ocurrencia de A. Automáticamente, el resultado ω_2 indica que A no ocurrió. La elección canónica es $\omega_1 = 1$ y $\omega_2 = 0$, lo que equivale a escribir 1 y 0 para indicar la ocurrencia o no ocurrencia de A, respectivamente.

Ejemplo 1.4.3 Consideremos el lanzamiento de un dado y el suceso de interés B: *Sale un as.* La siguiente es una lista de espacios muestrales propuestos, algunos de los cuales son inadmisibles porque violan los principios básicos enunciados.

(i) $\Omega = \{$ mayor que 2, menor que 3 $\}$ es *Inadmisible:* Si el resultado es *menor que 3*, no podemos asegurar si salió un as o no.

(ii) $\Omega = \{1, 2, 3, 4, 5\}$ es *Inadmisible:* Puede salir un 6, el cual no está en la lista.

(iii) $\Omega = \{$ mayor que 1, menor que 2 $\}$ es *Admisible:* El suceso sale un as corresponde al subconjunto $\{$ menor que 2 $\}$.

(iv) $\Omega = \{$ 1, entre 2 y 5, 6 $\}$ es *Admisible:* Exactamente uno de los resultados debe ocurrir y el suceso de interés corresponde al primero.

(v) $\{1, 2, 3, 4, 5, 6\}$ es *Admisible.*

La elección (v) tiene la ventaja de estar adaptada a cualquier suceso cuya ocurrencia dependa exclusivamente del número que se obtiene al lanzar el dado, e.g. *Sale un número par* o *El número excede 4.*

Si se lanzan dos dados y consideramos como resultados posibles a *Sale un 6 en el primer dado*, *Sale un 6 en el segundo dado* y *Otros casos*, esta asignación es inadmisible ya que si sale un seis en ambos dados, los dos primeros resultados ocurren simultáneamente.

Consideremos una familia de sucesos expresada en términos de proposiciones lógicas. Usando los conectivos lógicos *y,o,* y la negación, se generan muchos otros sucesos. Por ejemplo, si se lanza

un dado y el suceso A_i es que salga un as en el i-ésimo lanzamiento, la siguiente tabla muestra algunos posibles sucesos de interés.

B : sale algún as en los tres primeros lanzamientos	A_1 o A_2 o A_3.
C : no sale un as en el segundo lanzamiento	no ocurre A_2.
D : salen ases en el segundo y cuarto lanzamiento	ocurren A_2 y A_4.
E : salen exactamente dos ases en los primeros tres lanzamientos	muy tedioso de escribir.

Para ciertos propósitos, incluyendo la formulación axiomática de la probabilidad, es conveniente traducir los sucesos originales al lenguaje de conjuntos. Los conectivos lógicos *y,o* y la negación traducen en unión, intersección y complementación respectivamente. La tabla anterior se reescribiría como sigue:

$B : A_1 \cup A_2 \cup A_3$
$C : A_2'$
$D : A_2 \cap A_4$
$E : [(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3)] \setminus (A_1 \cap A_2 \cap A_3)$

Un concepto clave en la teoría de probabilidad es el de sucesos *mutuamente excluyentes*, es decir, que la ocurrencia de uno de ellos torna imposible que algún otro ocurra. Esta condición equivale a *a lo más uno de los sucesos de la familia* puede ocurrir. Cuando *a lo menos uno de los sucesos en la familia* ocurre necesariamente, decimos que la familia es *exhaustiva*. Una familia de sucesos es a la vez mutuamente excluyente y exhaustiva si, en una misma realización del experimento, *exactamente uno de los sucesos de la familia* debe ocurrir. Una manera común de generar familias de este tipo es que cada suceso corresponda a un valor de una o más variables. Mirando a la familia de sucesos como una familia de conjuntos, tenemos la equivalencia mostrada en la siguiente tabla.

Notación probabilística	**Notación de conjuntos**
Mutuamente excluyentes	Disjuntos (intersección vacía)
Exhaustiva	Unión de los conjuntos es todo Ω
Mutuamente excluyente y exhaustiva	Partición de Ω

El modelo probabilístico es un modelo matemático que se representa por la terna $(\Omega, \mathcal{A}, P)$, donde Ω es el espacio muestral, $\mathcal{A}$ es una familia de subconjuntos de Ω (que contiene a la familia de sucesos de interés), y P es una regla que asigna a cada $A \in \mathcal{A}$ un número real entre 0 y 1, a la que llamaremos *distribución de probabilidad*.

1.4.3. Variables

Frecuentemente las preguntas que dan origen a los sucesos de interés se pueden reformular en términos de los valores que toman algunas variables. En tal caso, la ocurrencia del suceso está enteramente determinada por los valores de las variables, y la incerteza sobre ellas se transmite a estos sucesos. A cada variable se le puede asociar una serie de proposiciones o sucesos, por ejemplo, la temperatura supera los 15 grados, la temperatura es inferior a 5 grados, la temperatura está entre 6 y 14 grados, sale un número par al lanzar el dado, gana un candidato particular, etc. Por otra parte, un suceso puede depender de varias variables simultáneamente. Por ejemplo el suceso: *el paciente es obeso* depende del peso, de la talla y de otras variables; el suceso *la suma de los números ob-*

tenidos en tres lanzamientos de un dado es mayor que 14 depende de los valores de tres variables, correspondiendo cada una al número que se obtiene en un lanzamiento determinado.

Cuando hay una única variable de interés, el espacio muestral más natural es simplemente un listado de los posibles valores de esta variable.

Consideremos una *población* finita de *individuos*, cada de los cuales tiene definidos los valores de k variables, a las que denotamos por $X_1, X_2, \ldots, X_k$. Los términos "individuo" y "población" se utilizan para tener una percepción más concreta, pero matemáticamente los individuos de una población son simplemente los elementos de un conjunto arbitrario. Si se enumeran los individuos de la población de 1 a N, todos los valores se pueden organizar como un arreglo rectangular, en que cada fila corresponde a un individuo y cada columna a una variable. Si denotamos por x_{ij} al valor de la variable X_j para el i-ésimo individuo, la i-ésima fila de este arreglo es $(x_{i1}, x_{21}, \ldots, x_{ik})$.

Para ilustrar las ideas, consideramos la Tabla 1.4.1, que muestra las 10 primeras líneas de un archivo computacional de 500 líneas. Cada una de ellas indica la comuna de residencia, el nivel socio-económico (mayor número indica mayor ingreso), el número de integrantes del grupo familiar, el número de consultas médicas efectuadas a lo largo de un a no, el sexo y el peso para el individuo correspondiente.

Identificador	X_1 : Comuna	X_2 : Nivel Socio Económico	X_3 : Tama no Familia	X_4 : N Consultas Médicas	X_5 : Sexo	X_6 : Peso (kg)
1	A	1	3	3	M	74.8
2	A	1	3	2	F	54.2
3	A	1	4	4	M	69.7
4	A	3	4	2	F	58.4
5	C	3	3	8	M	64.6
6	C	4	3	1	F	64.5
7	B	2	3	6	M	72.1
8	A	3	2	2	F	66.0
9	C	3	1	4	M	71.6
10	A	2	2	2	M	72.9

Cuadro 1.4.1: Primeras 10 líneas de un archivo de datos.

Así, el primer individuo es un hombre de 74.8 kg, que vive en la comuna A, de nivel socio-económico bajo. Su familia consta de tres personas y realizó tres visitas al médico el a no pasado. Las variables en nuestro ejemplo ilustran la diversidad que encontramos en la vida real. Ellas se clasifican primariamente de acuerdo al conjunto E de valores posibles, pero también se toma en cuenta las estructuras adicionales definidas sobre E.

Denotemos a la variable por X y por E a su conjunto de valores posibles. Decimos que X es finita si $\text{card}\,(E) < \infty$. Cuando $\text{card}\,(E) = 2$ decimos que la variable es *binaria* o *dicotómica*. Si los valores $x \in E$ son no numéricos, se les denomina *categorías* y se dice que X es *categórica* o *cualitativa*. El sexo, el color, el nivel socio económico, la preferencia por un candidato y la región de residencia son algunos ejemplos. A veces las categorías se codifican como números para efectos computacionales, e.g hombre =1, mujer =2, pero carece de sentido efectuar operaciones aritméticas con estos códigos. Cuando las categorías poseen un orden natural y queremos enfatizar este aspecto, decimos que la variable es *ordinal*. Ejemplos de variables ordinales son el nivel socioeconómico, el

grado de dureza, el grado de acuerdo con una medida gubernamental, etc.

Cuando $E \subseteq \mathbb{R}$, se dice que X es *cuantitativa*. Ellas se denomina *discreta* si E es finito o numerable. Lo más común es que una variable discreta sea un *recuento*, es decir, el número de veces que algo ocurre, en cuyo caso, el conjunto E de valores de la variable está contenido en $\{0, 1, 2, 3, \ldots\}$. Cuando el número total está acotado por n, por ejemplo, si X es el número de transistores defectuosos en un lote de tama no n, $E = \{0, 1, \ldots, n\}$. Un recuento binario tiene sólo valores 1 y 0, que se pueden interpretar como presencia o ausencia de una característica determinada, y se la denomina variable *indicatriz* o *indicadora*. Toda variable binaria se puede recodificar como una variable indicatriz. Por ejemplo, la variable binaria sexo se transforma en indicatriz si le asignamos el código 1 a una mujer y 0 a un hombre. La suma de todos los valores de esta variable indicatriz sobre la población entrega el número total de mujeres y el promedio coincide con la proporción de mujeres en la población.

Cuando no se conoce, a priori, una cota superior para los recuentos es usual tomar $E = \{0, 1, 2, 3, \ldots\}$. El número de hijos de una pareja y el número de llamadas telefónicas efectuadas en un lapso de 5 minutos son dos ejemplos donde se da esta situación.

Como el número de decimales en cualquier medición siempre es finito, una variable numérica X observable es siempre discreta. Sin embargo, cuando tiene sentido imaginar valores intermedios entre cualquier par (x_1, x_2) de valores de X, es útil aceptar la existencia de una variable subyacente Y, que toma valores $y \in [a, b] \subseteq \mathbb{R}$, tal que x se puede interpretar como una buena aproximación de y. Se dice que la variable Y es *continua*. La mayoría de los modelos científicos emplea variables continuas, e.g., edad, peso, estatura, nivel de colesterol, concentración de calcio, temperatura, velocidad y longitud. Habitualmente se ignora la distinción entre la variable subyacente Y y la variable observada X, de modo que se actúa como si X fuese continua.

Ejemplo 1.4.4 Si el experimento consiste en medir una temperatura, el resultado suele describirse por un número real. Sin embargo, podemos hacer las siguientes consideraciones:

- La elección de escala afecta este número (por ejemplo, 0 grados Celsius, 32 grados Farenheit y 273 grados Kelvin corresponden a una misma temperatura).
- Si tomamos en consideración el hecho que el instrumento de medición tiene una precisión finita, el resultado se puede describir más fielmente como un intervalo en $\mathbb{R}$. Por ejemplo, si la precisión es de un decimal, un valor de 36.7 grados corresponde realmente al suceso que la verdadera temperatura está en el intervalo $[36{,}65, 36{,}75)$.
- Se puede considerar un experimento ideal en que el resultado sea la temperatura exacta, pero claramente ella no es observable.

Cuando no se desea imponer una cota superior o inferior a priori, basta tomar $b = \infty$ o $a = -\infty$ respectivamente. Mediante un cambio lineal de variable, o sea, una transformación lineal afín, se reduce el estudio de estas variables a $E = \mathbb{R}$, $E = \mathbb{R}^+$ y $E = [0, 1]$. La clasificación de las variables en la Tabla 1.4.1 es:

X_1 : Comuna	Categórica
X_2 : Nivel Socio-económico	Ordinal
X_3 : Tama no familia:	Recuento
X_4 : Número de consultas médicas	Recuento
X_5 : Sexo	Categórica
X_6 : Peso	Continua

Un comentario final. Cuando la población es finita, se puede concebir una tabla para la población total. Si Ω es el conjunto de todas las filas, y se identifica a la fila $\omega \in \Omega$ con el elemento de la población, una variable asigna un valor a cada ω y, en consecuencia, se puede interpretar como una función definida sobre Ω, que es justamente la definición abstracta del concepto de variable.

Una ventaja del lenguaje de variables es que su uso es mucho más habitual que el de conjuntos. Además, puede que sea claro que los sucesos de interés correspondan a una variable, pero no cuáles son exactamente los sucesos de interés. Por ejemplo, nos puede interesar cuál es el valor de la temperatura, pero no tener claro si el suceso que la temperatura exceda 30 grados es de interés.

1.4.4. Particiones, familias generadas y variables

Con un espacio muestral finito Ω, hay asociadas dos familias especiales de subconjuntos de Ω :

(i) La clase de los *sucesos elementales* $\{\omega\}, \omega \in \Omega$.

(ii) La clase de todos los subconjuntos de Ω.

La primera clase constituye la partición más fina de Ω, mientras que todo suceso en (ii) es una unión disjunta de algunos sucesos elementales. Cuando el resultado puede identificarse con el valor de una variable finita, los sucesos elementales corresponden a la obtención de un valor determinado de la variable, mientras que los sucesos en (ii) son aquellos cuya ocurrencia o no, está determinada por el valor que se obtenga para la variable.

Una familia $(A_1, \ldots, A_k)$ de subconjuntos básicos del espacio muestral Ω induce una partición de Ω que consta de 2^k términos. Cada término es la intersección de k subconjuntos, coincidiendo el i-ésimo subconjunto en esta intersección con A_i o su complemento A_i'. Las uniones finitas de los elementos de la partición inducida constituyen la *familia de sucesos generada por* $A_1, \ldots, A_k$, cuya cardinalidad es 2^{2^k}. De esta forma, dos sucesos inducen una partición del espacio muestral en 4 sucesos y la familia generada consta de 16 sucesos. Para tres sucesos, la partición inducida y la familia generada constan de 8 y 256 sucesos respectivamente. La partición inducida por los sucesos A y B es $(A \cap B, A \cap B', A' \cap B, A' \cap B')$. Para tres o más sucesos resulta tedioso detallar los sucesos que forman la partición inducida por estos sucesos, sin contar con una notación más conveniente.

Con esta motivación, consideramos la variable indicatriz de A_i, que toma el valor $x_i = 1$ si A_i y el valor 0 en caso contrario. El vector binario $\boldsymbol{x} = (x_1, \ldots, x_k)$ determina cuáles sucesos básicos ocurren y cuáles no lo hacen, siendo también verdadera la afirmación recíproca. El conjunto formado por los 2^k arreglos $\boldsymbol{x}$ constituye un espacio muestral alternativo, que denotamos por $\mathcal{X}$. Cada elemento de $\mathcal{X}$ está en correspondencia uno a uno con un suceso de la partición generada por los A_i, al cual denotamos por $E_{\boldsymbol{x}}$ y los subconjuntos de $\mathcal{X}$ están en correspondencia uno a uno con la familia de sucesos generada por los A_i. Con esta notación y tomando $A_1 = A$ y $A_2 = B$, tenemos

$E_{11} = A \cap B$, $E_{10} = A \cap B'$, $E_{01} = A' \cap B$ y $E_{00} = A' \cap B'$. A continuación mostramos como escribir algunos sucesos generados por A y B como uniones de los $E_{\boldsymbol{x}}$ y en términos de condiciones que satisfacen los valores x_1 y x_2.

Ocurre B	$E_{11} \cup E_{10}$:	$x_2 = 1$
Ocurre exactamente uno de los dos sucesos	$E_{10} \cup E_{01}$:	$x_1 + x_2 = 1$
Ocurre al menos uno de los dos sucesos	$E_{11} \cup E_{10} \cup E_{01}$:	$x_1 + x_2 > 0$
No ocurre ninguno de los dos sucesos	E_{00} :	$x_1 + x_2 = 0$
Ocurren ambos sucesos	E_{11} :	$x_1 = 1, x_2 = 1$

Para tres sucesos A_1, A_2, A_3, la ocurrencia de dos o más de ellos corresponde al nuevo suceso $\{(x_1, x_2, x_3)/x_1 + x_2 + x_3 \geq 2\}$ de $\mathcal{X}$, que a su vez corresponde al subconjunto $[(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3)] \setminus (A_1 \cap A_2 \cap A_3)$ de Ω.

Como $\text{card}\ (\mathcal{X})$ tiene 2^k elementos y la suma de sus probabilidades es igual a 1, basta especificar $2^k - 1$ números no negativos cuya suma no exceda 1 para determinar las probabilidades de los 2^{2^k} sucesos generados por $A_1, A_2, \ldots, A_k$. Así, para $k = 3$, 7 números determinan 256 probabilidades, mientras que para $k = 4$, 15 números determinan 65536 probabilidades. Las probabilidades de los sucesos generados pueden también calcularse a partir de aquellas asociadas a $2^k - 1$ sucesos adecuadamente seleccionados. Los sucesos A_i y todas sus intersecciones, de a 2, de a 3, ..., de a k, sirven para este fin, aun cuando esto dista de ser obvio.

1.5. Axiomas

La teoría de probabilidad, considerada como rama de las matemáticas, descansa en una serie de axiomas y de términos que no se definen. Dentro de la teoría, no se hace uso alguno del significado o la interpretación del número real que representa la probabilidad. El cálculo de probabilidades es el conjunto de reglas de operación que permite determinar la probabilidad de ciertos sucesos, a partir de los valores de las probabilidades de otros. Los axiomas son reglas básicas, a partir de las cuales se deducen las reglas de operación.

1.5.1. Aditividad y medida

Las interpretaciones frecuentista y subjetiva son radicalmente diferentes, por lo que es una grata sorpresa que exista una teoría unificada. Esto es posible porque el enfoque matemático consiste en imponer ciertos axiomas y obtener luego conclusiones mediante un razonamiento lógico. La utilidad de este enfoque requiere que no haya contradicción entre los axiomas elegidos y las nociones intuitivas. Con la interpretación frecuentista las probabilidades son proporciones límites, lo que sugiere que las reglas de operación con probabilidades sean análogas a las referentes a operaciones con proporciones.

Una propiedad clave que satisfacen las proporciones es la *aditividad*. Para escribir esto rigurosamente, consideremos una partición finita $(A_1, \ldots, A_k)$ de $A \subseteq \Omega$, donde Ω es un conjunto finito. La aditividad significa que la proporción de elementos de Ω que están contenidos en A es la suma de las proporciones correspondientes a los conjuntos A_i.

Con la notación de (1.2.8), la aditividad de las proporciones se escribe como

$$\text{Prop}\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \text{Prop}(A_i).$$

Muchos conceptos geométricos y físicos, tales como longitud, área, volumen, peso y carga eléctrica, se pueden representar como una función aditiva definida sobre una clase de conjuntos. Por ejemplo, si cortamos un hilo en k pedazos y medimos la longitud de cada uno, la suma de estos números coincide con la longitud original del hilo; si cortamos un pedazo de carne en k pedazos, los pesamos por separado y sumamos los pesos, se recupera el peso original. Si bien los valores de la longitud, el área, el volumen y el peso son todos positivos, ellos pueden ser positivos o negativos en el caso de la carga eléctrica. Un caso semejante es el de una empresa con k sucursales. La ganancia total de la empresa será la suma de las ganancias de cada sucursal (aditividad), pero algunas de estas ganancias podrían ser eventualmente negativas (pérdidas).

En el caso de proporciones no tiene interés considerar particiones infinitas, pero no ocurre lo mismo con los ejemplos geométricos y físicos. Por ejemplo, un círculo no es una unión finita de rectángulos, pero se puede escribir como una unión numerable. Una *medida* es una función m definida sobre una clase de subconjuntos $\mathcal{A}$ de un conjunto Ω, que cumple el axioma de σ-aditividad, también denominada aditividad numerable:

$$m\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} m(A_i). \tag{1.5.1}$$

La longitud, el área, el volumen, el peso, la carga eléctrica, el número de elementos y la proporción, son todos ejemplos de medidas. A nivel de estas notas no enfatizaremos la aditividad numerable.

Decimos que una medida m es *positiva* si $m(A) \geq 0$ para todo $A \in \mathcal{A}$. En el caso del área de una figura (o sea un subconjunto del plano) existen subconjuntos de interés cuya área es infinita. Si todos los subconjuntos de interés están contenidos en una región acotada Ω, el área de Ω es finita y lo propio acontece con todos sus subconjuntos. Cuando la medida m satisface $m(\Omega) < \infty$ se dice que ella es *finita*. Si $m(\Omega) = 1$ se dice que ella es *normalizada*.

La operación de contar está relacionada con una medida positiva, donde a cada subconjunto de un conjunto finito Ω se le asocia su cardinalidad, i.e. el número de elementos que contiene. La aditividad de la cardinalidad es obvia; por ejemplo el número de alumnos de un colegio se puede obtener sumando los tama nos de todos los cursos. Matemáticamente,

$$\text{card}\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \text{card}\,(A_i). \tag{1.5.2}$$

La condición que los subconjuntos A_i sean disjuntos es acá necesaria para evitar contar dos veces el mismo elemento. La proporción $m(A)$ de elementos contenidos en A, dada por

$$m(A) = \frac{\text{card}\,(A)}{\text{card}\,(\Omega)},$$

es también una medida normalizada. En general, toda medida positiva finita se puede normalizar, dividiéndola por la medida de Ω. En el caso de la longitud, el área, el volumen y el peso, la normalización se puede alcanzar con un simple cambio de unidades.

La aditividad no se puede extender a familias no numerables de conjuntos sin generar resultados paradojales. Por ejemplo, el área de un círculo es positiva, mientras que el área de un conjunto con un solo punto es 0. De valer la aditividad en este caso, la suma de muchos ceros sería un número positivo.

1.5.2. Axiomas de probabilidad

Con la formulación conjuntista, la distribución de probabilidad es una función con valores reales, definida sobre una familia $\mathcal{A}$ de conjuntos del espacio muestral Ω. En el caso finito $\mathcal{A}$ está constituida por todos los subconjuntos de Ω. Con estas definiciones,

La probabilidad es una medida positiva y normalizada

Esta afirmación es equivalente a imponer los siguientes **Axiomas de Probabilidad:**

La distribución de probabilidad P es una función definida sobre una clase $\mathcal{A}$ de subconjuntos de Ω que satisface las siguientes condiciones:

$$\text{Aditividad :} \qquad P\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \text{card } (A_i). \tag{1.5.3}$$

$$\text{Positividad :} \qquad P(A) \geq 0, \text{ para todo } A. \tag{1.5.4}$$

$$\text{Normalización :} \qquad P(\Omega) = 1. \tag{1.5.5}$$

$$\text{Aditividad numerable :} \qquad P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{1.5.6}$$

Cuadro 1.5.1: Axiomas de Probabilidad

La aditividad significa que si un suceso se descompone en un número finito de sucesos mutuamente excluyentes, su probabilidad es la suma de las probabilidades de los sucesos en esta descomposición. La aditividad numerable es análoga para particiones infinitas. Tomando $A_i = \phi$, para $i > n$ y demostrando previamente que $P(\phi) = 0$, se deduce que la aditividad numerable garantiza la aditividad, existiendo contraejemplos para la afirmación recíproca.

La σ-aditividad permite calcular probabilidades bajo ciertos procesos límites. Cuando el límite B_∞ de una sucesión de sucesos B_n existe, se requiere la σ-aditividad para garantizar que $P(B_\infty)$ coincide con el límite de las probabilidades $P(B_n)$. Los casos más importantes donde el límite existe corresponden a sucesiones encajonadas de conjuntos, en cuyo caso el límite es la unión de todos los conjuntos para sucesiones crecientes y la intersección de todos ellos en el caso decreciente. Un ejemplo se sucesión creciente es $B_n = \bigcup_{i=1}^{n} A_i$, siendo $B_\infty = \sum_{i=1}^{\infty} A_i$ el límite respectivo.

Ejemplo 1.5.1 Considere lanzamientos sucesivos de una moneda con probabilidad de cara p, con $0 < p < 1$. Sea C_n el suceso *no sale cara en los primeros n lanzamientos*. En el Capítulo 2 se demostrará que $P(C_n) = p^n$, cuyo valor límite es 0. Por otra parte

C_n es una sucesión decreciente y, en consecuencia, $C = \lim C_n = \bigcap_{n=1}^{\infty} C_n$. El suceso C ocurre si nunca sale cara. El axioma de σ-aditividad implica que la probabilidad de este conjunto es 0.

La σ-aditividad es esencial para estudiar problemas donde una variable aleatoria toma valores enteros no negativos, pero hay una cota superior natural. La σ-aditividad garantiza que la suma de las probabilidades de todos los resultados coincide con la probabilidad que ocurra alguno de ellos y por tanto es igual a 1. Por cierto la suma es realmente el valor de una serie.

1.5.3. Propiedades básicas

A partir de los axiomas se puede obtener muchas propiedades útiles. Algunas valen para toda medida, otras para toda medida positiva y otras para toda medida positiva normalizada. El tratamiento axiomático nos entrega una herramienta poderosa para intuir las propiedades probabilísticas básicas. Simplemente usamos un modelo concreto para el cual comprendemos bien alguna medida positiva normalizada, evitando utilizar características muy especiales de esa medida. Por ejemplo, si una figura está contenida dentro de otra, el área de la primera no puede exceder el área de la otra. Esta propiedad intuitiva vale para cualquier medida positiva y se denomina *monotonicidad*. Formalmente, la función de conjunto m es *monótona* si

$$C \subseteq D \Rightarrow m(C) \leq m(D), \tag{1.5.7}$$

- Concepto probabilístico: *La probabilidad de un suceso imposible es nula.*
  Propiedad general: Si algún conjunto tiene medida finita, entonces la medida del conjunto vacío es igual a 0.
  Demostración: Inmediata a partir de $A = A \cup \emptyset \Rightarrow m(A) = m(A) + m(\emptyset)$.

- Concepto probabilístico: *Monotonicidad.* Si la ocurrencia de C implica la de D, entonces $P(C) \leq P(D)$.
  Propiedad general: Toda medida positiva es monótona.
  Demostración: Consideremos la unión disjunta $D = C + (D \setminus A)$. Por aditividad $m(D) = m(C) + m(D \setminus C)$ y por positividad $m(D \setminus C) \geq 0$. Usando un diagrama de Venn e identificando m con el área, es fácil visualizar los pasos de la demostración en términos muy intuitivos.

- Concepto probabilístico: *La probabilidad que ocurra algún suceso es menor o igual a la suma de las probabilidades respectivas.*
  Propiedad general: Para toda medida positiva, la medida de una unión numerable de conjuntos se puede acotar por la suma de sus probabilidades.
  Demostración: Si la desigualdad se cumple para uniones finitas, la σ-aditividad permite pasar al límite. Basta demostrar, entonces, que

$$m\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} m(A_i)$$

para toda medida positiva m. Por inducción se reduce la demostración al caso $k = 2$. Para obtener una idea intuitiva es útil construir diagramas de Venn para $k = 2$ y $k = 3$ y tomar a m como el área.

Considerando un conjunto finito Ω y tomando m como la cardinalidad, la desigualdad se produce al contar algunos elementos más de una vez. Si identificamos los elementos de Ω con nombres de personas y A_i como una lista de algunos de estos nombres, la desigualdad nos dice que el total de nombres puede ser mayor que la suma de los números de cada lista. Por cierto, si no hay nombres repetidos, la desigualdad se transforma en igualdad. La ausencia de repeticiones es equivalente a la intersección vacía de los conjuntos en esta familia de conjuntos.

- Concepto probabilístico: *Fórmula para la probabilidad que ocurra algún conjunto de una familia dada.*
  Problema general: Fórmula para la medida de una unión de conjuntos.

  - Caso $k = 2$:
    $$m(A_1 \cup A_2) = m(A_1) + m(A_2) - m(A_1 \cap A_2)$$
    La demostración es sencilla, siendo lo esencial considerar la partición $(A_1 \setminus A_2) \cup (A_2 \setminus A_1) \cup (A_1 \cap A_2)$. La desigualdad (1.5.13) para $k = 2$ se obtiene como corolario. Notamos, además, que $m(A_1 \cup A_2) = m(A_1) + m(A_2)$ si y sólo si $m(A_1 \cap A_2) = 0$. Usando la analogía con área, si se corta un rectángulo en dos pedazos, el área del borde entre ellos es nula, de modo que se puede incorporar el borde a cada uno de los pedazos sin alterar el área total.
  - Caso general
    $$m(\bigcup_{i=1}^{k} A_i) = \sum_{j=1}^{k} (-1)^{j-1} \alpha_j, \tag{1.5.8}$$
    donde α_j es la suma de la probabilidades de todas las intersecciones de j conjuntos. Por ejemplo, para $k = 3$ se obtiene
    $$\begin{aligned} m(A_1 \cup A_2 \cup A_3)) &= m(A_1) + m(A_2) + m(A_3) \\ &\quad -m(A_1 \cap A_2) - m(A_1 \cap A_3) - m(A_2 \cap A_3) \\ &\quad +m(A_1 \cap A_2 \cap A_3) \end{aligned}$$
    Cuando m es la cardinalidad de un conjunto, (1.5.15) es una identidad combinatorial que se conoce bajo el nombre de principio de unión-exclusión. La razón es que la fórmula se puede interpretar como una manera de descontar repeticiones por exceso, para posteriormente corregirlo, repitiéndose el ciclo varias veces.

Para facilitar las referencias posteriores entregamos una lista de las fórmulas probabilísticas que hemos demostrado en un marco más general.

$$P\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i) \tag{1.5.9}$$

$$P(\phi) = 0, \tag{1.5.10}$$

$$C \subseteq D \Rightarrow P(C) \leq P(D) \tag{1.5.11}$$

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \tag{1.5.12}$$

$$P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i) \tag{1.5.13}$$

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \tag{1.5.14}$$

$$P(\bigcup_{i=1}^{k} A_i) = \sum_{j=1}^{k} (-1)^{j-1} \alpha_j, \tag{1.5.15}$$

donde α_j es la suma de la probabilidades que ocurran j sucesos simultáneamente, al tomar exactamente j sucesos a la vez.

1.5.4. Ejemplos

El primer ejemplo ilustra la idea de descomposiciones aditivas, en conjunto con argumentos de simetría.

Ejemplo 1.5.2 Calcular la probabilidad de obtener exactamente un par al lanzar 3 dados balanceados. La siguiente tabla entrega una lista exhaustiva de resultados.

111	112	113	114	115	116	121	122	123	124	125	126
131	132	133	134	135	136	141	142	143	144	145	146
151	152	153	154	155	156	161	162	163	164	165	166
211	212	213	214	215	216	221	222	223	224	225	226
231	232	233	234	235	236	241	242	243	244	245	246
251	252	253	254	255	256	261	262	263	264	265	266
311	312	313	314	315	316	321	322	323	324	325	326
331	332	333	334	335	336	341	342	343	344	345	346
351	352	353	354	355	356	361	362	363	364	365	366
411	412	413	414	415	416	421	422	423	424	425	426
431	432	433	434	435	436	441	442	443	444	445	446
451	452	453	454	455	456	461	462	463	464	465	466
511	512	513	514	515	516	521	522	523	524	525	526
531	532	533	534	535	536	541	542	543	544	545	546
551	552	553	554	555	556	561	562	563	564	565	566
611	612	613	614	615	616	621	622	623	624	625	626
631	632	633	634	635	636	641	642	643	644	645	646
651	652	653	654	655	656	661	662	663	664	665	666
18	18	18	18	18	18	18	18	18	18	18	18

Si estos resultados son equiprobables – lo que justificaremos en el Capítulo 2 – la probabilidad buscada es $\frac{90}{216} = \frac{5}{12}$. Una forma alternativa de cálculo, que no hace uso de la lista completa, es utilizar las simetrías del problema y la aditividad. Sea A el suceso *sale un par*, A_i el suceso *sale un par de* i y A_{ij} el suceso *sale un par de* i *y un número*

j. Entonces, $A = \sum_{i=1}^{6} A_i$, y por simetría $P(A) = 6P(A_1)$. A su vez $A_1 = \sum_{i=2}^{5} A_{1j}$ y la simetría implica $P(A_1) = 5P(A_{12})$. Así $P(A) = 30P(A_{12})$. En términos del conjunto Ω de los 216 arreglos, el suceso A_{12}, que corresponde a 2 ases y 1 dos, se identifica con el conjunto $\{211, 121, 112\}$, de modo que él corresponde a 3 resultados favorables. Esto muestra que

$$P(A_{12}) = \frac{3}{216}, \quad P(A) = 90 \times \frac{1}{216} = \frac{5}{12}.$$

Ejemplo 1.5.3 (Probabilidad geométrica) La elección de un punto al azar en una región acotada de un plano se obtiene normalizando el área, es decir, la probabilidad de un subconjunto es la razón entre su área y el área total del plano. Por ejemplo, la probabilidad que un punto elegido al azar en un cuadrado caiga dentro del círculo inscrito es $\frac{\pi}{4}$. Si se puede realizar repetidamente este experimento, la proporción de veces que el punto cae dentro del círculo, multiplicada por 4, permite aproximar el valor de π experimentalmente.

Un cálculo similar muestra que si se elige un punto al azar en un disco, la probabilidad que la distancia al origen sea inferior a la mitad del radio es $\frac{1}{4}$. Por otra parte, el disco se puede escribir como

$$\{(r\cos\theta, r\sin\theta) / \ 0 \leq r \leq R, \ 0 \leq \theta < 2\pi\}.$$

La idea de elegir un punto al azar en un diámetro del disco, e independientemente hacer una rotación al azar, se traduce en la elección del par (r, θ) al azar dentro del rectángulo $[0, R] \times [0, 2\pi)$. Bajo este supuesto, la probabilidad que la distancia al origen sea inferior a la mitad del radio es $\frac{1}{2}$. Esto ilustra los peligros de atacar estos problemas de manera puramente intuitiva.

De manera análoga, la probabilidad que un punto al azar en cierto intervalo A cumpla con ciertas condiciones es el cuociente entre la longitud del conjunto de puntos que satisfacen la condición y la longitud de A. Para una región acotada en el espacio, lo propio vale con la longitud reemplazada por el volumen.

Ejemplo 1.5.4 Consideremos un experimento en que el resultado es un número real en el intervalo $[0, M]$. Se nos indica que la probabilidad de un intervalo cualquiera es proporcional al área bajo cierta curva positiva y continua, entre las rectas verticales $x = 0$ y $x = M$. Esta área representa efectivamente una medida acotada sobre los subconjuntos de $[0, M]$, pero no está normalizada. Si la curva es el gráfico de una función continua $h \geq 0$, entonces

$$m([a, b]) = \int_a^b h(x)dx \quad \text{y} \quad P([a, b]) = \frac{\displaystyle\int_a^b h(x)dx}{\displaystyle\int_0^M h(x)dx}.$$

Por ejemplo, si $M = 1$ y $h(x) = x^2(1 - x)$, $\int_0^1 h(x)dx = \frac{1}{12}$ y

$$P([a, b]) = 4b^3 - 3b^4 - 4a^3 + 3a^4.$$

La probabilidad que el resultado sea inferior a un número b se obtiene tomando $a = 0$, i.e. $4b^3 - 3b^4$. Esta es una función estrictamente creciente en b, lo que está de acuerdo con la propiedad monótona de la probabilidad. Ella alcanza el valor 1 para $b = 1$, lo que es simplemente la propiedad de normalización.

Ejemplo 1.5.5 Sea $X = N^\circ$ de artículos defectuosos en un lote de 100 unidades. Los valores posibles de X son $0, 1, 2, \ldots, 100$ y el espacio de probabilidad correspondiente sería $\Omega = \{0, 1, 2, \ldots, 100\}$. Sería catastrófico para la calidad del equipo que los 101 elementos de Ω fueran equiprobables. Se necesitan en principio 100 números para determinar las probabilidades relevantes. Para facilitar los cálculos de probabilidades del tipo $P(X \leq a)$, $P(X > b)$, $P(c < X < d)$, es preferible tabular las probabilidades de los 100 sucesos: $X \leq x$, $x = 0, \ldots, 99$ (obviamente $P(X \leq 100) = 1$). Esta idea es enteramente análoga a las distribuciones acumuladas de proporciones. Denotando $P(X \leq x)$ por $F(x)$ y a $P(X = x)$ por $p(x)$ se obtiene nuevamente un sistema triangular de ecuaciones:

$$\begin{aligned}
F(0) &= p(0) \\
F(1) &= p(0) + p(1) \\
F(2) &= p(0) + p(1) + p(2) \\
&\vdots \\
F(x) &= p(0) + p(1) + \cdots + p(x) \\
&\vdots \\
F(99) &= p(0) + p(1) + \cdots + p(99),
\end{aligned}$$

cuya solución es

$$p(x) = F(x) - F(x-1). \tag{1.5.16}$$

Los sucesos $0 \leq X \leq 2$ y $2 \leq X \leq 3$ no son mutuamente excluyentes pues ambos ocurren cuando $X = 2$. Por lo tanto $P(0 \leq X \leq 2 \text{ ó } 2 \leq X \leq 3) = P(0 \leq X \leq 3)$ $= P(0 \leq X \leq 2) + P(2 \leq X \leq 3) - P(X = 2)$. Además

$$P(a < X \leq b) = F(b) - F(a). \tag{1.5.17}$$

Advertencia: Es importante notar que la primera desigualdad en el lado izquierdo de (1.5.17) es estricta, pero la segunda no lo es. Por ejemplo $P(5 < X < 6)$ no es $F(6) - F(5)$ sino 0, ya que no hay números enteros estrictamente entre 5 y 6. Análogamente, $P(5 \leq X \leq 6)$ no es $F(6) - F(5)$ sino $F(6) - F(4)$. La clave está en reescribir, en caso de necesidad, la desigualdad en forma canónica, es decir con las desigualdades adecuadas. Por ejemplo $3 \leq X < 6 \Leftrightarrow 2 < X \leq 5$, de modo que $P(3 \leq X < 6) =$ $P(2 < X \leq 5) = F(5) - F(2)$. Observe que este procedimiento usa fuertemente el hecho que los valores de X son números enteros.

1.6. Modelo de Probabilidad Numerable

1.6.1. Caso general

Sea Ω numerable y sea P una distribución de probabilidad dada. Se define la función de probabilidad por $p(\omega) = P(\{\omega\})$, $\omega \in \Omega$.

Los sucesos básicos $\{\omega\}$ constituyen una partición numerable de Ω y todo $A \subset \Omega$ es una unión numerable de los sucesos básicos $\{\omega,\ \omega \in A\}$. Por σ-aditividad,

$$P(A) \quad = \quad \sum_{\omega \in A} p(\{\omega\}). \tag{1.6.1}$$

Esto indica que la probabilidad de un suceso sigue siendo la suma de las probabilidades de los resultados favorables.

Aplicando (1.6.1), con $A = \Omega$, se tiene

$$\sum_{\omega \in \Omega} p(\omega) = 1, \quad p(\omega) \geq 0. \tag{1.6.2}$$

Todas las ecuaciones (1.2.1) – (1.2.8) rigen por definición o como consecuencia lógica.

Si se enumeran los términos de A, la suma en (1.6.1) es el valor de una serie. La no negatividad de los términos garantiza que este valor no depende de la enumeración elegida. Además, (1.6.2) garantiza la convergencia. El caso finito sale como corolario, donde no se requiere la σ-aditividad, sino la aditividad finita.

1.6.2. Enteros no negativos

Si el resultado del experimento es un número entero no negativo k, para el cual no queremos imponer una cota superior, lo habitual es elegir Ω como el conjunto de enteros no negativos $\{0, 1, 2, \ldots\}$. Esta situación ocurre frecuentemente cuando la variable es un recuento, e.g. número de accidentes, de llamadas telefónicas, de llegadas a una intersección, de clientes en una cola, etc. Las igualdades (1.6.2) y (1.6.1) se transforman en

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

y

$$P(A) = \sum_{k \in A} p(k),$$

respectivamente. Para verificar si

$$p(k) = 0{,}7 \times 0{,}3^{k-1}, \ \ k > 0, \ \ p(0) = 0$$

define una función de probabilidad válida, basta verificar que los valores son no negativos y calcular la suma. Si el experimento consiste en extraer artículos de un lote hasta que aparezca el primer

defectuoso y k es el número total de artículos que se extrae, $P(X > 3)$ coincide con la probabilidad que las primeras tres extracciones entreguen artículos no defectuosos. Un cálculo directo da

$$\begin{aligned} P(X \geq 3) &= \sum_{k=4}^{\infty} p(k) && (1.6.3)\\ &= 0{,}7 \times 0{,}3^3 \sum_{k=4}^{\infty} 0{,}3^{k-4} && (1.6.4)\\ &= 0{,}7 \times 0{,}3^3 \sum_{j=0}^{\infty} 0{,}3^{j} && (1.6.5)\\ &= 0{,}3^3 && (1.6.6)\\ &= 0{,}027 && (1.6.7) \end{aligned}$$

Si la serie converge pero la suma es $c \neq 1$, basta normalizar la función dividiendo cada término de la serie por c. Basta, entonces, indicar el valor de la función de probabilidad salvo por una constante de proporcionalidad y determinarla usando (1.6.2). Por ejemplo, de la serie de Taylor de la función exponencial se deduce que $c = \sum_{k=0}^{\infty} \frac{1}{k!}\lambda^k = e^{\lambda}$, de modo que

$$p(k) = \frac{e^{-\lambda}}{k!}\lambda^k, \quad \lambda > 0, \quad k \geq 0$$

es una legítima función de probabilidad. La distribución de probabilidad correspondiente se denomina distribución de Poisson y está determinada por el parámetro ajustable λ.

1.6.3. Familias paramétricas y series de potencia

Es muy excepcional conocer los valores exactos de la función de probabilidad. Lo habitual es que exista información empírica previa sobre las frecuencias relativas de los distintos valores posibles de una variable aleatoria. Para que el modelo probabilístico tenga relevancia práctica se procura elegir *la forma* de función de probabilidad p de tal modo que se asemeje a la función de probabilidad empírica (donde las proporciones empíricas reemplazan a las probabilidades). El uso de *familias paramétricas*, como la de Poisson, permite ajustar la función de probabilidad a los datos mediante la elección de uno o más números reales, que se denominan *parámetros.*

Muchas familias paramétricas se pueden deducir a partir de series de potencia conocidas. Sea

$$G(z) = \sum_{k=0}^{\infty} c_k z^k, \ |z| < r, \tag{1.6.8}$$

una serie de potencias con radio de convergencia r. El caso especial, en que todos los coeficientes c_k son nulos excepto un número finito de ellos, da origen a un polinomio, para el cual $r = \infty$.

Si $c_k \geq 0$ para todo k, la función $p(\cdot, \theta)$ definida por

$$p(k, \theta) = \frac{c_k \theta^k}{G(\theta)}, \ \theta < r, \ k \geq 0, \tag{1.6.9}$$

es una función probabilidad válida para $0 \leq \theta < r$.

Ejemplo 1.6.1 Una aplicación de (1.6.8) y (1.6.9) a las conocidas expansiones

$$\begin{aligned} e^z &= \sum_{k=0}^{\infty} \frac{1}{k!} z^k, \ |z| < \infty, \\ z(1-z)^{-1} &= \sum_{k=1}^{\infty} z^k, \ |z| < 1 \end{aligned}$$

conduce a las funciones de probabilidad

$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \ \ \lambda > 0.$$

y

$$p(k, \theta) = (1-\theta)\theta^{k-1}, \ \ k > 0, \ 0 < \theta < 1.$$

La primera genera la distribución de Poisson, y la segunda recibe el nombre de geométrica. Para $\theta = 0{,}3$ se obtiene la función de probabilidad en el ejemplo de los artículos defectuosos.

Ejemplo 1.6.2 La función $G(z) = (1+z)^n$ es un polinomio y el coeficiente de z^k es $c_k = \binom{n}{k}$, que es no negativo. Por lo tanto,

$$\begin{aligned} p(k, \theta) &= \binom{n}{k} \frac{\theta^k}{G(\theta)} \\ &= \binom{n}{k} \frac{\theta^k}{(1+\theta)^n}, \ \theta \geq 0, \end{aligned}$$

define una familia paramétrica. Si $\alpha = \frac{\theta}{1+\theta}$, ésta se puede reescribir como

$$\binom{n}{k} \alpha^k (1-\alpha)^{n-k}, \ 0 \leq \alpha \leq 1.$$

1.7. Problemas

1. Sean tres sucesos E, F y G. Encuentre expresiones para los siguientes sucesos en lenguaje de conjuntos.

   a.- Sólo ocurre E.

   b.- Ocurren tanto E como G, pero no así F.

   c.- Al menos uno de los sucesos ocurre.

   d.- Al menos dos de los sucesos ocurren.

   e.- Los tres sucesos ocurren.

   f.- Ninguno de los tres sucesos ocurre.

   g.- A lo más uno de ellos ocurre.

   h.- A lo más dos de ellos ocurren.

   i.- Exactamente dos de ellos ocurren.

2. Pruebe la desigualdad de Boole:

$$P(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$$

3. Demuestre que la probabilidad que ocurra exactamente uno de los sucesos E o F es igual a:

$$P(E) + P(F) - 2P(E \cap F).$$

4. Demostrar que si P y Q son dos medidas de probabilidad definidas sobre un mismo espacio, entonces $aP + bQ$ es también una medida de probabilidad para algunos números no negativos a y b tales que $a + b = 1$. Dar un ejemplo encontrando valores a y b que cumplan las condiciones.

5. Una caja contiene una ficha roja, una verde y una azul. Considere el siguiente experimento: se saca una ficha de la caja, ésta es devuelta y se extrae una segunda ficha. Describir un espacio muestral apropiado. Repetir lo anterior si la ficha se extrae sin reposición.

6. Se lanza un dado hasta que aparece un seis. ¿Cuál es el espacio muestral de este experimento?. Si E_n denota el suceso que son necesarios n lanzamientos para completar el experimento, ¿qué elementos del espacio muestral están contenidos en E_n?. ¿Qué es $(\bigcup_{i=1}^{\infty} E_n)^c$?

7. Formular un modelo matemático para los siguientes experimentos, describiendo el espacio muestral e indicando las probabilidades asociadas a cada uno de sus elementos.

   a.- Se lanza cinco veces una moneda.

   b.- Se lanza un dado cinco veces.

   c.- Se lanza cinco veces un dado cuyas caras están marcadas 1, 1, 2, 2, 3, 4.

8. En una tienda existen tres camisas de distinto tipo para la venta.

   a.- Si dos hombres compran una camisa cada uno, ¿cuántas posibilidades de compra hay?

   b.- Si dos camisas son vendidas, ¿cuántas posibilidades de venta hay?

9. Se seleccionan dos cartas al azar en un juego de naipes. ¿Cuál es la probabilidad que una de ella sea un as y la otra no esté entre 1 y 7?

10. Cinco fichas son aleatoriamente distribuidas en tres cajas A, B y C. Evaluar la probabilidad de los siguientes sucesos:

    a.- La caja A está vacía.

    b.- Sólo la caja A está vacía.

    c.- Exactamente una caja está vacía.

    d.- Al menos una caja está vacía.

    e.- No hay cajas vacías.

    f.- Dos cajas están vacías.

    g.- La caja A o la caja B están vacías.

11. Repetir el ejercicio anterior con n fichas y tres cajas. Verificar la expresión general para el ejercicio anterior.

12. Se ordena un grupo de 30 personas al azar y se les va preguntando de uno a uno el día de su nacimiento. Calcule la probabilidad que no haya dos personas con el mismo cumplea nos entre las primeras (i) 10 (ii) 20 personas.

13. Suponga que de un mazo de n cartas marcadas de 1 a n, se extraen cartas aleatoriamente y éstas van siendo ordenadas según el orden de extracción. Sea A el suceso que la carta 1 aparezca en la primera posición y sea B el suceso que la carta 2 aparezca en la segunda posición.

    a.- Demuestre que $P(A) = P(B) = 1/n$.

    b.- Demuestre que $P(A \cap B) = 1/n(n-1)$.

    c.- Demuestre que $P(A \cup B) = (2n-3)/n(n-1)$

14. Suponga que 4 tarjetas marcadas $1, 2, 3, 4$ se mezclan y luego se colocan al azar en 4 posiciones fijas. Sea X el número de coincidencias, i.e, el número de veces que una tarjeta marcada i queda en la posición i. Demuestre por enumeración directa de los 24 resultados posibles que $P(X = x)$ es la siguiente:

k	0	1	2	3	4
$P(X=k)$	$\frac{9}{24}$	$\frac{8}{24}$	$\frac{6}{24}$	0	$\frac{1}{24}$

15. Si un número de 3 dígitos (000 a 999) es elegido al azar, encontrar la probabilidad que exactamente un dígito sea mayor que 5.

    $Resp$: 0,432

16. Suponga que h hombres y m mujeres se sientan aleatoriamente en $h+m$ asientos puestos en fila. Encontrar la probabilidad que todas las mujeres queden juntas.

    $Resp: \frac{h+1}{\binom{h+m}{h}}$.

17. Un experimento consiste en sacar diez cartas al azar de un naipe de 52 cartas.

    a.- Si la extracción se hace con reemplazo, encontrar la probabilidad que no hayan dos cartas con el mismo valor numérico.

    b.- Si la extracción se hace sin reemplazo, encontrar la probabilidad que al menos nueve cartas sean de la misma pinta.

    $Resp: a) \frac{52*48*44*\cdots*18}{(52)^{10}} \quad b) \frac{4*\binom{13}{9}*39+4*\binom{13}{10}}{\binom{52}{10}}$.

18. Una caja contiene $2n$ helados, n de naranja, y n de limón. De un grupo de $2n$ personas, $a < n$ prefieren naranja, y $b < n$ prefieren limón, mientras que las restantes $2n-a-b$ personas no tienen preferencias. Demuestre que si los $2n$ helados se reparten al azar, la probabilidad que todas las preferencias sean respetadas es

$$\frac{\binom{2n-a-b}{n-a}}{\binom{2n}{n}}.$$

Desafíos

19. Se lanza un par de dados hasta que la suma de ellos sea cinco o siete. Encuentre la probabilidad que la suma cinco aparezca primero.

    $Hint$: Sea E_n la suma cinco aparece en el n-ésimo lanzamiento y cinco o siete no aparece en el lanzamiento $n-1$. Calcule $P(E_n)$ y argumente que $\sum_{n=1}^{\infty} P(E_n)$ es la probabilidad deseada.

20. (Problema de Banach) El matemático Banach mantenía dos cajas de fósforos, una en cada bolsillo y cada caja contenía n fósforos. Cada vez que él necesitaba un fósforo, seleccionaba aleatoriamente uno de los bolsillos. Cuando él encontró que la caja seleccionada estaba vacía, ¿cuál es la distribución del número de fósforos que quedaban en la otra caja ?

    $Hint$: Divida en dos casos de acuerdo a que el bolsillo derecho e izquierdo esté vacío, pero tenga cuidado con el caso en que ambos estén vacíos.

21. Generalice el Problema 14 al caso de n tarjetas.

22. Una urna contiene n tarjetas enumeradas de 1 a n. Se sacan al azar las tarjetas una por una y sin reemplazo. Si la tarjeta con el número r aparece en la r-ésima extracción, entonces diremos que ocurrió un encuentro. Probar que la probabilidad que al menos un encuentro ocurra es:

$$1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + \frac{(-1)^{n-1}}{n!} \to 1 - e^{-1}$$

    cuando $n \to \infty$.

    $Hint$: Usar el Problema 13 y $P(A_1 \cup A_2 \cup \cdots \cup A_n)$.

Capítulo 2

Probabilidad Condicional e Independencia

2.1. Probabilidad Condicional e Información

2.1.1. Introducción

Analicemos las dos situaciones siguientes.

- La probabilidad de obtener dos caras al lanzar dos veces una moneda equilibrada es $\frac{1}{4}$. Sin embargo, si alguien nos comunica que la primera moneda salió cara, la probabilidad relevante es intuitivamente mayor. Dado que sólo existe incertidumbre sobre el segundo lanzamiento, el valor $\frac{1}{2}$ parece reflejar mejor la situación.
- La probabilidad de obtener al menos un as al lanzar dos dados es $\frac{11}{36}$, pues los 36 pares (x_1, x_2) son equiprobables. Si alguien nos informa que la suma de los dos dados es 5, podemos desechar casi todos los resultados y quedarnos sólo con $\{(x_1, x_2)/\ x_1 + x_2 = 5\} = \{(1,4),(2,3),(3,2),(4,1)\}$. De estos 4 resultados hay 2 favorables, por lo que resulta tentador usar la fracción $\frac{2}{4} = \frac{1}{2}$ para representar nuestra incertidumbre. Cabe notar que la equiprobabilidad implícita no es evidente.
- La probabilidad que una persona elegida al azar de una población posea cierta característica genética coincide con la proporción α de personas de la población que la poseen. Si la persona se hace un test, que tiene un margen de error, y éste resulta positivo, interesa actualizar esta probabilidad α para tomar en cuenta esta información.

El argumento implícito en los primeros dos casos es que resultados que eran equiprobables siguen siéndolo luego de conocida cierta información. En estos tres ejemplos quedan de manifiesto los siguientes hechos:

1. La información afecta la probabilidad.
2. La información se traduce en que cierto suceso F ocurre.

3. Si sabemos que F ocurre, la ocurrencia de A implica la de $A \cap F$.

La probabilidad buscada en los tres ejemplos se puede traducir en la probabilidad que ocurra A dado que ha ocurrido F. Se la denotará por $P(A|F)$ y se leerá *probabilidad condicional de A dado F*.

2.1.2. Interpretación frecuentista

Identificando a A y F con los respectivos conjuntos de resultados favorables en un experimento dado, y suponiendo él es repetible muchas veces, podemos interpretar las probabilidades como proporciones. Sea $N = 10^6$ el número de repeticiones y supongamos que F se cumple en 400000 de ellas, mientras que A y F ocurrieron conjuntamente 300000 veces. Es claro que F ocurrió un 75 % de aquellas repeticiones en que el suceso A se cumplió. Esta fracción parece reflejar mejor la incerteza sobre A cuando se sabe que F ocurrió, que simplemente la probabilidad que ocurra A.

2.1.3. Caso equiprobable

Si los n puntos de Ω son equiprobables y sólo sabemos que F ocurre, parece natural usar F como nuevo espacio muestral y suponer que sus m puntos siguen siendo equiprobables. Notemos que

$$\frac{1}{m} = \frac{\frac{1}{n}}{\frac{m}{n}}.$$

2.2. Definición Formal de Probabilidad Condicional

Las interpretaciones discutidas en la sección anterior sugieren cómo definir la probabilidad condicional usando un enfoque axiomático.

Definición 2.2.1 Si $P(F) > 0$, la probabilidad condicional de A dado F, que se denota $P(A|F)$, está dada por

$$P(A|F) = \frac{P(A \cap F)}{P(F)}, \tag{2.2.1}$$

que es equivalente a la *regla multiplicativa*

$$P(A \cap F) = P(F)P(A|F). \tag{2.2.2}$$

En la práctica es más frecuente tener una idea de los valores de $P(A|F)$ y $P(F)$, por lo que la versión multiplicativa es la más útil.

Nota: Si alguien nos informa de un suceso, cuya ocurrencia era absolutamente segura, ello no debiera cambiar nuestras probabilidades. En otras palabras, $P(A|F)$ debiera coincidir con $P(A)$. Esto se desprende inmediatamente de la definición axiomática, tomando $F = \Omega$, bastando la condición $P(F) = 1$.

La utilidad de una definición formal se muestra en el próximo ejemplo.

Ejemplo 2.2.1 Se dispone de tres cartas: (1) con ambas caras blancas, (2) con ambas negras y (3) con una cara de cada color. Se elige una carta al azar y luego se pone sobre una mesa, eligiendo al azar una de sus caras. Si la cara mostrada es negra, calcule la probabilidad que la otra sea negra.

La intuición indica que hay sólo dos cartas posibles, y por la simetría del problema ambas son equiprobables, de modo que la probabilidad buscada es $\frac{1}{2}$. Si el lector tiene la paciencia de repetir muchas veces el experimento, se dará cuenta que del conjunto de repeticiones en que la cara mostrada es negra, mucho más de la mitad tiene la otra cara negra. Esto muestra que la intuición no siempre funciona.

Para analizar formalmente el problema, marquemos cada cara de las cartas (con tinta invisible), con las letras a y b. Hay entonces 6 resultados $1a, 1b, 2a, 2b, 3a, 3b$, que debieran ser equiprobables, por simetría. La información que la cara visible es negra se traduce en un suceso $B = \{2a, 2b, 3b\}$. La probabilidad que la carta visible sea negra y que la otra también lo sea es igual a la probabilidad del suceso $\{2a, 2b\}$. Por definición de probabilidad condicional la probabilidad buscada es

$$\frac{P(\{2a, 2b\})}{P(\{2a, 2b, 3b\})} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}.$$

Sean ahora X e Y las variables indicatrices de los sucesos A y F respectivamente. La regla multiplicativa (2.2.2) equivale a $P(X = 1, Y = 1) = P(X = 1)P(Y = 1|X = 1)$. La idea general es que el valor de una variable finita X afecta las probabilidades relativas a otra variable Y, lo que podemos escribir como $P(Y = y|X = x)$. En nuestro caso $P(Y = 1|X = 0) = P(A|F')$, $P(Y = 0|X = 1) = P(A'|F)$, y $P(Y = 0|X = 0) = P(A'|F')$. Intuitivamente, $P(A)$ debiera ser un valor intermedio entre $P(A|F)$ y $P(A|F')$, o sea, algún promedio ponderado de estos valores. Si estamos bastante seguros que F debe ocurrir, $P(A)$ debiera estar más cerca de $P(A|F)$ que de $P(A|F')$, por lo que debiera aparecer con más peso en el promedio ponderado. La descomposición aditiva $A = (A \cap F) + (A \cap F')$ implica $P(A) = P(A \cap F) + P(A \cap F')$. Aplicando la regla multiplicativa a los pares (A, F) y (A, F) permite obtener la fórmula exacta

$$P(A) = P(F)P(A|F) + P(F')P(A|F'). \tag{2.2.3}$$

En otras palabras, la probabilidad *marginal* (no condicional) de A es un promedio ponderado de las probabilidades condicionales. Si $P(F) = 0$, $P(A|F)$ no está definido pero le podemos asignar cualquier número entre 0 y 1, de modo que $P(F)P(A|F) = 0$. Algo análogo ocurre cuando $P(F') = 0$. Con esta convención, (2.2.3) vale sin restricciones. En términos de las variables indicatrices, ella se reescribe como $P(Y = 1) = P(X = 1)P(Y = 1|X = 1) + P(X = 0)P(Y = 1|X = 1)$. Usando los mismos argumentos, se demuestra que

$$P(Y = y) = P(X = 1)P(Y = y|X = 1) + P(X = 0)P(Y = 1|X = 0).$$

2.3. Independencia de dos sucesos

El concepto de independencia está intuitivamente asociado con *ausencia de efecto o de interacción*. Desde el punto de vista probabilístico, nos interesa expresar la idea que la ocurrencia o no de

un suceso no afecte la probabilidad que otro ocurra. En el lenguaje de variables, la idea es que el valor que toma una variable no afecte las probabilidades de los valores de otra variable. Esta idea aparece implícitamente en la interpretación frecuentista, pues se supone que lo que ocurra en una de las repeticiones del experimento no afecta a las otras. Ya hemos mencionado que lanzamientos sucesivos de una moneda o un dado, así como el muestreo con reposición, parecen cumplir con esta ausencia de interacción. El problema es cómo dar una definición formal de este concepto, dentro del marco axiomático.

Si queremos expresar que la ocurrencia o no de un suceso F no afecta la probabilidad que otro suceso A ocurra, parece natural imponer la condición

$$P(A|F) = P(A|F'),\ P(F) > 0,\ P(F') > 0. \tag{2.3.1}$$

La condición $P(F) > 0$, $P(F') > 0$ es equivalente a $0 < P(F) < 1$, la que se requiere para que queden bien definidas las probabilidades condicionales. Sin embargo, los casos excluidos corresponden a la ocurrencia de un suceso seguro, lo que no debiera afectar nuestras creencias sobre otros sucesos. Digamos provisionalmente que A es independiente de F cuando (2.3.1) se cumple. Por otra parte, los otros casos corresponden a la ocurrencia de algo seguro, lo que no debiera afectar nuestras creencias sobre la ocurrencia del suceso A. Para evitar imponer esto como condición, es más conveniente reformular (2.3.1) como sigue. Por (2.2.3), $P(A)$ es un promedio ponderado de $P(A|F)$ y de $P(A|F')$, de modo que la igualdad de dos de estas tres cantidades implica que todas son iguales. Por lo tanto, (2.3.1) equivale a

$$P(A|F) = P(A),\ P(F) > 0. \tag{2.3.2}$$

La definición de probabilidad condicional, hace que (2.3.2) equivalga a

$$P(A \cap F) = P(A)P(F), \tag{2.3.3}$$

donde la restricción $P(F) > 0$ ha desaparecido. Desde un punto de vista práctico, (2.3.1), (2.3.2) y (2.3.3) son efectivamente equivalentes.

Intercambiando A con F en la última ecuación, se obtiene $P(F \cap A) = P(F)P(A)$, que es idéntica con (2.3.3). Por esta razón decimos que la condición (2.3.3) es simétrica en A y F. Una consecuencia inmediata es que (2.3.1) equivale a

$$P(F|A) = P(F|A').$$

Por lo tanto, la afirmación *A es independiente de F*, es *matemáticamente equivalente* a *F es independiente de A*. Esta simetría muestra que hay que tener sumo cuidado en la interpretación de esta condición y de su opuesto. Por ejemplo, el precio de una acción hoy incide sobre el precio ma nana. Por simetría, esto indica que este precio futuro afecta el precio de hoy. Existe la tentación de intentar explicaciones sustantivas de este fenómeno, lo cual puede llevar fácilmente a contrasentidos. Esencialmente, la noción probabilística de de dependencia no discrimina entre A causa F y F causa A. En vista de lo anterior, lo habitual es usar (2.3.3) como definición de independencia, lo que cabría traducir como *A y F son independientes*.

Para facilitar las referencias posteriores escribimos la definición formal:

Definición 2.3.1 Los sucesos A y B son independientes si

$$P(A \cap B) = P(A)P(B) \tag{2.3.4}$$

Advertencia: Si dos sucesos de probabilidad positiva son mutuamente excluyentes, la ocurrencia de uno de ellos *garantiza* la no ocurrencia del otro, lo que constituye un caso extremo de dependencia. Pese a esto, suele producirse confusión entre estos conceptos. El siguiente ejemplo complementa estas aseveraciones intuitivas con una demostración rigurosa.

> **Ejemplo 2.3.1** Demostrar que dos sucesos son independientes y mutuamente excluyentes sólo si uno de ellos tiene probabilidad nula.
>
> Si A y B son independientes $P(A \cap B) = P(A)P(B)$. Si ellos son, además, mutuamente excluyentes $P(A \cap B) = 0$. El cumplimiento simultáneo de estas condiciones equivale a $P(A \cap B) = P(A)P(B) = 0$. Como $A \cap B \subseteq A$, esto se cumple si y sólo si $P(A)P(B) = 0$, lo que, a su vez, equivale a $P(A) = 0$ o $P(B) = 0$.

2.4. Teoremas Básicos

En esta sección enunciamos dos teoremas famosos, cuya demostración es notablemente sencilla dentro del enfoque axiomático. Aunque el enunciado habla de una familia numerable de conjuntos, el caso más importante, dentro del presente capítulo, es el caso finito. La única diferencia entre ambos casos es la necesidad del axioma de σ-aditividad.

Teorema 2.4.1 (Ley de probabilidades totales) Considere una familia, posiblemente infinita, de sucesos $(A_i,\ i = 1, 2, \ldots, I)$. Suponga que $P(A_i) > 0$, $i = 1, 2, \ldots, I$, y que exactamente uno de los sucesos A_i ocurre. Si Ω es el espacio muestral, las condiciones se naladas corresponden a la existencia de una partición de Ω con probabilidades positivas para cada elemento de la partición. Entonces, para cualquier suceso B se cumple:

$$\boxed{P(B) = \sum_{i=1}^{I} P(A_i)P(B|A_i)} \qquad \text{Ley de las Probabilidades Totales} \tag{2.4.1}$$

Demostración: Por definición de probabilidad condicional $P(A_i)P(B_j|A_i) = P(A_i \cap B_j)$. Pero

$$B = \sum_{i=1}^{I} A_i \cap B_j,$$

y el resultado es consecuencia de la aditividad. ■

Teorema 2.4.2 (Teorema de Bayes) Bajo las mismas condiciones del teorema 2.4.1, se cumple para cualquier $1 \leq r \leq I$, y cualquier suceso B con $P(B) > 0$, que

$$\boxed{P(A_r|B) = \frac{P(A_r)P(B|A_r)}{\sum_{i=1}^{I} P(A_i)P(B|A_i)} \quad \text{Teorema de Bayes}} \tag{2.4.2}$$

Demostración: Por (2.4.1) el denominador de (2.4.2) coincide con $P(B)$. Por otra parte, se tiene que $P(A_r)P(B|A_r) = P(A_r \cap B)$, de tal forma que el segundo miembro de (2.4.2) es $\frac{P(A_r \cap B)}{P(B)}$ y el resultado se obtiene por definición de la probabilidad condicional. ■

> **Nota Importante:** Recordar que los Teoremas 2.4.1 y 2.4.2 son válidos para I finito o infinito. En el primer caso no se requiere el axioma de σ-aditividad.

En ciertas aplicaciones del Teorema de Bayes se considera a $P(A_i)$ como la probabilidad a priori, es decir, previa a saber que B ocurrió. De esta forma, $P(A_i|B)$ se denomina probabilidad a posteriori, que es la relevante una vez que se sabe que B ocurrió. El denominador en (2.4.2) se cancela al calcular razones entre probabilidades a posteriori:

$$\frac{P(A_i|B)}{P(A_j|B)} = \frac{P(A_i)}{P(A_j)} \, \frac{P(B|A_i)}{P(B|A_j)}. \tag{2.4.3}$$

La razón entre dos probabilidades a posteriori se obtiene multiplicando la razón entre las probabilidades a priori correspondientes por el factor

$$\frac{P(B|A_i)}{P(B|A_j)},$$

que en aplicaciones estadísticas, se denomina *razón de verosimilitud*. En particular, tomando $I = 2$, $A_1 = A$, $A_2 = A'$, y aplicando (2.4.3) se obtiene:

$$\frac{P(A|B)}{1 - P(A|B)} = \frac{P(A)}{1 - P(A)} \, \frac{P(B|A)}{P(B|A')}. \tag{2.4.4}$$

Este resultado, que tiene numerosas aplicaciones, se puede expresar como:

> *Las chances a posteriori se obtienen multiplicando las chances a priori por la razón de verosimilitud.*

> **Ejemplo 2.4.1** Un médico examina la radiografía de tórax de un paciente y está indeciso en su diagnóstico entre cáncer al pulmón y tuberculosis. Sobre la base de información histórica, se estima que la probabilidad que el cáncer produzca una radiografía de este tipo es 0.6, la cual aumenta a 0.8 para la tuberculosis. En su experiencia, el médico estima que el 70 % de los pacientes que consultan por síntomas similares tiene cáncer y el 30 % tiene tuberculosis.

(a) ¿Cuál es la probabilidad que el paciente tenga cáncer?

Sea A_1: el paciente tiene cáncer, A_2: el paciente tiene tuberculosis y B: el paciente tiene una radiografía del tipo observado. Las probabilidades de las 4 ramas son:

Rama	Prob. marginal	Prob. condicional	Producto
A_1B	0,7	0,6	0,42
A_2B	0,3	0,8	0,24
A_1B'	0,7	0,4	0,28
A_2B'	0,3	0,2	0,06

La suma de las dos primeras da $P(B) = 0{,}66$. Por división

$$P(A_1|B) = \frac{42}{66}, \quad P(A_2|B) = \frac{24}{66}.$$

(b) Si la radiografía no hubiera sido del tipo que se observó, se presentaría nuevamente el problema de decidir entre cáncer y tuberculosis. Indique cuál es la probabilidad relevante y calcúlela.

La probabilidad adecuada es $P(A_1|B')$. La probabilidad de B' es la suma de las probabilidades de la tercera y la cuarta ramas, esto es, 0.28 +0.06=0.34. Alternativamente, podemos usar $P(B') = 1 - P(B) = 1 - 0{,}66 = 0{,}34$. La probabilidad buscada es $\frac{0{,}28}{0{,}34} = \frac{28}{34}$.

(c) Obtenga las chances de cáncer en cada una de los casos anteriores y deduzca las probabilidades respectivas.

Aplicamos ahora (2.4.4).

Caso (a):

$$\begin{aligned} \text{Chances de cáncer} &= \frac{7}{3} \times \frac{6}{8} = \frac{7}{4} \\ \text{Prob. de cáncer} &= \frac{7}{7+4} = \frac{7}{11} = \frac{42}{66} \end{aligned}$$

Caso (b):

$$\begin{aligned} \text{Chances de cáncer} &= \frac{7}{3} \times \frac{4}{2} = \frac{14}{3} \\ \text{Prob. de cáncer} &= \frac{14}{14+3} = \frac{14}{17} = \frac{28}{34}. \end{aligned}$$

2.5. Tablas de probabilidades conjuntas y marginales

2.5.1. Tablas para sucesos

Consideremos dos particiones finitas o numerables cualesquiera, $(A_1, \ldots, A_I)$ y $(B_1, \ldots, B_J)$, del espacio muestral Ω, en vez de (A, A'), (F, F'), o (B, B'). Estas dos particiones generan una *partición producto,* cuyos elementos son las intersecciones $A_i \cap B_j$. Ella está así constituida por los

sucesos básicos $A_i \cap B_j$. La representación gráfica natural de esta construcción es una tabla bidimensional, donde la i-ésima fila corresponde a un A_i y la j-ésima columna a un B_j. La intersección de esta fila y esta columna es la celda (i, j), la cual representa al suceso $A_i \cap B_j$. Las probabilidades de los sucesos $A_i \cap B_j$ se denominan *probabilidades conjuntas* y generan una tabla, cuyas celdas contienen estas probabilidades. La suma total es 1 y ellas permiten calcular todas las probabilidades de interés que sean formulables en términos de las dos particiones. En particular, la suma de las probabilidades de la columna encabezada por B_j coincide con $P(B_j)$, por ser $(A_i \cap B_j,\ i = 1, 2, \ldots, I)$ una partición de B_j. Análogamente, el total de la fila encabezada por A_i coincide con $P(A_i)$. Estas *probabilidades marginales* son representables por dos tablas unidimensionales. Es cómodo ubicar las probabilidades marginales $P(A_i)$ en una columna adicional, es decir, como margen derecho de la tabla. Del mismo modo, las probabilidades marginales $P(B_j)$ se ubican en una fila adicional, es decir, como margen inferior. La definición de probabilidad condicional implica

$$\begin{aligned} P(B_j|A_i) &= \frac{P(A_i \cap B_j)}{P(A_i)} \\ P(A_i|B_j) &= \frac{P(A_i \cap B_j)}{P(B_j)}, \end{aligned}$$

o sea,

> *la probabilidad condicional se encuentra dividiendo la probabilidad conjunta por la probabilidad marginal del suceso a la derecha del símbolo "|".*

Las probabilidades $P(A_i|B_j)$ se representan por tablas separadas para cada j, pero es cómodo agruparlas como columnas de una misma tabla. El total de cada columna es ahora igual a 1. Análogamente, las tablas que contienen las probabilidades $P(B_j|A_i)$ se ubican como filas de una tabla común, siendo 1 el total de cada fila.

A continuación mostramos cómo todo este proceso es, en realidad, más difícil de explicarlo que llevarlo a cabo.

Ejemplo 2.5.1 Para $I = 3$, $J = 4$ la tabla de sucesos conjuntos es

	B_1	B_2
A_1	$A_1 \cap B_1$	$A_1 \cap B_2$
A_2	$A_2 \cap B_1$	$A_2 \cap B_2$
A_3	$A_3 \cap B_1$	$A_3 \cap B_2$

Si las probabilidades conjuntas están dadas por la tabla

	B_1	B_2
A_1	0,1	0,3
A_2	0,1	0,2
A_3	0,2	0,1

se puede deducir de aquí las probabilidades marginales

Suceso	Probabilidad
A_1	0,4
A_2	0,3
A_3	0,3

Suceso	Probabilidad
B_1	B_2
0,4	0,6

que son mejor representadas simultáneamente como márgenes de la tabla de probabilidades conjuntas:

	B_1	B_2	
A_1	0,1	0,3	0.4
A_2	0,1	0,2	0.3
A_3	0,2	0,1	0.3
	0.4	0.6	1

Dividiendo cada celda por los número en los márgenes se obtienen las probabilidades condicionales:

	$P(B_1\|A_i)$	$P(B_2\|A_i)$	Total
A_1	$\frac{1}{4}$	$\frac{3}{4}$	1
A_2	$\frac{1}{3}$	$\frac{2}{3}$	1
A_3	$\frac{2}{3}$	$\frac{1}{3}$	1

	B_1	B_2
$P(A_1\|B_j)$	$\frac{1}{4}$	$\frac{3}{6}$
$P(A_2\|B_j)$	$\frac{1}{4}$	$\frac{2}{6}$
$P(A_3\|B_j)$	$\frac{2}{4}$	$\frac{1}{6}$
Total	1	1

Las probabilidades en las tablas anteriores pueden interpretarse como proporciones. Supongamos, por ejemplo, que $i = 1, 2, 3$ corresponde a nivel socio económico bajo, medio y alto, y que $j = 2$ significa estar a favor de un proyecto de rebaja de aranceles. De una encuesta a 1000 personas se pueden obtener proporciones que coinciden numéricamente con las probabilidades conjuntas. Se invita al lector a reinterpretar las demás tablas en este nuevo contexto.

2.5.2. Espacio producto y tablas para variables

En la práctica, las particiones $(A_1, \ldots, A_I)$ y $(B_1, \ldots, B_J)$ son inducidas por dos variables discretas X e Y respectivamente. Cuando la partición está formada por un suceso y su negación, e.g., (F, F'), (B, B'), la variable es binaria. Sea $\{x_1, \ldots, x_i, \ldots, x_I\}$ una enumeración del conjunto de valores $\mathcal{X}$ de X y sea $\{y_1, \ldots, y_j, \ldots, y_J\}$ una enumeración del conjunto de valores $\mathcal{Y}$ de Y. Definamos los sucesos A_i y B_j por $X = x_i$ e $Y = y_j$ respectivamente. Entonces $A_i \cap B_j \Leftrightarrow (X = x_i, Y = y_j)$. El espacio muestral más cómodo es el *espacio muestral producto* asociado al par de variables (X, Y), que es el producto cartesiano

$$\mathcal{X} \times \mathcal{Y} = \{(x_i, y_j),\ i = 1, \ldots, I;\ j = 1, \ldots, J.\}.$$

Las etiquetas i y j son arbitrarias e innecesarias. En vez de $X = x_i$ y $Y = y_j$, es preferible escribir $X = x$ y $Y = y$ respectivamente. El punto (x, y) corresponde a la realización *conjunta* de los sucesos $X = x$ e $Y = y$. Este calificativo se extiende a las probabilidad correspondientes, así como a la función de probabilidad p, la que se denota por $p_{X,Y}$ si se desea evitar confusiones. Si los sucesos de interés dependen sólo de la variable X, el espacio muestral natural es $\mathcal{X}$ y lo llamamos

espacio marginal asociado a X. El calificativo marginal se emplea también para los sucesos $X = x$, para sus probabilidades, así como para la función de probabilidad definida sobre $\mathcal{X}$, a la que denotamos por p_X, dada por $p_X(x) = P(X = x)$. Algo semejante ocurre con la variable Y.

Las variables aleatorias X e Y se pueden interpretar como resultados potenciales de la primera y segunda etapa de un experimento. Con esta notación, podemos reformular los resultados para sucesos en términos de una variable discreta.

La regla multiplicativa se traduce en

$$p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x). \tag{2.5.1}$$

Los teoremas fundamentales cobran un aspecto más amistoso:

Teorema 2.5.1 (Ley de probabilidades totales) Sea X una variable discreta con función de probabilidad positiva. Entonces

$$p_Y(y) = \sum_{x\in\mathcal{X}} p_X(x)p_{Y|X}(y|x). \tag{2.5.2}$$

Teorema 2.5.2 (Teorema de Bayes para una variable discreta) Sea X una variable discreta con función de probabilidad positiva.

$$p_{X|Y}(x_0|y) = \frac{p_X(x_0)p_{Y|X}(y|x_0)}{\sum_{x\in\mathcal{X}} p_X(x)p_{Y|X}(y|x)}. \tag{2.5.3}$$

2.6. Experimentos secuenciales

2.6.1. Construcción del espacio muestral

Se lleva a cabo un experimento E_0 obteniéndose un resultado x. De acuerdo a cual sea este resultado, se realiza un segundo experimento, que denotamos por E_x. No hay, a priori, ninguna relación entre los experimentos E_x. Llamamos $\mathcal{X}$ al espacio muestral, que suponemos numerable, asociado con el experimento E_0 y denotamos por p_X a su función de probabilidad. Del mismo modo, denotaremos por $\mathcal{Y}_x$ al espacio muestral correspondiente al experimento E_x.

Sea y el resultado de E_x. Si no conocemos x, el conjunto de valores posibles de la variable correspondiente Y es

$$\mathcal{Y} = \bigcup_{x\in\mathcal{X}} \mathcal{Y}_x.$$

El resultado del experimento bietápico es $(x, y) \in \mathcal{X} \times \mathcal{Y}$, pero algunos elementos de este conjunto pueden ser imposibles. Para no cambiar de espacio muestral, le asignamos probabilidad cero a tales puntos. Por ejemplo, en muestras sin reposición en que x e y identifican completamente cada ficha, los puntos (x, x) son imposibles.

Ejemplo 2.6.1 Sea E_0 el lanzamiento de una moneda. Si sale cara se elige un número al azar del conjunto $\{a_1, a_2, a_3\}$; si sale sello, se elige un número al azar del conjunto

$\{b_1, b_2\}$. Codificando cara =1, sello =2, tenemos $\mathcal{X} = \{1, 2\}$, $\mathcal{Y}_1 = \{a_1, a_2, a_3\}$ e $\mathcal{Y}_2 = \{b_1, b_2\}$. Si sólo observáramos el resultado del segundo experimento, el espacio muestral sería $\mathcal{Y} = \{a_1, a_2, a_3, b_1, b_2\}$.

Ejemplo 2.6.2 Se lanza un dado y luego se lanza una moneda tantas veces como el número que indica el dado. En este caso $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Codificando cara =1, sello =0, tenemos

x	$\mathcal{Y}_x$
1	$\{0, 1\}$
2	$\{0, 1\} \times \{0, 1\}$
3	$\{0, 1\} \times \{0, 1\} \times \{0, 1\}$
4	$\{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\}$
5	$\{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\}$
6	$\{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\}$

En este caso $\mathcal{Y}$ resulta altamente artificial. Si consideramos como resultado del segundo experimento al número total de caras, esto se simplifica a

x	$\mathcal{Y}_x$
1	$\{0, 1\}$
2	$\{0, 1, 2\}$
3	$\{0, 1, 2, 3\}$
4	$\{0, 1, 2, 3, 4\}$
5	$\{0, 1, 2, 3, 4, 5\}$
6	$\{0, 1, 2, 3, 4, 5, 6\}$

Si el interés está sólo en el lanzamiento del dado, entonces $\mathcal{Y} = \{0, 1, 2, 3, 4, 5, 6\}$.

2.6.2. Identificación con probabilidades condicionales y regla multiplicativa

Por comodidad usaremos el lenguaje de variables. El supuesto clave es que tanto el espacio muestral como las probabilidades asociadas con el experimento E_x son conocidos. Denotemos por $\pi_x(\cdot)$ la función de probabilidad asociada a E_x. Desde un punto de vista frecuentista, la proporción de veces que se observa x tiende a $p_X(x) = P(X = x)$ y la proporción de veces que se obtiene (x, y) tiene como límite a $p_X(x)\pi_x(y)$. Pero sabemos que este último límite coincide con $p_{X,Y}(x, y) = P(X = x, Y = y)$. Por definición de probabilidad condicional

$$\pi_x(y) = P(Y = y|X = x) = p_{Y|X}(y|x).$$

La idea básica es:

Identificar la función de probabilidad asociada al experimento E_x con la función de probabilidad condicional de Y dado $X = x$.

Ejemplo 2.6.3 Calculemos la probabilidad de obtener dos fichas blancas al extraer dos fichas, sin reposición, de una urna que contiene dos fichas blancas y una negra. Si x e y son los colores (b o n) de la primera y segunda ficha, $\mathcal{X} = \mathcal{Y} = \{b, n\}$. Por (2.5.1) la probabilidad buscada es

$$p_{X,Y}(b, b) = p_X(b)P(Y = b|X = b).$$

En el cálculo de $p_X(b)$ podemos ignorar el hecho que habrá una segunda extracción. Por equiprobabilidad se obtiene $P_X(b) = \frac{2}{3}$. Por otra parte, dado $X = b$, se genera físicamente una nueva urna compuesta por una ficha de cada color. El experimento E_b consiste en extraer una ficha al azar de esta urna y anotar su color, de modo que $\pi_b(b) = \frac{1}{2}$. Por lo tanto $P(Y = b|X = b) = \frac{1}{2}$. De aquí $p(b, b) = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$. El lector debe verificar que el mismo resultado se obtiene considerando el espacio equiprobable de las 6 muestras ordenadas.

En el ejemplo anterior, la representación bietápica del experimento es natural en el contexto del problema planteado. En otros ejemplos, esto no ocurre, pero la representación bietápica sigue siendo válida en un nivel puramente conceptual. Cabe recordar que las probabilidades obtenidas no dependen de la representación elegida; esta última es una herramienta de cálculo que puede o no ser útil.

Por ejemplo, en la extracción de una muestra al azar ordenada y sin reposición, podemos pensar que la primera etapa determina el conjunto de valores obtenido, mientras que la segunda genera un orden particular. Podemos también revertir el orden del tiempo y considerar como primera etapa la segunda ficha extraida. En general, todo problema con espacio muestral $\mathcal{X} \times \mathcal{Y}$ y función de probabilidad $p_{X,Y}$ se puede representar secuencialmente. Simplemente se inventa un experimento E_x con espacio muestral

$$\mathcal{Y}_x = \{y \in \mathcal{Y} / p_{X,Y}(x, y) > 0\}. \tag{2.6.1}$$

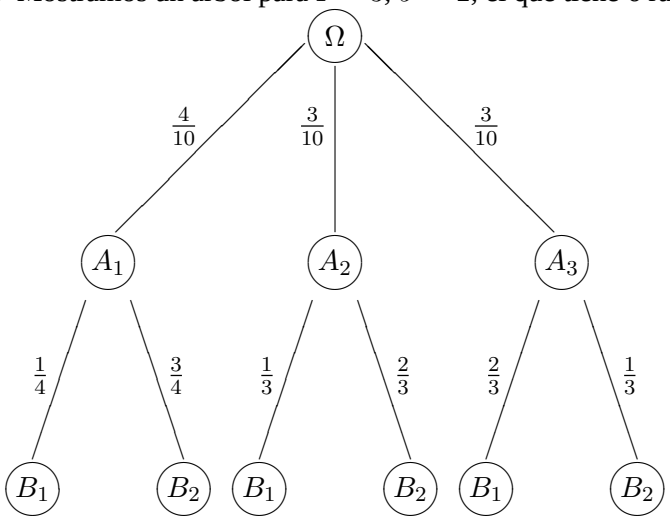
2.6.3. Representación por árboles

Un árbol es una estructura matemática formada por nodos y arcos dirigidos. Una variable discreta X genera un nodo para cada valor x. Previamente existe un nodo origen, rotulado por O, el que se une a cada uno de estos nodos generando *arcos dirigidos* que emanan del origen, a los que se denota por Ox. Al nodo Ox se le asigna la probabilidad $p_X(x) = P(X = x)$. La suma de los valores asignados a todos los arcos emergentes del nodo origen es, por tanto, igual a 1.

Consideremos ahora una segunda variable discreta Y. A partir de cada nodo x se dibujan arcos emergentes con nodos terminales rotulados por los valores de y, lo que genera un nuevo árbol a partir de cada nodo rotulado por x. Juntando todos estos árboles con el árbol original, se forma uno más grande en que aparecen *ramas*, constituidas por los arcos Ox y xy. Las ramas está en correspondencia uno a uno con los pares de valores (x, y) y con los nodos terminales. Hay que distinguir acá entre nodo y rótulo del nodo. Pueden haber muchos nodos terminales con el rótulo y, pero a cada uno de estos nodos llega un solo arco, que proviene de un nodo primario dado. Cada rama se puede interpretar como un resultado del experimento bietápico. El origen del árbol se puede asociar con el suceso seguro Ω. El producto de los números asignados a los arcos de la

rama Oxy es $p_X(x)p_{Y|X}(y|x)$, que coincide con $p_{X,Y}(x,y)$. La regla multiplicativa corresponde, así, a multiplicar los números de una rama. Esto es fácilmente extensible a k variables $X_1, \ldots, X_k$, lo que veremos en la próxima sección.

Ejemplo 2.6.4 Mostramos un árbol para $I = 3$, $J = 2$, el que tiene 6 ramas.



Cuando el experimento se describe secuencialmente, podemos asignar E_0 al origen y E_x al nodo x. Las probabilidades que emanan de ese nodo corresponden, en este caso, a las probabilidades $\pi_x(y)$.

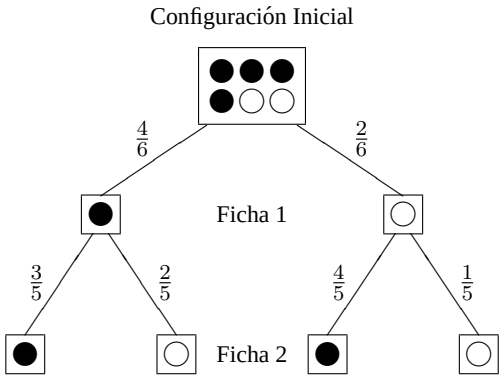
Ejemplo 2.6.5 Sea una urna con 4 fichas negras y 2 blancas. Se extrae una muestra ordenada de tamaño 2. Considere los colores del par de fichas extraido como resultado del experimento.

(a) Calcule la probabilidad de cada resultado.

(b) Calcule la probabilidad que la segunda ficha sea negra.

(c) Calcule la probabilidad que la primera ficha sea negra, dado que la segunda también lo es.

El diagrama de árbol que se muestra en la figura es una forma razonable de abordar este problema. De este modo, las probabilidades de cada rama se obtienen de multiplicar los números sobre cada arco, las que corresponden a lo pedido en (a).



Así, $P(nn) = \frac{4}{6} \times \frac{3}{5} = \frac{6}{15}$, $P(nb) = \frac{4}{6} \times \frac{2}{5} = \frac{4}{15}$, $P(bn) = \frac{2}{6} \times \frac{4}{5} = \frac{4}{15}$ y $P(bb) = \frac{2}{6} \times \frac{1}{5} = \frac{1}{15}$. La probabilidad pedida en (b) se obtiene simplemente de sumar las probabilidades de las ramas que terminan en $\bullet$, lo que da $\frac{6}{15} + \frac{4}{15} = \frac{2}{3}$. Finalmente, lo pedido en (c) es una aplicación del Teorema de Bayes. El resultado es $\frac{3}{5}$, y los detalles se dejan al lector como ejercicio.

2.6.4. Relación entre tablas y árboles

Hay una correspondencia uno a uno entre el conjunto de ramas, el conjunto de nodos terminales, el conjunto de sucesos $A_i \cap B_j$, y el conjunto $\mathcal{X} \times \mathcal{Y}$. Esto indica que se puede elegir Ω como el conjunto de ramas del árbol o como el conjunto de nodos terminales. Si se obtiene la probabilidad de cada rama por multiplicación, y se organizan estos productos en la tabla de probabilidades conjuntas, se puede obtener los márgenes. El margen inferior entrega las probabilidades buscadas, mientras que los números en el margen derecho deben coincidir con los valores $P(A_i)$, que son un dato del problema.

A partir de la tabla de probabilidades conjuntas, el margen derecho entrega las probabilidades de los nodos primarios. Dividiendo la probabilidad de cada celda por el número correspondiente en esta marginal se encuentran las probabilidades de los arcos que conectan un nodo primario con uno secundario.

Ejemplo 2.6.6 En el Ejemplo 2.6.5, sean X e Y el color de las fichas extraidas la primera y segunda vez, respectivamente. Se tiene entonces que $\mathcal{X} = \mathcal{Y} = \{b, n\}$, y las probabilidades conjuntas se obtienen de efectuar la multiplicación en cada rama:

	$Y = n$	$Y = b$	Total
$X = n$	$\frac{4}{6} \times \frac{3}{5} = \frac{12}{30}$	$\frac{4}{6} \times \frac{2}{5} = \frac{8}{30}$	$\frac{2}{3}$
$X = b$	$\frac{2}{6} \times \frac{4}{5} = \frac{8}{30}$	$\frac{2}{6} \times \frac{1}{5} = \frac{2}{30}$	$\frac{1}{3}$
Total	$\frac{2}{3}$	$\frac{1}{3}$	1

La probabilidad de un suceso cualquiera que depende de las variables X e Y se puede calcular en dos pasos:

1. Identificar las ramas favorables, i.e. aquellas para las cuales el suceso ocurre.
2. Multiplicar los números de los arcos de estas ramas para obtener la probabilidad de cada rama favorable.
3. Sumar las probabilidades del punto anterior.

Si lo que se desea es obtener la distribución marginal de Y, entonces:

1. Multiplicar los números de los arcos de estas ramas para obtener la probabilidad de cada rama.
2. Sumar las probabilidades de todas las ramas con nodo terminal y.

Ejemplo 2.6.7 Retomamos acá el Ejemplo 2.2.1 de las tres cartas. Mostramos que el problema se puede también resolver aplicando el Teorema de Bayes. Sea X el número de la carta, y sea $Y = b$ o $Y = n$ según sea blanco o negro el color mostrado. El árbol con ramas (x, y) tiene 6 ramas, aunque 2 de ellas tienen probabilidad nula.

Rama	$p_X(x)$	$p_{Y\|X}(y\|x)$	$p_{X,Y}(x,y)$
$1b$	$\frac{1}{3}$	1	$\frac{1}{3}$
$1n$	$\frac{1}{3}$	0	0
$2b$	$\frac{1}{3}$	0	0
$2n$	$\frac{1}{3}$	1	$\frac{1}{3}$
$3b$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$
$3n$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Organizando los valores de la última columna se obtiene la función de probabilidad conjunta de X e Y:

	$x=1$	$x=2$	$x=3$
$y=b$	$\frac{1}{3}$	0	$\frac{1}{6}$
$y=n$	0	$\frac{1}{3}$	$\frac{1}{6}$

y de aquí la tabla de funciones de probabilidad condicional de X dado $Y = y$:

y	$p_{X\|Y}(1\|y)$	$p_{X\|Y}(2\|y)$	$p_{X\|Y}(3\|y)$
b	$\frac{2}{3}$	0	$\frac{1}{3}$
n	0	$\frac{1}{3}$	$\frac{2}{3}$

Ejemplo 2.6.8 Suponga que en el Ejemplo 2.6.2 se han obtenido dos caras. Calcule la función de probabilidad del número que salió en el dado, condicional en esta información.

Rama	Prob. marginal	Prob. condicional	Producto
$X=1, Y=3$	$\frac{1}{6}$	0	0
$X=2, Y=3$	$\frac{1}{6}$	0	0
$X=3, Y=4$	$\frac{1}{6}$	0	0
$X=4, Y=4$	$\frac{1}{6}$	$\frac{1}{16}$	$\frac{4}{64}$
$X=5, Y=4$	$\frac{1}{6}$	$\frac{5}{32}$	$\frac{10}{64}$
$X=6, Y=4$	$\frac{1}{6}$	$\frac{15}{64}$	$\frac{15}{64}$

La función de probabilidad condicional es proporcional a la última columna. Omitiendo puntos de probabilidad nula se tiene:

x	4	5	6
$p_{X\|Y}(x\|4)$	$\frac{4}{29}$	$\frac{10}{29}$	$\frac{15}{29}$

2.7. Experimentos multietápicos

2.7.1. Cálculo de probabilidades conjuntas

Sea X_i la variable que representa el resultado potencial de la i–ésima etapa, y sea $\mathcal{X}_i$ el conjunto de m_i valores posibles de esta variable. Es conveniente escribir el resultado en la forma $\boldsymbol{x} = (x_1, x_2, \ldots, x_k) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k$. El número total de resultados,

$$\text{card } (S) = m = \prod_{i=1}^{k} m_i,$$

puede ser muy grande, tornando inviable una asignación directa de la probabilidad para cada $\boldsymbol{x}$. Por ejemplo, si se lanza una moneda 100 veces, el número de resultados posibles asciende a la astronómica cifra de 2^{100}. Consideremos la descripción secuencial:

Etapa 1.	Se realiza experimento E_0, obteniéndose $X_1 = x_1$.
Etapa 2.	Se realiza experimento E_{x_1}, obteniéndose $X_2 = x_2$.
$\vdots$	$\vdots$
Etapa r.	Se realiza experimento $E_{x_1 x_2 \cdots x_{r-1}}$, obteniéndose $X_r = x_r$.
$\vdots$	$\vdots$
Etapa k.	Se realiza experimento $E_{x_1 x_2 \cdots x_{k-1}}$, obteniéndose $X_k = x_k$.

Para $r > 0$ usamos la notación $\boldsymbol{x}_r = (x_1, x_2, \ldots, x_r)$ y $\boldsymbol{X}_r = (X_1, X_2, \ldots, X_r)$, de modo que $P(X_r = x_r | X_j = x_j,\ j < r) = P(X_r = x_r | \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1})$. Considerando a $\boldsymbol{x}_{r-1}$ como el resultado de una primera macro-etapa, la función de probabilidad sobre el espacio muestral asociado con $E_{\boldsymbol{x}_{r-1}}$ coincide con las función probabilidad condicional de $(X_r | \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1})$.

Teorema 2.7.1 *Sean $X_1, X_2, \ldots$ variables aleatorias. Sea $\alpha_1 = \beta_1 = P(X_1 = x_1)$ y*

$$\alpha_r = P(X_r = x_r | \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1}), \quad \beta_r = P(\boldsymbol{X}_r = \boldsymbol{x}_r).$$

Entonces

$$\begin{aligned} &\beta_r = \beta_{r-1}\alpha_r \\ &\beta_r = \prod_{i=1}^{r} \alpha_i, \quad r = 1, 2, \ldots \end{aligned} \tag{2.7.1}$$

Demostración: La segunda igualdad en (2.7.1) se obtiene aplicando la primera recursivamente y $\alpha_1 = \beta_1$. La primera es consecuencia directa de la definición de probabilidad condicional:

$$\begin{aligned} \alpha_r &= P(X_r = x_r | \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1}) \\ &= P(X_r = x_r, \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1} | \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1}) / P(\boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1}) \\ &= P(\boldsymbol{X}_r = \boldsymbol{x}_r) / P(\boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1}) \\ &= \frac{\beta_r}{\beta_{r-1}}. \qquad \blacksquare \end{aligned}$$

Observaciones:

- Si el resultado $\boldsymbol{x}$ se interpreta como la rama de un árbol que pasa por los nodos $x_1, x_2, \ldots$, la probabilidad $\alpha_r = P(X_r = x_r | \boldsymbol{X}_{r-1} = \boldsymbol{x}_{r-1})$ se asigna al arco que une a x_{r-1} con x_r. Ella corresponde a la probabilidad del resultado x_r en el experimento $E_{\boldsymbol{x}_{r-1}}$.

- Si el suceso $X_i = x_i$ se reemplaza por un suceso cualquiera A_i, (2.7.1) se satisface con $\alpha_1 = \beta_1 = P(A_1)$ y

$$\alpha_r = P(A_r | \bigcap_{i=1}^{r-1} A_i), \quad \beta_r = P(\bigcap_{i=1}^{r} A_i).$$

Ejemplo 2.7.1 Para 4 variables X_1, X_2, X_3 y X_4, (2.7.1) genera las tres igualdades:

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) P(X_2 = x_2 | X_1 = x_1)$$

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, X_3 = x_3) &= P(X_1 = x_1, X_2 = x_2) \\ &\quad \times P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \end{aligned}$$

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) &= P(X_1 = x_1) \\ &\quad \times P(X_2 = x_2 | X_1 = x_1) \times P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \\ &\quad \times P(X_4 = x_4 | X_1 = x_1, X_2 = x_2, X_3 = x_3) \end{aligned}$$

A continuación, mencionamos varias representaciones simbólicas de estas igualdades. La estructura básica es clara, pero hay que indicar de alguna manera el orden de las variables o sucesos.

- En términos de las variables, escribimos

$$\begin{aligned} (X_1 X_2) &= (X_1)(X_2|X_1) \\ (X_1 X_2 X_3) &= (X_1)(X_2|X_1)(X_3|X_1 X_2) \\ (X_1 X_2 X_3 X_4) &= (X_1)(X_2|X_1)(X_3|X_1 X_2)(X_4|X_1 X_2 X_3) \end{aligned}$$

- Eliminando símbolos redundantes, esto se simplifica a

$$\begin{aligned} (12) &= (1)(2|1) \\ (123) &= (1)(2|1)(3|12) \\ (1234) &= (1)(2|1)(3|12)(4|123) \end{aligned}$$

- Para cuatro sucesos A, B, C, D podemos escribir

$$\begin{aligned} (CB) &= (C)(B|C) \\ (CBD) &= (C)(B|C)(D|CB) \\ (CBDA) &= (C)(B|C)(D|CB)(A|CBD). \end{aligned}$$

La segunda ecuación, por ejemplo, representa la igualdad

$$P(C \cap B \cap D) = P(C) P(B|C) P(D|C \cap B).$$

Ejemplo 2.7.2 Ilustraremos las relaciones entre diversas funciones de probabilidad asociadas con tres variables aleatorias. En primer lugar, mostramos la notación a través de algunos ejemplos:

$$\begin{aligned} P(X=x,Y=y,Z=z) &= p_{X,Y,Z}(x,y,z) \\ P(X=x,Y=y) &= p_{X,Y}(x,y) \\ P(X=x) &= p_X(x) \\ P(Y=y|X=x) &= p_{Y|X}(y|x) \\ P(Z=z|X=x,Y=y) &= p_{Z|X,Y}(z|x,y) \\ P(Y=y,Z=z|X=x) &= p_{Y,Z|X}(y,z|x) \end{aligned}$$

El axioma de aditividad permite establecer relaciones usando sumas con respecto a los argumentos adecuados en las funciones de probabilidad. Por ejemplo:

$$\begin{aligned} p_{X,Y}(x,y) &= p_{X,Y,Z}(x,y,+) \\ p_X(x) &= p_{X,Y}(x,+) \\ &= p_{X,Y,Z}(x,+,+) \end{aligned}$$

La aditividad y el axioma de normalización (probabilidad del espacio muestral es 1) producen relaciones como

$$\begin{aligned} p_{X,Y,Z}(+,+,+) &= 1 \\ p_{X,Y}(+,+) &= 1 \\ p_X(+) &= 1 \end{aligned}$$

Aplicando el axioma de normalización a las funciones de probabilidad condicionales se obtienen igualdades como

$$\begin{aligned} p_{Y|X}(+|x) &= 1 \\ p_{Z|X,Y}(+|x,y) &= 1 \\ p_{Y,Z|X}(y,+|x) &= p_{Y|X}(y|x) \\ p_{Y,Z|X}(+,+|x) &= 1 \end{aligned}$$

Ejemplo 2.7.3 Se extraen 3 fichas sin reemplazo de una urna con 4 fichas blancas y 5 fichas negras. Definamos las variables

$$X = x \ \ (\text{color } 1^{era} \text{ ficha})$$

$$Y = y \ \ (\text{color } 2^{da} \text{ ficha})$$

$$Z = z \ \ (\text{color } 3^{era} \text{ ficha})$$

Interesa calcular $p(b,b,n)$.

Experimento E_0: Extraer ficha de la urna [4b,5n]. P(blanca)=$\frac{4}{9}$. Experimento E_b: Extraer ficha de la urna [3b,5n]. P(blanca)=$\frac{3}{8}$. Experimento E_{bb}: Extraer ficha de la urna [2b,5n]. P(negra)=$\frac{5}{7}$.

Las probabilidades indicadas corresponden a la proporción de fichas blancas en la urna que se indica. La rama del árbol

tiene asociada la probabilidad $\frac{60}{504} = \frac{4}{9} \times \frac{3}{8} \times \frac{5}{7}$. En términos de funciones de probabilidad, los números de los arcos son

$$p_X(b) = \frac{4}{9}, \quad p_{Y|X}(b|b) = \frac{3}{8}, \quad p_{Z|X,Y}(n|b,b) = \frac{5}{7}.$$

2.7.2. Dos casos particulares

La definición de las probabilidades usando la representación multietápica no representa, en general, un ahorro en el número de probabilidades independientes que hay que especificar. Para calcular este número, hay que recordar las igualdades de suma total igual a 1. Denotemos por M_r a $m_1 \times \cdots \times m_{r-1} \times m_r$. Como el número total de arreglos $\boldsymbol{x}$ es M_k, una asignación directa requiere $M_k - 1$ probabilidades independientes. Utilizando el lenguaje de árboles, hay $(m_1 - 1)$ probabilidades independientes para los m_1 arcos que emanan del origen. De cada nodo de orden $r-1$, que representa a $\boldsymbol{x}_{r-1}$, emergen m_r arcos, lo que requiere especificar $m_r - 1$ probabilidades, para cada uno de M_{r-1} nodos, es decir, $M_{r-1}(m_r - 1) = M_r - M_{r-1}$. Sumando de $r = 1$ hasta $r = k$ se obtiene una suma telescópica, que coincide con $M_k - 1$.

La representación multietápica es particularmente atractiva cuando en el cálculo de las probabilidades α_r, no es necesario especificar toda la historia pasada. En esta sección describimos brevemente los casos más importantes. El primero se retoma en la próxima sección desde otro punto de vista.

- **Irrelevancia de toda la historia.** En este caso α_r depende sólo de x_r, de modo que podemos escribir $\alpha_r(x_r)$. La probabilidad asociada a un arco depende, entonces, sólo del nodo de llegada. Se requiere especificar un total de

  $$(m_1 - 1) + \cdots + (m_k - 1)$$

  probabilidades. Por ejemplo,

  $$P(X_1{=}1, X_2{=}3, X_3{=}4, X_4{=}3, X_5{=}2) = \alpha_1(1)\alpha_2(3)\alpha_3(4)\alpha_4(3)\alpha_5(2).$$

  En la próxima sección veremos que el supuesto de historia irrelevante coincide con el de independencia de variables aleatorias. Si $X_1, X_2, \ldots$ tienen la misma distribución, basta especificar $m_1 - 1$ probabilidades.

- **La historia influye sólo a través del valor de la última variable.** Esto quiere decir que α_r depende sólo de x_{r-1} y x_r, para $r > 1$, de modo que escribimos $\alpha_1(x_1)$ y $\alpha_r(x_{r-1}, x_r)$, $r > 1$. Por ejemplo,

  $$P(X_1{=}1, X_2{=}3, X_3{=}4, X_4{=}3, X_5{=}2) = \alpha_1(1)\alpha_2(1,3)\alpha_3(3,4)\alpha_4(4,3)\alpha_5(3,2).$$

La propiedad descrita acá es conocida como *propiedad markoviana,* término que deriva del apellido de un eminente matemático ruso. Si se identifica a r como una versión discreta del tiempo y a x_r como el estado de un sistema en el tiempo r, las funciones α_r determinan el mecanismo de evolución probabilística del sistema. Lo más habitual es que el conjunto $\mathcal{X}_r$ de valores para x_r se pueda elegir como el mismo para todo r. Si denotamos por S a este conjunto común, decimos que S es el *espacio de estados.*

Si $\text{card}\ (S) = m$, la función α_1 está determinada por $m-1$ probabilidades, y lo propio ocurre con cada función $\alpha_r(x_{r-1}, \cdot)$. En total se requiere especificar $(m-1)+(k-1)(m(m-1)) = (m-1)(1+m(k-1))$ números. Un caso muy importante es el de un proceso *homogéneo,* en el sentido que los α_r son todos idénticos de $r = 2$, en adelante. Basta entonces especificar α_1 y α_2, lo que da $m^2 - 1$ constantes en total.

Si las k variables tienen p valores cada una, el número de probabilidades independientes, para varios casos de interés, se muestra en la siguiente tabla:

Sin restricciones:	$p^k - 1$
Caso markoviano:	$(p-1) + p(p-1)(k-1)$
Caso markoviano homogéneo:	$p^2 - 1$
Independencia:	$k(p-1)$
Independencia y homogeneidad:	$p-1$

Para $p = 2$ estos números se reducen a $2^k - 1$, $(2k-1)$, 3, k y 1 respectivamente.

Ejemplo 2.7.4 Representación Markoviana del Problema de Urnas: La probabilidad que en la r-ésima etapa la ficha extraida sea de un color determinado no depende sólo del color de la última ficha extraida. Sin embargo, podemos definir el estado del sistema para que el modelo sea markoviano. Una elección natural es la composición de la urna inmediatamente antes de extraer una ficha, o sea, el número de fichas de cada color. Sea, entonces, $y_i = (y_{ib}, y_{in})$, con $y_{ib} = \text{N}^\circ$ de fichas blancas e $y_{in} = \text{N}^\circ$ de fichas negras, después de la i-ésima extracción. Sea $x_i = 1$ si la i-ésima ficha es blanca y $x_i = 0$ si ella es negra. Los valores de las variables X_i y la composición inicial de la urna determinan la evolución de su contenido.

Supongamos que la composición inicial es de 4 fichas blancas y 5 negras, es decir, $y_0 = (4, 5)$. Si se extraen dos fichas blancas seguidas de una negra, $X_1 = 1$, $X_2 = 1$, $X_3 = 0$, de donde $y_1 = (3, 5)$, $y_2 = (2, 5)$ e $y_3 = (2, 4)$. En el caso del muestreo sin reposición, el número de fichas decrece en 1 con cada extracción. Por esta razón, se puede también elegir como estado del sistema a un elemento del par y_i. Esto facilita la escritura, aunque hace más difícil la comprensión de la notación. La propiedad markoviana implica

$$P(\boldsymbol{y} = (3, 2, 2)) = P(Y_{1b} = 3)P(Y_{2b} = 2|Y_{1b} = 3)P(Y_{3b} = 2|Y_{2b} = 2).$$

Condicional en los sucesos a la derecha de |, podemos expresar los sucesos a la izquier-

da de | en función de los X_i:

$$\begin{aligned} P(\boldsymbol{y}=(3,2,2)) &= P(X_1=1)\times P(X_2=1|Y_{1b}=3)\times P(X_3=0|Y_{2b}=2) \\ &= P(X_1=1|Y_0=(4,5))\times P(X_2=1|Y_1=(3,5)) \\ &\quad \times P(X_3=0|Y_2=(2,5)) \\ &= \frac{4}{9}\times\frac{3}{8}\times\frac{5}{7}. \end{aligned}$$

Hay 8 trayectorias posibles y se puede calcular la probabilidad de cada una usando una regla multiplicativa, al igual que en el caso particular descrito. Se deja el lector dibujar el árbol correspondiente, asignando las probabilidades a cada arco, y obteniendo las probabilidades de las ramas por multiplicación.

2.8. Noción general de independencia

2.8.1. Motivación

Si se lanzan 5 dados (equilibrados o no), la intuición indica que lo que muestra el tercer dado no afecta, en absoluto, como se comporta el quinto. Tampoco pareciera que lo que muestran los dos primeros influirá sobre la suma de los números de los otros tres. En general, no parece haber asociación entre los resultados de los cinco dados. Esta propiedad se parece a la de independencia, pero la definición formal (2.3.4) se queda muy corta. Es fundamental generalizarla a más de dos sucesos. El ejemplo de los dados sugiere la idea de independencia de variables. En efecto, el lanzamiento de 5 dados se puede ver como un experimento con 5 *etapas*, correspondiendo la i-ésima al lanzamiento del i-ésimo dado. El *resultado* natural de esta etapa es el número x_i que muestra el dado, al que consideramos como el *valor* o *realización* de una *variable* X_i.

Los sucesos que dependen sólo del resultado del i-ésimo dado son aquellos expresables en términos de la variable X_i. Aquellos que dependen sólo de los dados $i_1, i_2, \ldots, i_p$ son los expresables en términos de $(X_i,\ i\in\{i_1,i_2,\ldots,i_p\})$. Por ejemplo, el suceso A= *obtener el mismo número en los dados 4 y 5* es $X_4 = X_5$; el suceso B= *La suma de los números de los dados 3, 4 y 5 es mayor que 10* se escribe como $X_3+X_4+X_5>10$; el suceso C: *El número del segundo dado es mayor que el del primero* se escribe como $X_2 - X_1 > 0$, etc.

Intuitivamente, los sucesos B y C son independientes, pues dependen de conjuntos disjuntos de variables, es decir, $\{X_3, X_4, X_5\}\cap\{X_1, X_2\} = \emptyset$. El mismo argumento sugiere que $X_1 + X_2 = 6$, X_3 par, y $X_5 > X_4$ son sucesos independientes, pero aún no hemos definido la independencia de tres sucesos.

Los sucesos $X_i = j$ se pueden representar por un subconjunto A_{ij}, los que constituyen una partición del espacio muestral Ω (conjunto que aún no ha sido definido). Aquellos sucesos que dependen sólo de la i-ésima etapa son expresables como uniones de algunos de los $(A_{ij},\ j=1,\ldots,6)$. Para sucesos cualesquiera que dependan de los números que aparecen en los lanzamientos, ellos se

pueden expresar como uniones finitas de los conjuntos

$$B(\boldsymbol{x}) = \bigcap_{i=1}^{5} A_{ix_i},$$

que representan a los sucesos elementales $\boldsymbol{X} = \boldsymbol{x}$.

Es claro que el lenguaje de variables es mucho más atractivo que el de sucesos expresados como subconjuntos de un gran espacio Ω. Lo que haremos es proponer definiciones válidas para variables y luego mostrar como se recuperan las definiciones tradicionales de sucesos independientes.

2.8.2. Definiciones y teoremas

Definición 2.8.1 Las variables discretas $X_1, \ldots, X_k$ son independientes si

$$P(X_1 \in A_1, \ldots, X_k \in A_k) = \prod_{i=1}^{k} P(X_i \in A_i), \text{ para todo } A_i,\ i = 1, \ldots, k. \tag{2.8.2}$$

Las variables aleatorias en la sucesión $X_1, X_2, \ldots$ son independientes si para cualquier k finito, $X_1, \ldots, X_k$ son independientes.

Teorema 2.8.1 (Factorización) *Si $X_1, \ldots, X_k$ son variables discretas, la condición*

$$P(X_1 = x_1, \ldots, X_k = x_k) = \prod_{i=1}^{k} P(X_i = x_i), \text{ para todo } x_i,\ i = 1, \ldots, k, \tag{2.8.3}$$

es necesaria y suficiente para la independencia de las variables $X_1, \ldots, X_k$.

Definición 2.8.2 Las particiones generadas por ciertas variables son independientes si estas variables lo son.

Definición 2.8.3 Considere la partición generada por el suceso A_i, esto es, (A_i, A_i') Entonces, los sucesos $A_1, \ldots, A_k$ son independientes si las particiones generadas por estos procesos lo son.

El supuesto de independencia es muy fuerte y, a la vez, difícil de verificar. Sin embargo, resulta muy atractivo su uso, al menos inicialmente, o para disminuir la complejidad de los modelos. Por ejemplo, la independencia de los lanzamientos de tres dados permite especificar 15 probabilidades en vez de 215. En efecto, si

P(1^{er} dado muestra i) = α_i
P(2° dado muestra j) = β_j
P(3^{er} dado muestra k) = γ_k,

entonces $P(X_1 = i, X_2 = j, X_3 = k) = \alpha_i \beta_j \gamma_k$. Si los dados son parecidos, o si en vez de tres dados se trata de tres lanzamientos del mismo dada, $\alpha_i = \beta_i = \gamma_i$ y sólo se requiere asignar 5 números.

La independencia simplifica enormemente la obtención de la función de probabilidad conjunta. Por ejemplo, considere n monedas cargadas y codifique los resultados usando $x_i = 1$ para Cara y $x_i = 0$ para Sello. El resultado $\boldsymbol{x} = (x_1, \ldots, x_n)$ está contenido en $\mathcal{X} = \{0, 1\}^n$, cuya cardinalidad es n. Denotemos por p_i la probabilidad que la i–ésima moneda sea Cara y por q_i la probabilidad que ella sea Sello. Por supuesto $p_i + q_i = 1$ para $i = 1, \ldots, n$. El supuesto de independencia implica que p_i, $i = 1, \ldots, n$, determinan la función probabilidad. Además, es muy sencillo escribir la probabilidad de cualquier resultado. Por ejemplo, $P(\{(1, 1, 0, 0)\}) = p_1 p_2 q_3 q_4$. De acá se obtiene, mediante una suma, la probabilidad de cualquier resultado. Por ejemplo, la probabilidad de obtener exactamente 1 cara al lanzar las dos primeras monedas es la probabilidad del suceso $\{(1, 0), (0, 1)\}$, cuyos elementos tienen probabilidades $p_1 q_2$ y $q_1 p_2$. La probabilidad buscada es $p_1 q_2 + q_1 p_2$.

El ahorro de números es espectacular si las monedas son homogéneas, o sea, $p_1 = \cdots = p_n = p$. Basta el número $0 < p < 1$ para determinar las probabilidades de todos los resultados (para $n = 20$ ya hay más de un millón de éstos). La probabilidad de obtener exactamente 1 cara se reduce ahora a $2p(1 - p)$.

2.8.3. Resultados adicionales para dos variables

Para dos variables X e Y, la condición (2.8.3) se reduce a

$$P(X = x, Y = y) = P(X = x)P(Y = y) \text{ para todo } x, y. \tag{2.8.4}$$

Cuando x e y tienen dos valores cada uno, digamos 1 y 2, se tiene la situación especial en que $A_2 = A_1'$ y $B_2 = B_1'$. Escribiendo $A_1 = A$ y $A_2 = B$ se obtiene que las cuatro condiciones (2.8.4) son

$$\begin{array}{rcl} P(A \cap B) & = & P(A)P(B) \\ P(A \cap B') & = & P(A)P(B') \\ P(A' \cap B) & = & P(A')P(B) \\ P(A' \cap B') & = & P(A')P(B'). \end{array} \tag{2.8.5}$$

Por la Definición 2.3.1, estas condiciones equivalen a

$$\begin{array}{l} A \text{ y } B \text{ son independientes.} \\ A \text{ y } B' \text{ son independientes.} \\ A' \text{ y } B \text{ son independientes.} \\ A' \text{ y } B' \text{ son independientes.} \end{array} \tag{2.8.6}$$

En términos de la tabla de probabilidades conjuntas, la independencia equivale a que la probabilidad de una celda es el producto de los valores marginales, es decir, que sea una *tabla de multiplicación*. Para dos particiones (A, A') y (B, B'), una tabla general (sin imponer independencia) es

	B	B'	
A	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
A'	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
	$P(B)$	$P(B')$	1

En general hay 3 probabilidades conjuntas que se pueden elegir de manera arbitraria, sujeto sólo a la positividad y la suma igual a 1. Para márgenes fijos, cualquier probabilidad conjunta determina

todas las demás. Escribiendo $P(A) = a$, $P(B) = b$ y $P(A \cap B) = c$, la tabla general es

A	c			a
A'				$1-a$
	b	$1-b$		1

Si $c = ab$, un sencillo cálculo algebraico permite completar la tabla, obteniendo la tabla de multiplicación

A	ab	$a(1-b)$		a
A'	$(1-a)b$			$(1-a)(1-b)$
	b	$1-b$		1

Esto muestra que la independencia de A y B en (2.8.6) implica la independencia de los otros tres pares de sucesos. Por simetría, es claro que la independencia de cualquier par implica la de los otros tres. Este hecho se puede expresar sucintamente como un teorema:

Teorema 2.8.2 *Las definiciones 2.3.1 y 2.8.3 son equivalentes.*

Generalicemos ahora la equivalencia de (2.3.1), (2.3.2) y (2.3.3) a dos particiones o dos variables.

Teorema 2.8.3 *Las variables discretas X e Y son independientes si*

$$P(Y = y|X = x) \text{ no depende de } x, \tag{2.8.7}$$

o si

$$P(Y = y|X = x) = P(Y = y) \text{ para todo } x \text{ e } y. \tag{2.8.8}$$

Demostración: La condición (2.8.8) es inmediatamente equivalente al Teorema de Factorización. Además ella implica (2.8.7). Por el Teorema de Probabilidades Totales, $P(Y = y)$ es un promedio ponderado de los $P(Y = y|X = x)$. Luego (2.8.7) implica (2.8.8), lo que concluye la demostración.

En términos de árbol, (2.8.7) dice que el número asignado al arco xy depende sólo del nodo de llegada.

Ejemplo 2.8.1 Consideremos las siguientes tablas correspondientes a variables aleatorias independientes X e Y:

x	0	1	2
$P_X(x)$	0,5	0,3	0,2

y	0	1
$P_Y(y)$	0,4	0,6

Entonces la tabla conjunta es

x/y	0	1	$P_X(x)$
0	0,20	0,30	0,5
1	0,12	0,18	0,3
2	0,08	0,12	0,2
$P_Y(y)$	0,4	0,6	1

Las probabilidades condicionales de interés se obtienen como cuocientes entre las probabilidades conjuntas y marginales correspondientes. Por ejemplo

$$P_{Y|X}(1|2) = \frac{0{,}12}{0{,}2} = 0{,}6$$
$$P_{X|Y}(2|1) = \frac{0{,}12}{0{,}6} = 0{,}2$$
$$P_{X|Y}(2|0) = \frac{0{,}08}{0{,}4} = 0{,}2$$

Como era de esperar, debido a la independencia, dichas probabilidades condicionales coinciden con las no condicionales.

El siguiente resultado muestra cómo transformaciones de grupos disjuntos de variables aleatorias independientes resulta en variables aleatorias independientes, *sin importar las transformaciones empleadas*. Si $X_1, \ldots, X_5$ representan los resultados de 5 lanzamientos independientes de un dado, el siguiente teorema justifica algunas aseveraciones intuitivas hechas sobre la independencia de ciertos sucesos que dependen de conjuntos disjuntos de dados.

Teorema 2.8.4 Sean $X_1, \ldots, X_m, X_{m+1}, \ldots, X_{m+n}$ independientes, y defina las variables aleatorias Y, Z mediante

$$Y = g(X_1, \ldots, X_m),$$
$$Z = h(X_{m+1}, \ldots, X_{m+n}),$$

donde g y h son funciones de m y n argumentos respectivamente. Entonces Y y Z son también independientes.

Para concluir la sección, enunciamos, sin demostración, una caracterización alternativa de independencia de sucesos. Ella es la más popular en los textos de probabilidad, pero tiene la desventaja de no extenderse naturalmente a las variables aleatorias, que es el más usado en las aplicaciones usuales.

Teorema 2.8.5 *Sea $M = \{1, \ldots, k\}$. Los conjuntos $(A_1, \ldots, A_k)$ son independientes, según la Definición 2.8.3, si y sólo si se cumplen las siguientes igualdades*

$$P\left(\bigcap_{i\in E} A_i\right) = \prod_{i\in E} P\left(A_i\right), \text{ para todo } E \subseteq M, \text{con } \operatorname{card} E > 1. \tag{2.8.9}$$

2.9. Aplicaciones de independencia

2.9.1. Demostración de equiprobabilidad

Definición 2.9.1 Una variable aleatoria tiene distribución de probabilidad uniforme sobre el conjunto finito Ω si su función de probabilidad es constante. Se dice también que X se distribuye uniformemente sobre Ω.

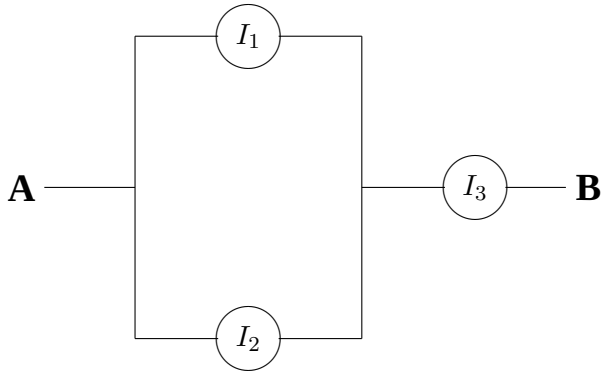
Teorema 2.9.1 *Si un experimento secuencial tiene etapas independientes y los resultados de cada etapa son equiprobables, entonces los resultados son equiprobables. En otras palabras, si X_i tiene distribución uniforme sobre Ω_i y $X_1, X_2, \ldots, X_k$ son independientes, entonces $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ se distribuye uniformemente sobre $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_k$.*

Demostración: Inmediata por el Teorema 2.8.1 (teorema de factorización).

El modelo de urna, con extracciones ordenadas sin reposición, es un caso particular con $n_i = m$ para todo i. El caso de un dado equilibrado corresponde a $m = 6$ y el de una moneda equilibrada a $m = 2$.

2.9.2. Aplicación a confiabilidad

Ejemplo 2.9.1 En el circuito que se indica en el diagrama siguiente



interesa calcular la probabilidad

$$\pi = P(\text{pasa corriente entre A y B}).$$

Este suceso depende del estado de los tres interruptores. Suponiendo independencia entre los interruptores, basta especificar la probabilidad p_i que el interruptor I_i deje pasar la corriente. El resto es un simple cálculo algebraico. Sea $X_i = 1$ si el interruptor I_i deja pasar la corriente y $X_i = 0$ en caso contrario. El resultado del experimento puede tomarse como (x_1, x_2, x_3) y el comportamiento probabilístico equivale a lanzar 3 monedas cargadas y anotar 1 o 0 según salga cara o sello. La lista de resultados y sus correspondientes probabilidades se indica a continuación. La presencia del signo $\surd$ en la última columna indica que la fila correspondiente identifica un resultado favorable con respecto al suceso de interés, es decir que pasa corriente entre A y B.

X_1	X_2	X_3	Probabilidad	
0	0	0	$q_1q_2q_3$	
0	0	1	$q_1q_2p_3$	
0	1	0	$q_1p_2q_3$	
0	1	1	$q_1p_2p_3$	$\surd$
1	0	0	$p_1q_2q_3$	
1	0	1	$p_1q_2p_3$	$\surd$
1	1	0	$p_1p_2q_3$	
1	1	1	$p_1p_2p_3$	$\surd$

La probabilidad buscada se obtiene sumando todas las filas marcadas por $\surd$:

$$\pi = q_1p_2p_3 + p_1q_2p_3 + p_1p_2p_3 = \alpha p_3,$$

con $\alpha = (q_1p_2 + p_1p_2 + q_1q_2)$. Pero α debe coincidir con la probabilidad que el subsistema formado por los dos primeros interruptores deje pasar la corriente. Por otra parte, la única forma que no pase corriente es que ni I_1 ni I_2 dejen que esto ocurra, lo que por independencia tiene probabilidad q_2q_3. Finalmente $\alpha = 1 - q_2q_3$, lo que se puede verificar algebráicamente a partir de la identidad $(p_1 + q_1)(p_2 + q_2) = 1$.

2.9.3. Aplicación a simulación

2.9.3.1. Tablas de números aleatorios

Un espacio muestral Ω de cardinalidad N está en correspondencia biunívoca con $\{1, 2, \ldots, N\}$ y con $\{0, 1, 2, \ldots, N-1\}$. Cuando los elementos de Ω son equiprobables, la probabilidad de cualquier suceso es una fracción con denominador N. Por conveniencia práctica, $N = 10^r$ es el caso más común debido a que el sistema numérico decimal tiene base 10.

Sea U una variable aleatoria con función de probabilidad constante sobre Ω. Físicamente, U es representable por la ficha extraida de una urna con N fichas. El muestreo con reposición desde tal urna genera una sucesión de variables $U_1, U_2, \ldots$ independientes e idénticamente distribuidas (i.i.d.), o sea, U_i y U_j tienen la misma distribución, para todo $i \neq j$. Cuando Ω es un conjunto numérico, decimos que los U_i son números aleatorios. Existen tablas con realizaciones de esta sucesión para $U_1, U_2, \ldots, U_M$, donde M es un número grande. La mayoría de estas tablas considera $\Omega = \{0, 1, 2, \ldots, 9\}$, agrupando los números de a 5. Esto facilita la lectura y tiene una ventaja adicional que explicamos a continuación.

Si interpretamos al arreglo de 5 números como un número de 5 cifras, o sea $Y_1 = 10^4U_1 + 10^3U_2 + 10^2U_3 + 10U_4 + U_5$, se verifica que Y_1 tiene una distribución uniforme entre 0 y 99999. En efecto, el Teorema 2.9.1 muestra que los valores $\boldsymbol{u} = (u_1, u_2, u_3, u_4, u_5)$ del vector aleatorio $(U_1, U_2, U_3, U_4, U_5)$ son equiprobables y $\boldsymbol{u}$ está en correspondencia biunívoca con $y_1 = 10^4u_1 + 10^3u_2 + 10^2u_3 + 10u_4 + u_5$. Anotando $Y_{t+1} = 10^4U_{5t+1} + 10^3U_{5t+2} + 10^2U_{5t+3} + 10U_{5t+4} + U_{5t+5}$, $t = 0, 1, 2, \ldots$ se obtiene una sucesión $Y_1, Y_2, \ldots$ de variables aleatorias uniformemente distribuidas entre 0 y 99999.

La independencia de los U_i y el Teorema 2.8.4 implican que $Y_1, Y_2, \ldots$ son i.i.d. Por otra parte, $Z_i = 10^{-5}Y_i$ tiene resultados equiprobables $0{,}00000, 0{,}00001, \ldots, 0{,}99998, 0{,}99999$. Esto difiere

muy poco de la elección de un punto al azar en un segmento recto de largo 1, el que se modela por una variable continua V con valores en $[0,1]$. Si se dispone de V y se trunca el número a 5 decimales se obtiene U y $|U-V| \leq 0{,}00001$.

Este procedimiento de agrupación de cifras permite usar una urna con 10 fichas en vez de una urna con 10^5 fichas, un ahorro substancial. Si la urna tiene fichas numeradas de 1 hasta N, y ellas se agrupan en arreglos de r fichas cada uno ($N=9$, $r=5$ en el caso recién analizado), se obtiene una sucesión i.i.d. de variables aleatorias uniformemente distribuidas sobre $\{0,1,2,\ldots,N^r-1\}$. El caso $N=2$ es especialmente importante a nivel computacional. Además, en este caso, $U_1, U_2, \ldots$ pueden generarse físicamente lanzando una moneda.

2.9.3.2. Simulación de variables i.i.d.

En la Sección 1.3.3 vimos cómo simular cualquier distribución de probabilidad finita. En el lenguaje de variables aleatorias, se dispone de U con función de probabilidad constante sobre el conjunto Ω de cardinalidad N. Físicamente, U es representable por la ficha extraida de una urna de N fichas y se genera la variable aleatoria X mediante $X=g(U)$. La función g se define identificando $\{u/g(u)=x\}$ con el conjunto de fichas para las que $X=x$. El muestreo con reposición genera las sucesiones de variables independientes $U_1, U_2, \ldots$ y $X_1, X_2, \ldots$. La variable X_i se obtiene de U_i por el mismo procedimiento usado para generar X a partir de U, es decir, $X_i = g(U_i)$. La independencia de los U_i y el Teorema 2.8.4 implican que $X_1, X_2, \ldots$ son independientes. Como U_i y U_j tienen la misma distribución, lo propio ocurre con X_i y X_j, de modo que las variables $X_1, X_2, \ldots$ son i.i.d.

Hemos demostrado así que se puede simular variables aleatorias finitas i.i.d. a partir de números aleatorios o de lanzamientos de una moneda equilibrada.

Ejemplo 2.9.2 Simular una muestra aleatoria de tamaño 200, con reemplazo, de una población subdividida en categorías A, B, C, D, E, F, con las proporciones individuales y acumuladas dadas en la siguiente tabla:

Categoría	Prob. categ.	Prob. acum.	100 × Prob. acum.
A	0.06	0.06	6
B	0.12	0.18	18
C	0.15	0.33	33
D	0.28	0.61	61
E	0.20	0.81	81
F	0.19	1.00	100

Se generan números al azar entre 00 y 99 usando una tabla de números aleatorios y se hace la asignación

1-6	A
7-18	B
19-33	C
34-61	D
62-81	E
82-99, 00	F

Por ejemplo, si los 10 números obtenidos de la tabla fueran 72, 75, 28, 93, 64, 02, 15, 08, 54 y 18, se obtienen las letras que se indica:

72	75	28	93	64	02	15	08	54	18
E	E	C	F	E	A	A	A	D	B

La simulación da una muestra con 3 personas de la categoría A, 1 de la B, 1 de la C, 1 de la D, 3 de la E y 1 de la F.

Ejemplo 2.9.3 Hoy en día las tablas de números aleatorios han sido reemplazadas por programas computacionales, que pueden generar miles de números al azar en fracciones de segundo. En vez de números enteros se generan decimales con un cierto número de dígitos. Si $(T_i,\ i = 1, 2, \ldots)$ son i.i.d. con distribución uniforme en $\{0, 1, \ldots, N - 1\}$, las variables $U_i = \frac{T_i}{N}$ son uniformes en el conjunto $\{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$. Si $N = 10^r$, los elementos de este conjunto son los números reales $0 \leq u_i < 1$, redondeados hacia abajo con sólo r dígitos. La variable U_i satisface la igualdad $P(a \leq U \leq b) = b - a$, con un excelente grado de aproximación. Los computadores y calculadoras científicas generan una sucesión $U_1, U_2, \ldots$ de variables i.i.d. con distribución uniforme en $[0, 1]$, lo que significa que satisfacen $P(a \leq U_i \leq b) = b - a$. El arte de la simulación consiste en simular sistemas probabilísticos más complejos usando un generador de variables uniformes.

A modo de ejemplo, para simular la muestra aleatoria del Ejemplo 2.9.2 usando un generador de uniformes, una posible regla es:

$X_i = A$	si	$0 \leq U_i \leq 0{,}06$
$X_i = B$	si	$0{,}06 < U_i \leq 0{,}18$
$X_I = C$	si	$0{,}18 < U_i \leq 0{,}33$
$X_i = D$	si	$0{,}33 < U_i \leq 0{,}61$
$X_i = E$	si	$0{,}61 < U_i \leq 0{,}81$
$X_i = F$	si	$0{,}81 < U_i \leq 1{,}00$

2.10. Problemas

1. Un dado se lanza dos veces, independientemente. Dado que los resultados de ambos lanzamientos fueron distintos, calcule la probabilidad condicional que

   (a) al menos uno de los números fue 6.

   (b) la suma de los números es 8.

2. En una pregunta con alternativas, la probabilidad que un alumno sepa la respuesta es p. Habiendo m alternativas, si el alumno sabe la respuesta, responde correctamente con probabilidad 1; en caso contrario, el alumno escoge una respuesta al azar. Dado que el alumno dio la respuesta correcta, ¿cuál es la probabilidad que él haya sabido la respuesta?

3. Suponga que el número de accidentes en un día de semana cualquiera entre Lunes y Jueves tiene la siguiente función probabilidad: $p(0) = 0{,}7$, $p(1) = 0{,}2$, $p(2) = 0{,}1$. Análogamente, de Viernes a Domingo estas probabilidades cambian a $p(0) = 0{,}5$, $p(1) = 0{,}3$, $p(2) = 0{,}2$. Suponga que el número de accidentes en días distintos son independientes.

   (a) Describa el espacio muestral adecuado para el problema y utilice la hipótesis de independencia para asignar la probabilidad de cada punto del espacio muestral.

   (b) Calcule la probabilidad que el número total de accidentes en una semana sea (i) Igual a 2. (ii) Al menos 2.

4. Un modelo probabilístico muy simple para estudiar el tiempo atmosférico clasifica cada día como *seco* o *húmedo*. Se supone luego que el tiempo de ma nana será igual al de hoy con probabilidad $0{,}8$. Sabiendo que el día 15 de Mayo fue seco:

   (a) Asigne las probabilidades a cada uno de los 8 escenarios posibles para el tiempo en los próximos 3 días.

   (b) Calcule la probabilidad que el segundo día sea seco.

   (c) Calcule la probabilidad que exactamente dos días sean secos.

5. Dos deportistas disparan sucesivamente a un blanco. Las probabilidades de acertar en el primer disparo son 0.4 y 0.5 respectivamente. Estas probabilidades se incrementan en 0.05 para cada uno, en los disparos sucesivos. ¿Cuál es la probabilidad que el primer disparo haya sido efectuado por el primer deportista dado que el blanco fue acertado en el quinto disparo?.

6. Considere una urna que contiene doce fichas de las cuales ocho son blancas. Una muestra de cuatro fichas es elegida sin reemplazo.

   (a) Calcule la probabilidad que la primera y la tercera ficha extraidas sean blancas.

   (b) Calcule la probabilidad que exactamente tres de las fichas sean blancas.

   (c) ¿Cuál es la probabilidad condicional que la primera y la tercera ficha extraidas sean blancas, dado que la muestra contenía exactamente tres fichas blancas?.

   (d) Repita lo anterior suponiendo que después de cada extracción la ficha se restituye a la urna.

7. Tres cajas A, B y C contienen instrumentos nacionales (N) e importados (I). La composición de A, B y C es 2N y 4I, 8N y 4I, y 1N y 3I respectivamente. Se selecciona al azar un instrumento de una caja elegida al azar.

   (a) ¿Cuál es la probabilidad de obtener un instrumento nacional?.

   (b) Si el instrumento seleccionado es nacional, calcule la probabilidad que provenga de la caja A.

8. Con las mismas cajas del Problema 7, suponga que se selecciona un instrumento al azar de cada una de las cajas y que exactamente dos de ellos resultan ser nacionales. ¿Cuál es la probabilidad que éste provenga de la caja A?.

9. Una compa nía de seguros clasifica a las personas en una de tres categorías : bajo riesgo, riesgo medio y alto riesgo. Sus registros indican que la probabilidad que las personas tengan un accidente durante el a no son 0.05, 0.15, 0.30, respectivamente. Si el 20 % de la población es de bajo riesgo, el 50 % de riesgo medio, y el 30 % de alto riesgo, ¿cuál es la proporción de personas que tienen accidentes en un a no fijo?. Si la póliza tomada por A no tuvo accidentes en 1992, ¿cuál es la probabilidad que esta persona haya sido de bajo riesgo en ese a no?.

10. Suponga que un dado se lanza una vez. Si N es el resultado del lanzamiento, entonces $P(N = i) = p_i$, $i = 1, 2, 3, 4, 5, 6$. Si $N = i$ una moneda equilibrada se lanza i veces. Encontrar la probabilidad condicional que N sea impar dado que se obtuvo al menos una cara.

    $Resp: \frac{\frac{1}{2}p_1+\frac{7}{8}p_3+\frac{31}{32}p_5}{\frac{1}{2}p_1+\frac{3}{4}p_2+\frac{7}{8}p_3+\frac{15}{16}p_4+\frac{31}{32}p_5+\frac{63}{64}p_6}$.

11. Suponga que lanzamos una moneda n veces con probabilidad p de obtener una cara y q de obtener un sello en cada lanzamiento. Suponga además que todos los lanzamientos son independientes. Sea S_n la variable aleatoria que cuenta el número de caras obtenidas en los n lanzamientos. Encuentre $P(S_n \geq 3 | S_n \geq 1)$.

    $Resp: \frac{1-q^n-npq^{n-1}-\frac{1}{2}n(n-1)p^2q^{n-2}}{1-q^n}$.

12. Suponga que un dado equilibrado se lanza una vez. Si sale un número impar, una moneda honesta se lanza repetidamente; si sale un número par una moneda sesgada con probabilidad de obtener cara $p \neq \frac{1}{2}$ se lanza repetidamente (los lanzamientos de la moneda son independientes en cada caso). Si los n primeros resultados son caras, ¿cuál es la probabilidad que una moneda insesgada haya sido usada?.

    $Resp: \frac{\frac{1}{2^{n+1}}}{\frac{1}{2^{n+1}}+\frac{1}{2}p^n}$.

13. Suponga se tiene una urna con bolitas blancas y negras, sumando un total de n bolitas, y se extraen bolitas con reemplazo de dicha urna. Si se hacen k extracciones y se observan k bolitas blancas, ¿cuál es la probabilidad que la urna tenga sólo bolitas blancas?

14. La probabilidad que un pan de pascua contenga exactamente k pasas está dada por $p_k = \frac{\lambda^k e^{-\lambda}}{k!}$, $k = 0, 1, 2, \ldots$. Suponga que por cada pasa que contiene este pan de pascua, una moneda con probabilidad de cara p se lanza. Si sale cara, Ud. se come la pasa, y si sale sello, Ud. la guarda para algún amigo. ¿Cuál es la probabilidad que Ud. coma exactamente n pasas?. ¿Cuál es el rango de valores para n?.

15. Sea Q_n la probabilidad que en n lanzamientos de una moneda regular no aparezcan corridas de tres caras sucesivas. Demuestre que:

$$Q_n = \frac{1}{2}Q_{n-1} + \frac{1}{4}Q_{n-2} + \frac{1}{8}Q_{n-3},$$

sabiendo que $Q_0 = Q_1 = Q_2 = 1$. Encontrar Q_8.

16. Como un modelo simplificado para predecir el tiempo atmosférico, se supone que el tiempo (seco o húmedo) para ma nana será igual al de hoy con probabilidad p. Si el tiempo atmosférico el primero de Enero fue seco, demostrar que P_n, la probabilidad que llueva n días después, satisface la relación

$$P_n = (2p-1)P_{n-1} + (1-p), \qquad n \geq 1,$$

con $P_0 = 1$. Demuestre además que

$$P_n = \frac{1}{2} + \frac{1}{2}(2p-1)^n, \qquad n \geq 0.$$

17. Una jaula A contiene cinco aves blancas y siete aves negras. La jaula B contiene tres blancas y doce negras. Se lanza una moneda al aire. Si el resultado es cara, entonces un ave de A es seleccionada, mientras que si el resultado es sello, se selecciona un ave de la jaula B. Suponga que el ave seleccionada es blanca. ¿Cuál es la probabilidad que la moneda haya mostrado cara?.

$Resp: \frac{12}{37}$

18. Una urna contiene N fichas negras y A fichas azules. Se selecciona una ficha al azar, y se la devuelve a la urna, junto con C fichas adicionales del mismo color. Se selecciona ahora una segunda ficha al azar. Demuestre que la probabilidad que la primera ficha era negra, dado que la segunda fue azul es $\frac{N}{(N+A+C)}$.

19. Hay tres monedas en una caja. Una de ellas tiene dos caras, la otra es normal, y la tercera muestra cara con probabilidad 75 %. Dado que cuando se elige una de las tres monedas al azar y se lanza el resultado es cara, calcule la probabilidad que ésta sea la moneda de dos caras.

$Resp: \frac{4}{9}$.

20. Dos bolas se eligen aleatoriamente desde una urna que contiene ocho blancas, cuatro negras y dos amarillas. Suponga que ganamos \$2 por cada bola negra seleccionada, perdemos \$1 por cada bola blanca seleccionada, y que no hay cambios si se selecciona una bola amarilla. Determine los posibles valores que se pueden obtener, y calcule las probabilidades correspondientes.

$$Resp: \begin{array}{c|ccccccc} k & 4 & 2 & 1 & 0 & -1 & -2 \\ \hline P(k) & \frac{6}{91} & \frac{8}{91} & \frac{32}{91} & \frac{1}{91} & \frac{16}{91} & \frac{28}{91} \end{array}$$

21. Un comprador de transistores adquiere éstos en lotes de 20, y es su política inspeccionar cuatro transistores elegidos aleatoriamente desde un lote y aceptar el lote solamente si los cuatro están buenos. Si cada componente de un lote es, independientemente, defectuosa con probabilidad 0.1, ¿cuál es la proporción de lotes rechazados?.

    $Resp$: 0,3439

Desafíos

22. Una maleta contiene a esferas blancas y b negras. Las esferas se eligen de la maleta de acuerdo a la siguiente regla:

    a.- Una esfera se elige al azar y se elimina.

    b.- Una segunda esfera se elige a continuación. Si su color es distinto al de la primera, ésta es sustituida en la maleta y se repite el proceso del comienzo. Si el color es igual al de la primera, la esfera se elimina y se comienza desde el punto b.

    En otras palabras, las esferas son muestreadas y eliminadas hasta que ocurre un cambio de color, en tal caso la última esfera es devuelta a la maleta, y el proceso comienza de nuevo. Denote por P_{ab} la probabilidad que la última esfera en la maleta sea blanca. Demostrar que:

    $$P_{ab} = \frac{1}{2}$$

    Hint: Use inducción sobre $k \equiv a + b$.

23. Un dado A tiene cuatro caras rojas y dos caras blancas, por otra parte un dado B tiene dos caras rojas y cuatro caras blancas. Una moneda es lanzada una vez. Si el resultado es cara, el juego continua con el dado A; si es sello, el dado B es usado.

    a.- Demuestre que la probabilidad que salga una cara roja es $\frac{1}{2}$.

    b.- Si en los primeros dos lanzamientos aparece la cara de color rojo, ¿cuál es la probabilidad que en el tercer lanzamiento la cara sea roja?.

    c.- Si el rojo aparece en los dos primeros lanzamientos, ¿cuál es la probabilidad que se haya usado el dado A?.

    $Resp$: $b)\frac{3}{5}; c)\frac{4}{5}$

24. Supóngase que los días son clasificados en “Soleados” y “Nublados”, y que las condiciones del clima en ma nanas sucesivas forman una cadena de Markov con probabilidades de transición estacionarias. Suponiendo que la matriz de transición sea:

	Soleado	Nublado
Soleado	0.7	0.3
Nublado	0.6	0.4

a.- Si un día esta nublado, ¿cuál es la probabilidad que esté nublado al día siguiente?.

b.- Si un día es soleado, ¿cuál es la probabilidad que los dos días que siguen sean soleados?.

c.- Si un día esta nublado, ¿cuál es la probabilidad que al menos uno de los tres días siguientes esté soleado?.

Capítulo 3

Variables Aleatorias

En este capítulo desarrollamos con mayor profundidad algunos temas que ya fueron presentados en los capítulos previos. Así, en la Sección 1.4.3 discutimos el concepto de variable en términos de una población finita, distinguiendo tipos de variables. Por otra parte, una muestra al azar de una población finita transforma las probabilidades de los sucesos en proporciones dentro de la población finita. De esta forma, la construcción y descripción de distribuciones de probabilidad está íntimamente ligada al estudio de poblaciones en esta población; un paso al límite arroja luz sobre las variables continuas. Un subproducto importante del estudio de poblaciones finitas es que permite visualizar concretamente a una variable como una función definida para una población, lo que hace más natural la definición abstracta de variable aleatoria. La primera sección trata la descripción de proporciones para variables discretas y continuas, lo que proporciona una base intuitiva para atacar problemas probabilísticos.

3.1. Descripción de Proporciones en una Población

Continuamos acá el estudio iniciado en la Sección 1.4.3 sobre el concepto de variable en el contexto de una población finita. Examinamos ahora la descripción de poblaciones para distintos tipos de variables. Para ilustrar las ideas continuamos el ejemplo de dicha sección, donde se muestran las 10 primeras líneas de un archivo computacional. Supondremos ahora una población de gran tama no, digamos cien mil personas, de la cual se ha extraido una muestra al azar de 500 personas. Dada la peque na fracción de muestreo, hay poca diferencia entre el muestreo sin y con reposición. Adoptando este último supuesto, para cada columna de la tabla de datos, las 500 componentes pueden ser consideradas como una realización de 500 variables i.i.d., cuya distribución común coincide con la distribución de proporciones en la población. Por razones de espacio, la Tabla 3.1.1 muestra sólo las 100 primeras líneas del archivo de datos, pero algunos resultados se obtienen sobre la base de la muestra completa de tama no 500.

Por la manera de generar la información hay simetría entre los individuos, es decir, una reordenación arbitraria de las filas de la tabla no debiera afectar las conclusiones. Por otra parte, el número de filas de la tabla coincide con el tama no de la muestra, el que está sujeto a limitaciones de tiempo y presupuesto. En consecuencia, conviene caracterizar el comportamiento de las variables prescin-

diendo del tama no de la población. *Los promedios aritméticos y las proporciones son resúmenes sencillos que tienen estas características deseables.*

Por descripciones entendemos tanto a números, tablas numéricas o a los gráficos correspondientes, los que varían según el tipo de variable. Los ejemplos que se exhiben a continuación se refieren a los datos de la Tabla 3.1.1. La clasificación de variables se aplica también a las variables aleatorias y las probabilidades se describen de manera análoga a las proporciones.

Identificador	Comuna	Nivel Socio Económico	Tama no Familia	N Consultas Médicas	Sexo	Peso (kg)
1	A	1	3	3	M	74.8
2	A	1	3	2	F	54.2
3	A	1	4	4	M	69.7
4	A	3	4	2	F	58.4
5	C	3	3	8	M	64.6
6	C	4	3	1	F	64.5
7	B	2	3	6	M	72.1
8	A	3	2	2	F	66.0
9	C	3	1	4	M	71.6
12	A	2	2	2	M	72.9
13	A	1	6	5	F	46.3
14	B	2	3	4	F	56.3
15	A	1	6	4	F	52.2
16	B	1	5	4	F	62.0
17	B	5	1	4	F	66.3
18	A	2	3	5	M	77.3
19	B	1	7	9	M	79.4
20	A	1	5	2	M	70.1
21	A	2	4	6	F	63.9
22	A	2	2	3	F	61.5
23	A	1	5	0	F	57.8
24	A	3	1	5	M	69.3
25	A	2	3	5	M	86.3
26	A	2	1	2	M	78.3
27	B	3	1	1	M	73.9
28	A	2	4	5	F	55.0
29	B	2	3	4	F	72.3
30	B	4	1	1	M	76.6
31	A	2	2	4	M	71.0
32	A	1	3	1	F	57.7
33	B	2	2	4	M	71.8
34	A	1	3	2	M	73.7
35	A	2	1	2	M	77.7
36	A	3	2	7	F	58.5
37	C	4	2	3	F	58.9
38	A	1	4	3	F	67.0
39	A	1	6	3	F	57.5
40	C	1	5	7	M	79.9
41	B	3	4	7	M	74.9
42	B	3	1	4	F	54.8
43	C	4	3	1	M	79.7
44	B	2	3	4	F	72.1

continúa en la siguiente página

Identificador	Comuna	Nivel Socio Económico	Tama no Familia	N Consultas Médicas	Sexo	Peso (kg)
45	A	1	3	2	F	50.4
46	C	1	4	4	F	67.0
47	B	1	6	5	M	76.0
48	B	2	5	2	F	64.0
49	C	1	7	1	M	76.6
50	A	1	2	3	F	65.3
51	A	1	2	4	F	64.2
52	A	2	4	2	M	78.6
53	A	1	4	0	F	60.4
54	B	1	4	6	F	57.5
55	C	5	2	1	M	79.6
56	B	1	5	4	F	54.4
57	A	1	5	7	F	58.4
58	A	1	4	7	M	73.7
59	A	1	5	3	M	73.8
60	C	3	2	4	M	75.4
61	A	2	1	4	M	75.0
62	A	2	1	8	F	55.4
63	A	2	2	0	M	71.4
64	B	2	4	3	F	58.2
65	A	1	3	2	M	87.2
66	A	2	1	2	M	72.9
67	A	3	3	7	M	78.3
68	A	1	3	7	M	81.5
69	C	5	1	3	M	83.6
70	B	1	1	1	F	57.9
71	A	1	2	0	F	58.4
72	A	2	5	4	M	70.0
73	A	1	3	6	M	69.6
74	B	5	3	3	F	57.7
75	A	1	5	4	F	56.8
76	C	3	1	2	F	48.1
77	C	5	1	4	F	54.9
78	B	4	1	2	M	79.6
79	B	1	4	2	M	69.5
80	C	3	2	2	F	59.8
81	A	1	4	5	F	67.6
82	B	1	5	6	F	58.2
83	A	1	4	5	F	52.7
84	C	4	2	1	F	68.2
85	A	2	1	2	F	54.3
86	A	1	4	1	F	55.9
87	C	3	2	3	F	62.0
88	A	1	6	6	F	57.9
89	B	4	1	5	F	64.3
90	A	3	2	8	M	71.8
91	B	4	1	7	M	79.6
92	A	2	2	3	F	61.5
93	C	5	1	5	F	52.9

continúa en la siguiente página

Identificador	Comuna	Nivel Socio Económico	Tama no Familia	N Consultas Médicas	Sexo	Peso (kg)
94	B	1	3	3	F	54.4
95	A	1	4	5	F	59.6
96	A	1	5	9	F	59.7
97	A	1	5	2	F	56.4
98	B	1	6	7	M	70.6
99	A	4	2	4	F	54.7
100	A	1	6	4	F	61.6

Cuadro 3.1.1: Variables para subpoblación de 100 individuos

Para una variable categórica, la descripción es obvia. Simplemente se indica la proporción o porcentaje para cada categoría. Para una variable binaria basta la proporción correspondiente a una de las dos categorías. Cuando las categorías están ordenadas se pueden calcular, además, proporciones acumuladas. La representación gráfica depende mucho del ingenio, siendo tradicional los diagramas de barra o de torta, que frecuentemente aparecen en periódicos y revistas.

Una variable discreta se puede tratar como ordinal, siendo tradicional utilizar líneas o barras delgadas, para enfatizar que los valores intermedios carecen de sentido, e.g. 2.5 miembros en una familia. Los gráficos en la Figura 3.1.1 representan al tama no de grupo familiar y número de visitas médicas.

No es conveniente hacer lo mismo con una variable continua X, por la proliferación de barras y el hecho que si x se expresa con muchos decimales, todas las proporciones serán muy peque nas. De hecho, si el valor x no aparece en la tabla, la proporción correspondiente será igual a cero. Esto muestra que los valores individuales no tienen interés directo, y que lo relevante son las proporciones correspondientes a ciertos intervalos. Para resumir la información, conviene tomar una partición o una sucesión creciente de intervalos. En ambos casos se elige una sucesión ordenada de números reales: $-\infty = t_0 < t_1 < \cdots < t_{j-1} < t_j < \cdots < t_{r-1} < t_r = \infty$. La partición generada es $(A_1, \ldots, A_r)$, con $A_j = (t_{j-1}, t_j]$. La sucesión creciente está formada por los conjuntos $B_j = (-\infty, t_j]$, $j = 1, \ldots, r$. Denotando por q_j y Q_j a las proporciones correspondientes a los intervalos A_j y B_j respectivamente, se tienen las relaciones

$$Q_j = Q_{j-1} + q_j, \qquad q_j = Q_j - Q_{j-1}, \qquad Q_j = \sum_{m=1}^{j} q_m.$$

Podemos construir ahora dos gráficos asociados.

- **Proporciones acumuladas.** Se grafican los puntos (t_j, Q_j) para $j = 1, \ldots, r-1$ (uniéndolos opcionalmente por segmentos lineales).

- **Histograma.** Se construye una función constante dentro de cada A_j, de modo que su gráfico tiene forma de escalera. Se elige el valor d_j que toma la función dentro de A_j, como

  $$d_j = c \times \frac{q_j}{t_j - t_{j-1}},$$

  donde c se calcula de tal forma que la proporción q_j coincida con el *área* bajo el pelda no correspondiente. El gráfico de esta función se denomina *histograma.* Se sugiere al lector verificar que el área total bajo la escalera es 1 y que la función cuyo gráfico es el histograma

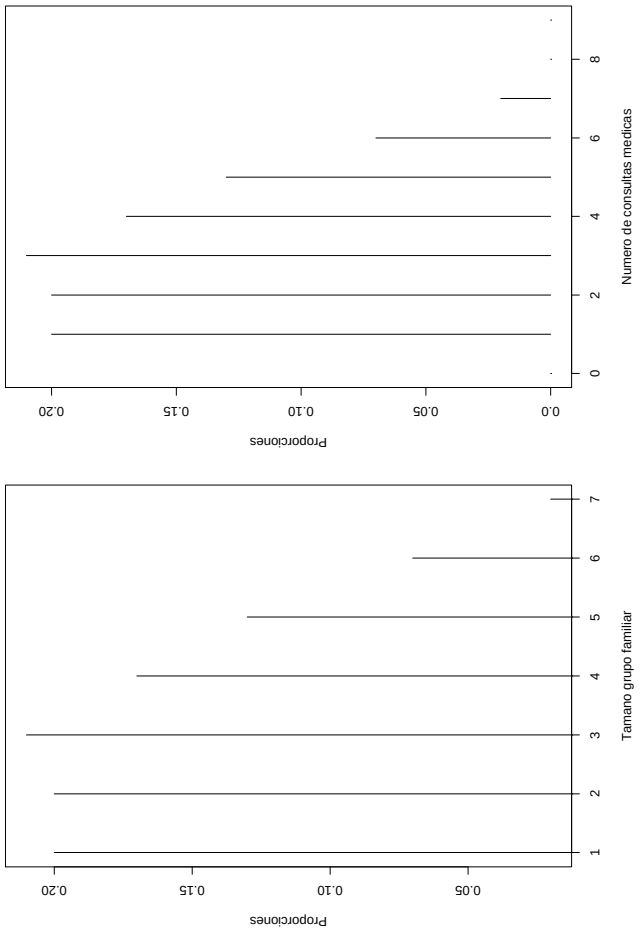


Figura 3.1.1: Tama no de grupo familiar y número de visitas médicas en Tabla 3.1.1.

coincide con la derivada de la función, cuyo gráfico es la poligonal descrita para las proporciones acumuladas.

La Figura 3.1.2 muestra el histograma de la variable peso, construido a partir de la población de 500 individuos, de los cuales la Tabla 3.1.1 muestra a 100 de ellos. Superpuesta al histograma hay una curva suave, que posteriormente vincularemos a la función densidad de probabilidad. Las áreas bajo esta curva también aproximan a las proporciones en un intervalo dado. La Figura 3.1.3 repite lo anterior, separadamente para hombres y mujeres. Invitamos al lector a proponer una explicación para la forma de estos gráficos.

Elijamos a los t_j como los valores distintos que una variable X alcanza en la tabla de datos, ordenados de menor a mayor, y supongamos que la tabla tiene muchas filas. Entonces, los valores consecutivos de la variable estarán muy próximos uno de otro y el gráfico de proporciones acumuladas se aproximará bien por una curva suave, que crece desde 0 hasta 1, a medida que aumenta el valor de x de la variable. Tal curva es el gráfico de cierta función F, y la proporción $\pi(a, b)$ de

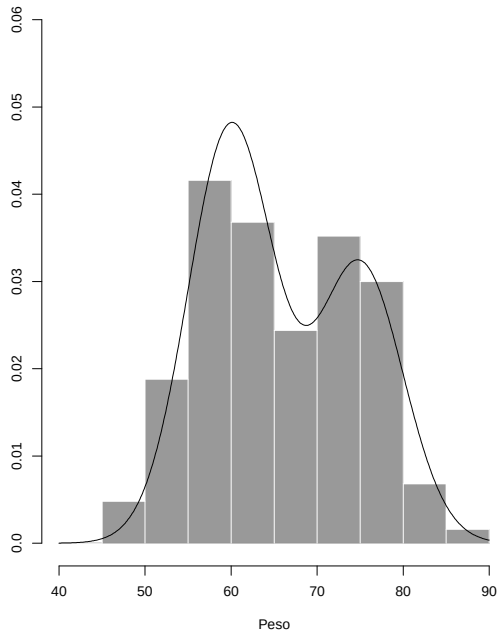


Figura 3.1.2: Histograma de la variable peso en la Tabla 3.1.1.

individuos que satisfacen $a < x \leq b$ se puede aproximar por

$$\pi(a, b) = F(b) - F(a). \tag{3.1.1}$$

Procediendo del mismo modo, el histograma se aproximará bien por el gráfico de cierta función no negativa f, tal que áreas bajo la curva aproximen a las proporciones. Las curvas superpuestas a los histogramas en las Figuras 3.1.2 y 3.1.3 son gráficos de una función f. Analíticamente las áreas bajo una curva son integrales, de modo que

$$\pi(a, b) = \int_a^b f(x)dx. \tag{3.1.2}$$

3.2. Variable Aleatoria y su Distribución de Probabilidad

3.2.1. Variable Aleatoria como Función

Hasta ahora, las variables han aparecido primariamente para ayudar a definir el resultado de un experimento y, por tanto, en la elección del espacio muestral Ω. Tanto los elementos $\omega \in \Omega$ como los subconjuntos (sucesos) de interés suelen describirse en términos de los valores $x_1, x_2, \ldots, x_n$ de ciertas variables *originales*, a las que denotamos por las letras mayúsculas correspondientes.

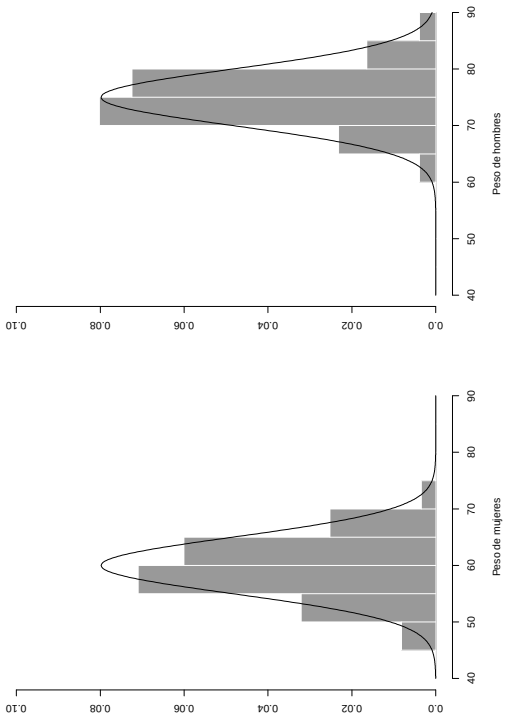


Figura 3.1.3: Histograma de la variable peso, separado por sexo, en la Tabla 3.1.1.

Cuando el resultado ω coincide con el valor x de X, se asignan directamente probabilidades a los elementos del conjunto $\mathcal{X}$ de valores posibles de esta variable. Cuando no es claro cómo calcular las probabilidades $p_X(x) = P(X = x)$, una posible vía de solución es escribir $x = h(\omega)$, donde ω es el resultado de cierto experimento, asignar probabilidades a los subconjuntos de Ω y *deducir* $P(X = x)$. Si Ω es finito o numerable, basta asignar las probabilidades $p(\omega) = P(\{\omega\})$ y obtener $p_X(x)$ como suma de las probabilidades de los casos favorables. En otras palabras,

$$p_X(x) = \sum_{\omega \in \Omega \ / \ h(\omega)=x} p(\omega)$$

En el importante caso de la elección de *un* individuo, al azar, de una población finita de tama no N, lo natural es tomar $\Omega = \{1, 2, \ldots, N\}$, donde ω identifica al individuo seleccionado. Por hipótesis, cada ω tiene probabilidad $\frac{1}{N}$ y de aquí se deduce que $P(X = x)$ coincide con una proporción poblacional. Si imaginamos una tabla de datos para toda la población, cada variable se representa por una columna de esa tabla o por una función que le asigna a ω el valor de la variable para el individuo con identificador ω.

Cuando las variables de interés son $X_1, X_2, \ldots, X_k$, es natural elegir $\omega = (x_1, x_2, \ldots, x_k)$. Con esta elección, la variable X_i corresponde a la función que asigna a cada arreglo de largo k su i-ésima componente. Por otra parte, el valor y de cualquier variable de interés Y debe estar determinada por ω, es decir, debe existir una función g para la cual $y = g(x_1, x_2, \ldots, x_n)$. En este esquema, denominamos a las X_i variables *originales o primarias*, mientras que a Y la denominamos

variable *derivada o secundaria* y la denotamos por $Y = g(X_1, X_2, \ldots, X_n)$. Cuando las variables X_i son discretas, i.e. el número de valores posibles es finito o numerable, se tiene

- El espacio muestral Ω es numerable.
- Toda variable derivada es discreta.
- Denotando por $p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_k)$ a $P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$, la función de probabilidad $p_Y(y)$ se obtiene mediante

$$p_Y(y) = \sum_{\boldsymbol{x} \ / \ g(\boldsymbol{x})=y} p(\boldsymbol{x}).$$

Si se toma a Ω como el conjunto de los arreglos $\boldsymbol{x}$ posibles, la función g tiene dominio Ω.

Motivados por la discusión anterior, entregamos una definición abstracta de variable aleatoria.

Definición 3.2.1 Una *variable aleatoria* es una función definida sobre el espacio muestral Ω, con valores en el conjunto $\mathcal{X}$.

Notación: Si el valor de la variable aleatoria se denota por una letra minúscula, la variable se denota por la letra mayúscula correspondiente. Normalmente se utilizan las últimas letras del alfabeto. Con esta convención escribimos $x = X(\omega)$. El suceso: "el valor x de X pertenece al conjunto B se denota por $X \in B$. Cuando $B = \{x\}$ se simplifica la notación a $\{X = x\}$ o $X = x$. Hacemos notar que utilizamos la misma letra para denotar un valor x incierto (antes de conocer el resultado del experimento). Una vez conocido el resultado ω, el valor de la variable es $x = X(\omega)$, donde usamos la letra X para representar una función. El suceso correspondiente a que el valor x de la variable X satisfaga un conjunto de condiciones se escribe reemplazando x por X. Por ejemplo, $P(X^2 - 5X + 6 \leq 0)$ es la probabilidad que el valor x de la variable X satisfaga $x^2 - 5x + 6 \leq 0$, o sea, $2 \leq x \leq 3$. Así $P(X^2 - 5X + 6 \leq 0) = P(2 \leq X \leq 3)$.

Definición 3.2.2 Dada una variable aleatoria X definida sobre Ω, con valores en $\mathcal{X}$, la distribución de probabilidad *inducida* por X sobre $\mathcal{X}$ se define, para un evento $B \subset \mathcal{X}$ como:

$$P_X(B) = P(X^{-1}(B)), \tag{3.2.1}$$

donde $X^{-1}(B) = \{\omega \in \Omega / \ X(\omega) \in B\}$ es un evento en Ω, y P es la distribución de probabilidad definida sobre Ω.

Se denomina también a P_X *distribución de probabilidad* o *distribución* de la variable aleatoria X.

Con la convención notacional adoptada $P_X(B) = P(X \in B)$, es decir, es la probabilidad que el valor de X esté contenido en B. Conocer la distribución de probabilidad de una variable aleatoria equivale a conocer la probabilidad que el valor de X esté contenido en B, para todo suceso B. La relación básica es que el suceso B en el espacio muestral $\mathcal{X}$ ocurre si y sólo si ocurre el suceso $X^{-1}(B))$ en el espacio muestral Ω. Este último corresponde a la ocurrencia de un resultado $\omega \in \Omega$, tal que $X(\omega) \in B$.

En general, especificar P_X directamente es una tarea difícil. En el Capítulo 1 vimos que en el caso particular que $\mathcal{X}$ es un conjunto finito o numerable, las probabilidades quedan determinadas por la función de probabilidad p_X, definida por

$$p_X(x) = P_X(\{x\}) = P(\{\omega \in \Omega / \ X(\omega) = x).$$

La función p_X debe ser no negativa y la suma de sus valores ser igual a 1.

Ejemplo 3.2.1 Considere 10 lanzamientos sucesivos de una moneda (en forma independiente). Si esto es todo lo que sabemos, lo más natural es escribir el resultado como la 10-tupla $(C, C, S, C, S, S, S, C, C, S)$ o similar, lo que equivale a elegir como espacio muestral a $\Omega = \{C, S\}^{10}$, que contiene $2^{10} = 1024$ elementos. El número de sucesos, es decir, el número de subconjuntos de Ω, asciende a la escalofriante cifra de 2^{1024}. Afortunadamente, la probabilidad de cualquiera de ellos es calculable si conocemos las probabilidades de los 1024 sucesos elementales $\{\omega\}$ y no todos los sucesos son de interés. Típicamente, aquellos de interés se pueden expresar en términos del valor x de alguna variable. Las preguntas más habituales se relacionan con el número de caras o sellos obtenidos. Por ejemplo: *El número de caras es superior al numero de sellos* o *El número de sellos es superior a 7* se expresan en términos de la variable X : *número total de caras*, por $X > 5$ o $X \leq 2$ respectivamente. En estas circunstancias, parece atractivo utilizar $\mathcal{X} = \{1, 2, \ldots, 10\}$ como un espacio muestral alternativo a Ω, dada su menor complejidad. Sin embargo, resulta poco claro como asignar probabilidades a los valores $x \in \mathcal{X}$, mientras que la probabilidad de cada ω es más fácil de obtener.

Como el valor de x está determinado por el de ω, debe existir una función h, tal que $x = h(\omega)$. Así, por ejemplo, tenemos que $h(C, C, S, C, S, S, S, C, C, S) = 5$, $h(S, S, S, S, S, S, S, S, S, S) = 0$, $h(S, C, C, C, S, C, S, C, S, S) = 5$, etc. La definición abstracta identifica a la variable X con esta función. Un peque no cambio de notación simplifica la escritura. Denotemos por x_i a la i -ésima componente del arreglo $\omega \in \Omega$ y consideremos x_i como valor de una variable X_i, que toma el valor C si aparece cara en el i-ésimo lanzamiento, y S si sale sello. $X_1, X_2, \ldots, X_{10}$ son las variables originales y sus valores determinan ω. De esta forma $X = g(X_1, X_2, \ldots, X_{10})$. Podemos definir otras variables derivadas:

- La variable Y_i, que asigna el valor $y_i = 1$, cuando la i-ésima moneda sale cara y sello en caso contrario. En este caso, $y_1, y_2, \ldots, y_{10}$ determinan, a su vez, ω.
- Z = número total de sellos. Para cada ω, los valores de x y z satisfacen $x+z = 10$, de modo que $Z = g(X) = 10 - X$.
- La variable X se puede escribir más fácilmente en función de los Y_i que de los X_i. En efecto $X = \sum_{i=1}^{10} Y_i$.

Supongamos que la moneda tiene probabilidad p de salir cara, y $q = 1-p$ de salir sello. El supuesto de independencia entre los lanzamientos de la moneda implica $p(\omega) =$

$P(\{\omega\}) = p^x q^{10-x}$, donde $x = h(\omega) = X(\omega)$ es el número de caras. De aquí,

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= \sum_{\omega \in \Omega /\ h(\omega)=x} p(\omega) \\ &= \binom{10}{x} p^x q^{10-x}, \end{aligned}$$

donde $x \in \{0, 1, \ldots, 10\}$, el conjunto imagen de X.

Ejemplo 3.2.2 Considere el juego de LOTO, y X definido como el número de aciertos en una cartilla seleccionada al azar. El espacio muestral Ω consiste de $\binom{36}{6} = 1{,}947{,}792$ posibles cartillas, mientras que $\mathcal{X}$ está simplemente dado por $\{0, 1, \ldots, 6\}$. Es razonable suponer que los elementos de Ω son todos equiprobables, de modo que el cálculo de la función de probabilidad inducida $p_X(x)$ para $x \in \mathcal{X}$ se reduce a contar casos favorables. Por ejemplo,

$$p_X(3) = \frac{\binom{6}{3} \times \binom{30}{3}}{\binom{36}{6}} = \frac{20 \times 4060}{1{,}947{,}792} = 0{,}0417$$

Más generalmente, y usando idénticos argumentos, se puede concluir que

$$p_X(x) = \frac{\binom{6}{x} \times \binom{30}{6-x}}{\binom{36}{6}}, \quad \text{para } x = 0, 1, \ldots, 6.$$

3.2.2. Conjunto de valores de una variable aleatoria como espacio muestral

En los capítulos previos, el rol básico de las variables es definir el espacio muestral. A menudo, la descripción del problema no aporta información alguna sobre, ya sea las probabilidades sobre el espacio muestral Ω, o las probabilidades inducidas sobre $\mathcal{X}$. Una forma de abordar el problema es, simplemente, desentenderse de Ω, y tomar $\mathcal{X}$ como el espacio muestral. En otras palabras, dada una única variable de interés X, la elección canónica del espacio muestral es $\Omega = \mathcal{X}$. Se identifica entonces la distribución P sobre Ω con la distribución inducida P_X sobre $\mathcal{X}$. Formalmente, esto es un caso particular de la definición general en que X es la función identidad, pero tal punto de vista es bastante inútil. Toda variable aleatoria Y se puede representar por $g(X)$, para cierta función g.

Si un estudio previo nos entrega proporciones empíricas, podemos adoptar a estas frecuencias como aproximaciones de las probabilidades sobre $\mathcal{X}$. Una manera de obtener una distribución de probabilidad consiste en postular una familia paramétrica de probabilidades sobre $\mathcal{X}$, y usar los datos previos para estimar los parámetros y, por tanto, seleccionar a un miembro de esta familia como la distribución buscada. Los procedimientos de estimación forman parte de la *Inferencia Estadística* (no contenidos en este texto), la que constituye una fértil área de aplicación de la teoría de probabilidad.

Si hay n variables de interés, la elección canónica es que Ω sea un conjunto de arreglos $(x_1, \ldots, x_n)$. En este caso, la variable X_i corresponde a la función que asigna a cada arreglo su i-ésima componente y toda variable aleatoria Y se puede escribir como $g(X_1, \ldots, X_n)$, para una función g adecuada. Cuando las variables X_i son discretas, y $\omega = (x_1, \ldots, x_n)$, lo mismo se aplica a cualquier variable aleatoria Y, de modo que P_Y queda determinada por su función de probabilidad p_Y. Si $Y = g(X_1, \ldots, X_n)$, $p_Y(y)$ es la suma de las probabilidades de los $(x_1, \ldots, x_n)$ tales que $g(x_1, \ldots, x_n) = y$.

Ejemplo 3.2.3 Considere un pan de pascua seleccionado al azar, y sea X definido como el número de pasas contenidos en el pan de pascua. El espacio muestral Ω para este caso es el conjunto de todos los posibles panes de pascua que pudimos haber seleccionado inicialmente (esto depende de la población objetivo de panes de pascua). El conjunto $\mathcal{X}$ queda representado por $\{0, 1, 2, \ldots\}$. Una familia paramétrica bastante popular es

$$p_X(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

para $x = 0, 1, 2, \ldots$, y para algún valor de $\lambda > 0$ (es fácil verificar que estos valores son positivos y suman 1). Para usar esta fórmula en la práctica, uno requiere estimar el valor de λ. Veremos más adelante que el contenido promedio de pasas obtenido para un conjunto de panes es una estimación razonable.

Los siguientes ejemplos corresponden a $\mathcal{X}$ no numerable. El primero está relacionado con el Ejemplo 1.5.3.

Ejemplo 3.2.4 Considere un dardo lanzado al azar sobre un tablero circular de radio unitario. Sea X la posición del dardo al hacer impacto con el tablero, como se indica en la Figura 3.2.4.

El espacio muestral natural es acá el disco unitario, cuya representación cartesiana sugiere la elección de $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 \,/\, x_1^2 + x_2^2 \leq 1\}$ como espacio muestral. El vector $\omega = (x_1, x_2)$ es el valor de una variable aleatoria, a la cual se la suele denominar *vector aleatorio*. La no numerabilidad de $\mathcal{X}$ hace imposible asignar probabilidades positivas a todos los puntos. Para evitar asimetrías muy marcadas hay que concluir que la probabilidad de cada punto es cero. Afortunadamente, los sucesos de interés no incluyen conjuntos de un sólo elemento, sino regiones de área positiva. Asignar probabilidades para subconjuntos no numerables de $\mathbb{R}^k$ es tema de otro capítulo. Sin embargo, una traducción adecuada de la idea de lanzamiento al azar, es que todos los subconjuntos de $\mathcal{X}$ de igual área sean equiprobables. De los axiomas de probabilidad se deduce que la probabilidad de un suceso $A \subset \mathcal{X}$ es proporcional al área de A, esto es:

$$P_X(A) = \frac{\text{área}(A)}{\text{área}(\mathcal{X})} = \frac{\text{área}(A)}{\pi} = \int_A \frac{1}{\pi} dx_1 dx_2.$$

Esta última expresión, tendrá una importante interpretación más adelante.

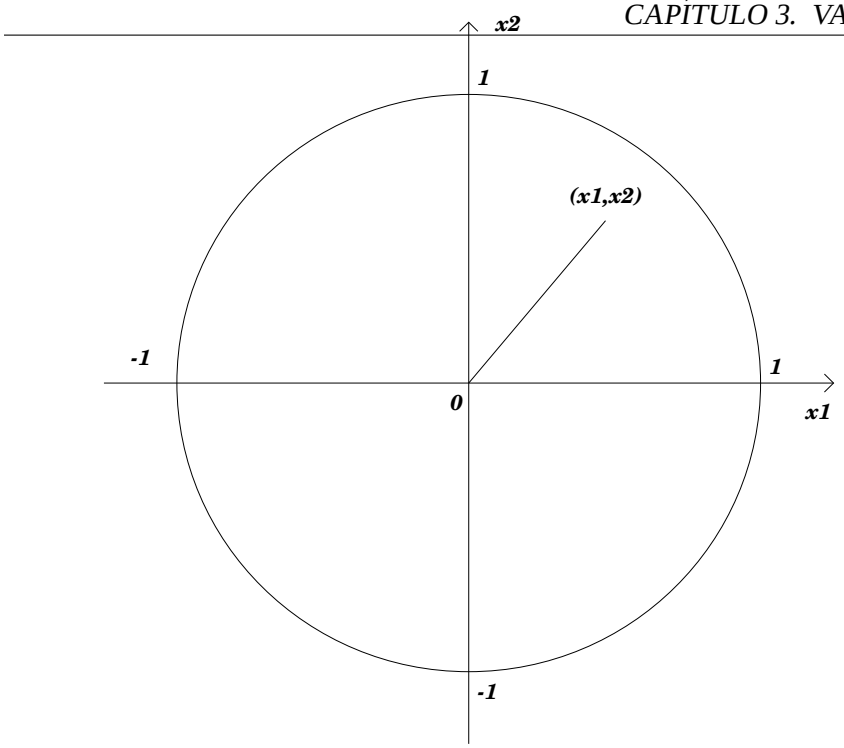


Figura 3.2.4: Representación esquemática del lanzamiento de un dardo.

Ejemplo 3.2.5 Una ampolleta tiene una probabilidad p de quemarse al instante de ser encendida. Si la ampolleta no se quema, entonces se sabe que la probabilidad que sobreviva $t > 0$ horas está dada por e^{-t}. ¿Cuál es la probabilidad que la ampolleta sobreviva 1 hora de funcionamiento?

Si denotamos por X el tiempo de vida de la ampolleta (esto es, el tiempo que tarda en quemarse), necesitamos calcular $P(X > 1)$. El espacio muestral se puede tomar como $\mathcal{X} = [0, \infty)$. Por las condiciones del problema, sabemos que $P(X = 0) = p$ (si la ampolleta se quema), y que $P(X > t|X > 0) = e^{-t}$ (cuando la ampolleta no se quema). Puesto que se quiere saber el valor de $P(X > 1)$, el teorema de probabilidades totales nos permite obtener que:

$$\begin{aligned} P(X > 1) &= P(X > 1|X = 0)P(X = 0) \\ &= +P(X > 1|X > 0)P(X > 0) \\ &= 0 \times p + e^{-1} \times (1 - p) = (1 - p) \times e^{-1}, \end{aligned}$$

que es lo que queríamos saber.

Note que, a diferencia del ejemplo anterior, hay un punto del espacio muestral que tiene una probabilidad positiva ($x = 0$).

3.3. Valores Esperados I

3.3.1. Motivación

La Ley de los Promedios o Ley de los Grandes Números es un resultado clave de la Teoría de Probabilidad. No existe la persona promedio, el alumno promedio o el árbol promedio, sino la altura promedio, el peso promedio, la renta promedio, el número promedio de accidentes, etc., que son valores de ciertas variables. Lo que se promedia son números reales, o bien elementos de un espacio vectorial (para el cual se puede hablar de combinaciones lineales). Como todo espacio vectorial de dimensión finita es representable por $\mathbb{R}^n$ y la suma y multiplicación en $\mathbb{R}^n$ se definen componente a componente, el caso fundamental es el de una variable con valores reales. La Ley de los Promedios, discutida informalmente en la Sección 1.2.1, refleja el hecho empírico que, bajo ciertas condiciones, los promedios exhiben una gran estabilidad. Si el valor de la variable cuantitativa en la i-ésima repetición se denota por y_i, lo que hacemos es considerar al número real y_i como el valor de una variable aleatoria. Las condiciones típicas bajo las cuales rige la Ley de los Promedios es que las repeticiones sean independientes y que el experimento se realice bajo condiciones semejantes. Esto se traduce formalmente en la condición

Las variables Y_i son i.i.d.

El promedio de n repeticiones es $t_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, que no es predecible exactamente, de modo que la incerteza se traduce en la distribución de la variable aleatoria

$$T_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Con esta formulación, la Ley de los Promedios se puede formular como un teorema, que se denomina Ley de los Grandes Números. Esencialmente, este teorema afirma que la distribución de T_n tiende a concentrase más y más en torno a cierto número μ, a medida que n aumenta:

$$P(\mu - \epsilon < T_n < \mu + \epsilon) \to 1, \text{ cuando } n \text{ tiende a } \infty.$$

Cuando este valor μ existe, él está determinado por la distribución común a todas las variables Y_i. Denotando por Y a una variable aleatoria, cuya distribución P_Y coincide con la de cada Y_i, el valor μ se denomina *media de la distribución* P_Y o *valor esperado* o *esperanza de la variable aleatoria* Y. Se plantea, entonces, el problema de dar una definición alternativa de μ o de $E(Y)$ que no requiera la repetición indefinida de un experimento. Aparte del ahorro de tiempo y energía, esto tiene la ventaja de que el concepto de media o valor esperado no depende de la interpretación frecuentista.

Para fijar las ideas consideremos el ejemplo pedestre, pero sencillo de llevar a cabo, – instamos al lector a hacerlo – que consiste en lanzar repetidamente un dado equilibrado. Si y_i es el número que muestra el dado en el i-ésimo lanzamiento, el gráfico de t_n versus n presenta inicialmente una gran inestabilidad, pero para valores grandes de n todos los puntos están muy cercanos a una recta horizontal, a una altura aproximada de 3.50. Si anotamos $z_i = 1$ si sale un seis y $z_i = 0$ en caso contrario, el promedio de los z_i coincide con la *proporción* p_n de veces que sale un seis en los

primeros n lanzamientos del dado. Por la interpretación frecuentista, p_n tiene como valor límite a la probabilidad que salga seis al lanzar *un* dado, de modo que el gráfico tiende nuevamente a una recta horizontal, esta vez con una altura igual a la probabilidad que salga seis en *un* lanzamiento del dado. Si Z representa una variable aleatoria con distribución igual a la de Z_i, tenemos el importante resultado:

$$E(Z) = P(Z = 1)$$

Notemos que $Z_i = h(Y_i)$, donde h es la función indicatriz del conjunto $\{6\}$. De esta forma, Z tiene la misma distribución que $h(Y)$ y, por tanto, el mismo valor esperado. Así,

$$E(Z) = E(h(Y)).$$

Esta profusión de paréntesis motiva la notación simplificada $E(Z) = Eh(Y)$. Es interesante resaltar que los promedios tienen perfecto sentido para cualquier función h con valores reales, sin importar la naturaleza de su dominio. Si el experimento consistiese en el lanzamiento de una moneda, con resultados $\omega = C$ y $\omega = S$, las repeticiones del experimento generarían una sucesión de letras que no se pueden promediar. Sin embargo, si para cada repetición uno gana $1000 si sale cara y pierde 500 si sale sello, la ganancia esperada, definida como límite de la ganancia promedio cuando el número de repeticiones tiende a infinito, es el valor esperado de la variable aleatoria definida sobre $\Omega = \{C, S\}$ por

$$W = 1000 \text{ si } \omega = C, \text{ y } W = -500 \text{ si } \omega = S.$$

La ganancia promedio en los primeros n juegos es

$$\begin{aligned} G_n &= \frac{1}{n}[1000 \times \text{número de caras} + (-500) \times \text{(n- número de caras)}] \\ &= [1000 \times \text{proporción de caras} + (-500) \times \text{proporción de sellos}] \\ &\to 1000 \times P(C) + (-500) \times P(S) \\ &= \sum p(\omega)h(\omega). \end{aligned}$$

donde

$$h(\omega) = 1000, \text{ si } \omega = C, \text{ y } h(\omega) = -500 \text{ si } \omega = S.$$

La función h coincide con la variable aleatoria W en la formulación abstracta.

En la interpretación subjetiva de la probabilidad $\Omega = \{\omega_1,\, i = 1, \ldots, k\}$ representa el conjunto de alternativas y $h(\omega)$ es la *utilidad* asociada con la alternativa ω. Esta utilidad no coincide, en general, con una ganancia monetaria, sino que es un concepto técnico. Por definición, ella es tal que uno debiera ser indiferente frente a la situación incierta que se presenta (por ejemplo, en un juego de azar o en una inversión financiera), y una utilidad cierta (segura) cuyo valor coincida con el valor esperado

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

La próxima sección discute fórmulas de cálculo.

3.3.2. Fórmulas para el valor esperado

Cuando el espacio muestral Ω es finito, la fórmula para el valor esperado es muy sencilla:

Definición 3.3.1 Sea Ω un espacio muestral numerable y sea X la variable aleatoria con valores $x = h(\omega)$, donde g es real valorada. El *valor esperado* o *esperanza* de X se denota por $E(X)$, y está dado por:

$$E(X) = \sum_{\omega \in \Omega} p(\omega) h(\omega), \tag{3.3.1}$$

donde la suma se interpreta como el valor de una serie cuando Ω es numerable. Si la serie no converge se dice que $E(X)$ no existe.

En particular, si $\omega = (x_1, \ldots, x_k)$ e $y = h(x_1, \ldots, x_k)$,

$$E(Y) = Eh(X_1, \ldots, X_k) = \sum_{(x_1,\ldots,x_k) \in \Omega} p(x_1, \ldots, x_k) h(x_1, \ldots, x_k). \tag{3.3.2}$$

Si la variable aleatoria X es discreta, siendo Ω arbitrario tenemos una definición alternativa:

Definición 3.3.2 El valor esperado o esperanza de una variable aleatoria X está dado por:

$$E(X) = \sum_{x \in \mathcal{X}} x p_X(x), \tag{3.3.3}$$

donde la suma se interpreta como el valor de una serie cuando X asume una cantidad numerable pero no finita de valores. Si la serie no converge se dice que $E(X)$ no existe.

Teorema 3.3.1 *Si Ω es numerable, las definiciones (3.3.1) y (3.3.2) son equivalentes.*

Demostración: La haremos sólo en el caso finito. Basta demostrar que las sumas (3.3.1) tienen el mismo valor. Como

$$p_X(x) = \sum_{h(\omega)=x} p(\omega),$$

(3.3.2) implica

$$\begin{aligned} E(X) &= \sum_{x \in \mathcal{X}} x \sum_{h(\omega)=x} p(\omega) \\ &= \sum_{x \in \mathcal{X}} \sum_{h(\omega)=x} x p(\omega) \\ &= \sum_{x \in \mathcal{X}} \sum_{h(\omega)=x} h(\omega) p(\omega) \end{aligned}$$

La última expresión es simplemente la suma en (3.3.1), efectuada en un orden distinto.

Las fórmulas (3.3.1) y (3.3.2) son ambas promedios ponderados de ciertos números. Estos números corresponden a los valores de una misma variable, pero, en general, (3.3.2) tiene menos términos (lo que no significa que sea más fácil de calcular). Si en vez de promediar valores de X interesara promediar valores de $Y = g(X)$, se tiene $y = v(\omega)$, donde $v(\omega) = g(h(\omega))$. Por lo tanto,

$$\begin{aligned} E(Y) &= \sum_{\omega\in\Omega} p(\omega)v(\omega) \\ &= \sum_{y\in\mathcal{Y}} yp_Y(y) \end{aligned}$$

Si $\mathcal{X}$ hubiera sido elegido como espacio muestral, la variable Y hubiese quedado expresada por la función g. Por (3.3.1) (con $\mathcal{X}$ en vez de Ω) se obtendría

$$E(Y) = \sum_{x\in\mathcal{X}} p_X(x)g(x),$$

que es nuevamente un promedio ponderado. Como $Y = g(X)$ se obtiene

$$E(g(X)) = \sum_{x\in\mathcal{X}} p_X(x)g(x).$$

En otras palabras el valor esperado de una función de la variable aleatoria X es un promedio ponderado, donde los números promediados son los valores de la función y los pesos son las probabilidades de los valores de la variable aleatoria. Por cierto, esto es, esencialmente, lo mismo que hicimos anteriormente, cambiando el par $(\Omega, \mathcal{X})$ por el par $(\mathcal{X}, \mathcal{Y})$.

Computacionalmente hablando, es más sencillo calcular el valor esperado de Y a partir de la función de probabilidad p_X, que a partir de p_Y. De hecho, $p_Y(y_0) = Eg(X)$, con g la función indicatriz de y_0.

Cuando no deseamos referirnos al espacio muestral Ω, es más conveniente *definir* directamente el valor esperado de una función real valorada de una variable aleatoria:

Definición 3.3.3 *Sea X una variable aleatoria con valores en un conjunto numerable $\mathcal{X}$. Sea g una función con dominio $\mathcal{X}$ y valores en $\mathbb{R}$. El valor esperado de $g(X)$ está dado por*

$$E(g(X)) = \sum_{x\in\mathcal{X}} p_X(x)g(x). \tag{3.3.4}$$

Teorema 3.3.2 *Si $Y = g(X)$, las definiciones 3.3.2 y 3.3.3 son equivalentes.*

Demostración: Idéntica a la del Teorema 3.3.1, salvo por cambios notacionales.

Los teoremas de equivalencia se pueden intuir directamente de la interpretación frecuentista. Basta pensar en n repeticiones del experimento y considerar la proporción de veces que aparece cada $\omega \in \Omega$, cada $x \in \mathcal{X}$ y cada $y \in \mathcal{Y}$. La extensión a espacios muestrales o variables aleatorias más generales, descansa en la idea que cualquier variable se puede aproximar adecuadamente por variables finitas.

Ejemplo 3.3.1 Suponga que X verifica $\mathcal{X} = \{-2, -1, 0, 1, 2\}$, con $p_X(x) = 0{,}1$, 0.2, 0.3, 0.2, 0.2 respectivamente. Considere $Y = g(X) = X^2$. Entonces $\mathcal{Y} = \{0, 1, 4\}$, y $p_Y(y) = 0{,}3, 0{,}4, 0{,}3$ respectivamente. Por otra parte, el valor esperado de Y, calculado directamente de la definición es:

$$E(Y) = 0 \times 0{,}3 + 1 \times 0{,}4 + 4 \times 0{,}3 = 1{,}6,$$

mientras que, usando (3.3.4) se llega a que

$$E(X) = 4 \times 0{,}1 + 1 \times 0{,}2 + 0 \times 0{,}3 + 1 \times 0{,}2 + 4 \times 0{,}2 = 1{,}6,$$

verificándose así el Teorema 3.3.2.

Cuando la función g es biyectiva, los cálculos se simplifican, pues en este caso tenemos que $\{x \in \mathcal{X} : g(x) = y\}$ es simplemente el singleton (o conjunto con sólo un punto) $\{g^{-1}(y)\}$, y por lo tanto,

$$p_Y(y) = p_X(g^{-1}(y)).$$

De esta forma (3.3.4) es inmediata.

Ejemplo 3.3.2 Sea X una variable aleatoria con función de probabilidad

$$\binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \ldots, n.$$

Entonces, la media $\mu = E(X)$ de la distribución de probabilidad P_X se calcula por

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} \\ &= np \end{aligned}$$

Del mismo modo,

$$\begin{aligned} E(X(X-1)) &= EX(X-1) \\ &= \sum_{k=0}^{n} k(k-1) \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2. \end{aligned}$$

Si escribimos $g(x) = x^2 = x + x(x-1)$, se tiene

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} [k(k-1)+k] \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} k(k-1) \cdot \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 + np \\ &= (np)^2 + np(1-p). \end{aligned}$$

Se observa que $E(X^2) > (E(X))^2$, a menos que $p = 0$ o $p = 1$. Finalmente, consideremos la función $g(x) = z^x$, donde z es un número real o complejo. Tenemos

$$\begin{aligned} E(z^X) &= \sum_{k=0}^{n} z^k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pz)^k (1-p)^{n-k} \\ &= (1-p+pz)^n. \end{aligned}$$

Ejemplo 3.3.3 Si Y tiene función de probabilidad

$$p_Y(y) = \frac{\lambda^y \exp(-\lambda)}{y!},\ y = 0, 1, 2 \ldots,$$

y t es un número real cualquiera,

$$\begin{aligned} E(\exp(tX)) &= \sum_{y=0}^{\infty} \exp(ty) \frac{\lambda^y \exp(-\lambda)}{y!} \\ &= \exp(-\lambda) \sum_{y=0}^{\infty} \frac{(\lambda \exp(t))^y}{y!} \\ &= \exp(-\lambda) \exp(\lambda \exp(t)) = \exp(\lambda(\exp(t) - 1)), \end{aligned}$$

la que está definida para cualquier real t.

3.3.3. Propiedades

A continuación listamos algunas propiedades del valor esperado, que no sólo son válidas para variables discretas. Se invita al lector a demostrarlas en el caso discreto.

Teorema 3.3.3

$$\text{Si } X = c, \text{ una constante, entonces } E(X) = c. \tag{3.3.5}$$

$$E(\sum_{i=1}^{n} c_i g_i(X)) = \sum_{i=1}^{n} c_i E(g_i(X)) \textit{ (linealidad)} \tag{3.3.6}$$

Ejemplo 3.3.4 Cuando se quiere adivinar el valor de X mediante un número real α, el error cometido es $X - \alpha$. Para deshacerse del potencial signo negativo podemos usar el valor absoluto o el cuadrado del error. Este último es más manejable analíticamente. En promedio, el cuadrado de error de predicción es $E(X-\alpha)^2$. Encontrar el valor de α que minimice este error cuadrático medio y el valor mínimo.

Solución:

$$\begin{aligned} E(X-\alpha)^2 &= E(X^2 - 2\alpha X + \alpha^2) \\ &= E(X^2) - 2\alpha E(X) + \alpha^2 \\ &= \alpha^2 - 2\mu\alpha + E(X^2) \\ &= (\alpha-\mu)^2 + E(X^2) - \mu^2. \end{aligned}$$

El polinomio en α se minimiza para $\alpha = \mu$ y el valor mínimo alcanzado tiene las expresiones alternativas

$$E(X-\mu)^2 = E(X^2) - \mu^2.$$

De esta forma, la media μ es la mejor predicción de X, siempre que aceptemos al error cuadrático medio como criterio de comparación.

Ejemplo 3.3.5 La distribución de X es simétrica con respecto al valor θ, si $X-\theta$ y $\theta - X$ tienen idéntica distribución. Probar que si $E(X) = \mu$ existe y la distribución de X es simétrica con respecto al valor θ, entonces $\mu = \theta$.

Demostración: La igualdad de distribuciones implica la igualdad de las medias. Por lo tanto, $E(X-\theta) = E(\theta - X)$. Por linealidad, $\mu - \theta = \theta - \mu$ y de aquí $\mu = \theta$.

Definiendo $Y_i = g_i(X)$, la propiedad de linealidad se escribe

$$E(\sum c_i Y_i) = \sum c_i E(Y_i), \tag{3.3.7}$$

que, de hecho, vale para variables Y_i arbitrarias (que no requieren ser funciones de una misma variable X). Para variables discretas, basta tomar $\boldsymbol{x} = (y_1, \ldots, y_k)$ y definir $g_i(\boldsymbol{x})$ como el valor de la i-ésima componente de $\boldsymbol{x}$. Tomando $c_i = 1$ se obtiene el caso más importante:

$$E(\sum Y_i) = \sum E(Y_i), \text{ esto es, } \textit{esperanza de la suma = suma de las esperanzas.} \tag{3.3.8}$$

3.3.4. Varianza y momentos

Definición 3.3.4 La *varianza* de la variable aleatoria X se define como

$$Var(X) = E\left(X - E(X)\right)^2 \tag{3.3.9}$$

siempre que la esperanza exista. En este caso, se define la *desviación estándar* de X como

$$\sigma(X) = \sqrt{Var(X)} \tag{3.3.10}$$

El Ejemplo 3.3.4 muestra que la varianza es el error cuadrático medio de la mejor predicción de X. Esto sugiere que a mayor varianza corresponde una mayor variabilidad de X o una mayor *dispersión* de su distribución. La unidad de medida de x, $\mu = E(X)$ y de $\sigma(X)$ son idénticas, mientras que las unidades de la varianza son los cuadrados de las unidades de los valores. El Ejemplo 3.3.4 entrega como subproducto la fórmula computacional

$$\text{Var }(X) = E(X^2) - (E(X))^2 = E(X^2) - \mu^2. \tag{3.3.11}$$

Finalmente, introducimos la noción de *momentos* de una variable aleatoria.

Definición 3.3.5 El *momento de orden* α de una variable aleatoria X, se define como

$$\mu_\alpha(X) = E(X^\alpha), \tag{3.3.12}$$

provisto que la esperanza correspondiente exista, y el *momento centrado de orden* α de X se define como

$$m_\alpha(X) = E((X - E(X))^\alpha). \tag{3.3.13}$$

El momento de orden 1 es simplemente, el valor esperado de X, mientras que el momento centrado de orden 1 es siempre 0, y el momento centrado de orden 2 es la varianza de X. Usualmente el interés se centra en momentos de orden k, donde k es un entero positivo.

Los valores esperados de funciones de una variable aleatoria se utilizan también en la definición de varias *funciones generadoras*, las que estudiamos en las Sección 3.8.2 y 3.8.3. El cálculo de Ez^X y Ee^{tX} en los Ejemplos 3.3.2 y 3.3.3 muestra que la función generadora de probabilidades de la distribución Binomial es $((1 - p + pz)^n$ y que el logaritmo de la función generadora de momentos es $\lambda(\exp(t) - 1)$.

Ejemplo 3.3.6 El Ejemplo 3.3.2 muestra cómo calcular directamente las cantidades $EX = np$ y $E(X(X-1)) = n(n-1)p^2$. Por linealidad, $x^2 = x(x-1)+x$ implica $EX^2 = n(n-1)p^2+np$, que coincide con lo obtenido directamente. Por (3.3.11), Var $X = np(1-p)$. Finalmente Var $\frac{X}{n} = \frac{p(1-p)}{n}$, que converge a 0 cuando n tiende a ∞. Este resultado es relevante para la Ley de los Grandes Números.

3.4. Función de Distribución Acumulada

Cuando los sucesos de interés dependen de una variable real, las preguntas relevantes se pueden formular, a menudo, en términos de intervalos, como por ejemplo: ¿Tendremos ma nana una temperatura superior a 5 grados? o ¿Será la inflación del próximo mes inferior a 1 %? o ¿Se mantendrá la variación del índice Dow-Jones estable entre -5 y +10 puntos?, etc. En estos casos $\mathcal{X} \subseteq \mathbb{R}$ y, de hecho, se puede tomar igual a $\mathbb{R}$, asignando probabilidad nula al complemento de $\mathcal{X}$. Si se asignan probabilidades a todos los intervalos, el axioma de σ-aditividad permite determinar automáticamente la probabilidad de todos los subconjuntos de $\mathbb{R}$ que aparecen en la realidad. En otras palabras, la distribución de probabilidad P_X queda completamente determinada en cuanto se conoce el valor de P_X para cada intervalo.

A primera vista, lo anterior requeriría especificar el tipo de intervalo, e.g. si el intervalo contiene o no su límite izquierdo a o su límite derecho b, así como si a o b son o no finitos. Para un tipo dado de intervalo, la probabilidad correspondiente depende naturalmente de a y de b, de modo que ella podría expresarse como $G_X(a, b)$ para cierta función G_X con dominio $\mathbb{R}^2$. Afortunadamente, podemos apelar a un procedimiento que es válido para cualquier medida positiva, que consiste en considerar previamente ciertas *probabilidades acumuladas* y deducir a partir de ellas la probabilidad de cualquier intervalo. Discutimos este enfoque en la próxima sección.

3.4.1. Definición y propiedades generales

Para un valor x cualquiera, están definidas las 4 probabilidades acumuladas $P(X \leq x)$, $P(X < x)$, $P(X \geq x)$ y $P(X > x)$. Como $((X \leq x), (X > x))$ y $((X < x), (X \geq x))$ son pares de sucesos complementarios, se cumplen automáticamente las identidades

$$\begin{aligned} P(X > x) &= 1 - P(X \leq x) \\ P(X \geq x) &= 1 - P(X < x). \end{aligned}$$

El problema se reduce así a asignar los valores de $P(X \leq x)$ y de $P(X < x)$ para cada x. Pero, para todo $x_0 \in \mathbb{R}$, el suceso $X < x_0$ es el límite, cuando n tiende a ∞, de la sucesión creciente de sucesos $X \leq x_0 - \frac{1}{n}$. La σ-aditividad implica que $P(X < x_0)$ satisface

$$P(X < x_0) = \lim_{n\to\infty} P(X \leq x_0 - \frac{1}{n}).$$

En consecuencia, basta conocer el valor de $P(X \leq x)$ para todo $x \in \mathbb{R}$. El resultado general, es que basta conocer una cualquiera de las probabilidades acumuladas para todo $x \in \mathbb{R}$. Esta discusión motiva la siguiente definición:

Definición 3.4.1 La *función de distribución acumulada* (f.d.a.), o simplemente, *función de distribución* de la variable aleatoria real valorada X, se define como:

$$F_X(x) = P_X(]-\infty, x]) = P(X \leq x), \quad \text{para } -\infty < x < \infty. \tag{3.4.1}$$

Para una función h definida sobre $\mathbb{R}$ y con valores en $\mathbb{R}$ utilizaremos la siguiente notación para los límites que se indican:

$$\begin{aligned} h(x_0^+) &\stackrel{\text{def}}{=} \lim_{x \to x_0^+} h(x), \\ h(x_0^-) &\stackrel{\text{def}}{=} \lim_{x \to x_0^-} h(x) \\ h(\infty) &\stackrel{\text{def}}{=} \lim_{x \to \infty} h(x) \\ h(-\infty) &\stackrel{\text{def}}{=} \lim_{x \to -\infty} h(x) \end{aligned}$$

La función F_X, para una variable aleatoria real, está definida en toda la recta real, y tiene las siguientes propiedades:

(a) $0 \leq F_X(x) \leq 1$ para todo $x \in \mathbb{R}$.

(b) F_X es no decreciente.

(c) Para todo $x \in \mathbb{R}$, los límites laterales $F_X(x^+)$ y $F_X(x^-)$ existen (pero no necesariamente coinciden).

(d) Para todo $x \in \mathbb{R}$, $F_X(x^-) = P(X < x)$.

(e) Para todo $x \in \mathbb{R}$, $F_X(x) = F_X(x^+)$, esto es, F_X es continua por la derecha.

(f) $F_X(\infty) = 1$ y $F_X(-\infty) = 0$.

(g) $P(X = x) = F_X(x) - F_X(x^-)$.

(h) $P(X \in]a, b]) = F_X(b) - F_X(a)$, y $P(X \in [a, b]) = F_X(b) - F_X(a^-)$.

La propiedad (a) se cumple por ser $F(x)$ una probabilidad. (b) es consecuencia de la monotonicidad de la probabilidad, pero se puede deducir directamente de la aditividad y la positividad como sigue: para $x_1 < x_2$ se tiene

$$\begin{aligned} F_X(x_2) &= P(X \in]-\infty, x_2]) = P(X \in]-\infty, x_1] \cup]x_1, x_2]) \\ &= P(X \in]-\infty, x_1) + P(X \in]x_1, x_2]) \\ &\geq P(X \in]-\infty, x_1]) = F_X(x_1) \end{aligned}$$

La propiedad (c) se satisface para toda función no decreciente. Las propiedades (d), (e) y (f) son consecuencia de la σ-aditividad, pero omitimos sus demostraciones. Finalmente, (g) y (h) son consecuencia de las propiedades anteriores y la aditividad. La continuidad por la derecha cambiaría a continuidad por la izquierda si $P(X \leq x)$ se reemplaza por $P(X < x)$.

Cuando X es el instante de falla de un equipo o de una componente, es común trabajar con la *función de confiabilidad*, definida por $S(x) = P(X > x)$, y que no es otra cosa que $1 - F_X(x)$.

Un resultado matemático importante, cuya demostración excede largamente los requisitos matemáticos de estas notas, es que dada cualquier función F que satisface (b), (e)

y (f), ella corresponde a la función de distribución acumulada de alguna variable aleatoria. Las propiedades (d),(g) y (h) permiten calcular las probabilidad de un intervalo y de un punto cualquiera.

Cuando existe un intervalo $S = [c, d]$, tal que $P(X \in S) = P(c < X < d) = 1$, los puntos c y d juegan el rol de $-\infty$ y $+\infty$ respectivamente. En particular, la condición (f) equivale a $F_X(c) = 0$ y $F_X(d) = 1$. Además, $F_X(x) = 0$ para todo $x < c$ y $F_X(x) = 1$ para todo $x \geq d$. Por (g), la función F_X es continua si y sólo si la probabilidad de cualquier conjunto de un elemento es nula. En este caso, $P_X([a, b]) = P_X(]a, b]) = P_X([a, b[) = P_X(]a, b[)$, para todo a, b.

3.4.2. Ejemplos

Ejemplo 3.4.1 Considere la siguiente tabla parcial de valores para F_X :

x	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
$F_X(x)$	0,30	0,38	0,45	0,52	0,58	0,62	0,65	0,68	0,70	0,71	0,72

Entonces:

- $P(X \leq 1{,}4) = 0{,}58$.
- $P(X > 1{,}7) = 0{,}32$.
- $P(1{,}4 < X \leq 1{,}7) = 0{,}68 - 0{,}58 = 0{,}10$.

Si se sabe que la función F es continua, se puede afirmar que

- $P(X < 1{,}4) = 0{,}58$.
- $P(X \geq 1{,}7) = 0{,}32$.
- $P(1{,}4 \leq X \leq 1{,}7) = P(1{,}4 \leq X < 1{,}7) = P(1{,}4 < X \leq 1{,}7) = 0{,}10$.

Ejemplo 3.4.2 La variable aleatoria X es el número de mujeres en un conjunto de 5 personas. Considere la siguiente tabla parcial de valores para F_X :

x	0	1	2	3	4	5
$F_X(x)$	,078	,337	,683	,913	,990	1,000

Entonces:

- $P(X \leq 2) = 0{,}683$.
- $P(X > 2) = 1 - {,}683 = {,}317$.
- $P(X \leq 2{,}5) = P(X \leq 2) = {,}683$.
- $P(X \geq 3) = P(X > 2) = {,}317$.
- $P(X = 3) = P(2 < X \leq 3) = {,}913 - {,}683 = {,}230$.

Ejemplo 3.4.3 Verificar que para todo k entero positivo, la función F definida por

$$F(x) = 1 - \sum_{j=0}^{k-1} \frac{x^j e^{-x}}{(k-1)!}, \qquad x > 0$$

y $F(x) = 0$ si $x \leq 0$, es la función de distribución acumulada de una variable aleatoria. En este caso $c = 0$ y $F(c) = F(0) = 0$. Un cálculo directo demuestra que la derivada de la función F es positiva para todo $x > 0$, de modo que F es creciente. Como e^x aumenta mucho más rápido que x^j, para todo $j \leq 0$, se tiene que $x^j e^{-x}$ converge a 0 cuando x tiende a ∞ y, por tanto, $F(\infty) = 1$.

3.4.3. Función de distribución acumulada para una variable aleatoria discreta

Recordemos que una variable aleatoria X se dice *discreta* si tiene un número finito o numerable de valores. Este es el caso de las variables en los Ejemplos 3.2.2 y 3.2.3. El adjetivo *discreta* se aplica también a su distribución de probabilidad P_X. Para una variable aleatoria discreta, P_X queda completamente determinada por su función de probabilidad p_X. Si el conjunto $\mathcal{X}$ de valores de X es un subconjunto de $\mathbb{R}$ tenemos la opción de elegir a $\mathbb{R}$ o a $\mathcal{X}$ como espacio muestral inducido por X. Sin pérdida de generalidad, supondremos que $p_X(x) > 0$ para todo $x \in \mathcal{X}$ (si no, simplemente eliminamos tal punto de $\mathcal{X}$). En este caso $\mathcal{X}$ se denomina *soporte* de X y es el menor subconjunto de $\mathbb{R}$ que cumple la propiedad $P_X(S) = 1$. Se lo puede escribir como

$$\mathcal{X} = \{x/\ p_X(x) > 0\}$$

La función de probabilidad p_X determina P_X sólo cuando X es una variable discreta. En particular, F_X se puede expresar como

$$F_X(x) = \sum_{\{y \in \mathcal{X}/y \leq x\}} p_X(y) \tag{3.4.2}$$

Las características de P_X se reflejan necesariamente en F_X. Cuando P_X es discreta, la función F_X sólo crece a saltos, coincidiendo el conjunto de puntos de salto con el soporte S. En otras palabras, su gráfico tiene forma de escalera, con un pelda no en cada punto del soporte, coincidiendo la altura de este pelda no con la probabilidad del punto de salto. Más formalmente, F_X es una función escalera, y con una discontinuidad de salto en cada punto $x \in \mathcal{X}$. La magnitud del salto es precisamente $p_X(x)$. Se ilustra esto en la Figura 3.4.5. Recíprocamente, si F_X es una función en escalera, la variable X es necesariamente discreta.

3.5. Variables Aleatorias Continuas y Función Densidad de Probabilidad

3.5.1. Definición y relación con la distribución acumulada

La analogía de la probabilidad con otras medidas positivas, como la masa de un cuerpo, sugiere utilizar la idea de densidad. Considerando a la masa como una medida positiva

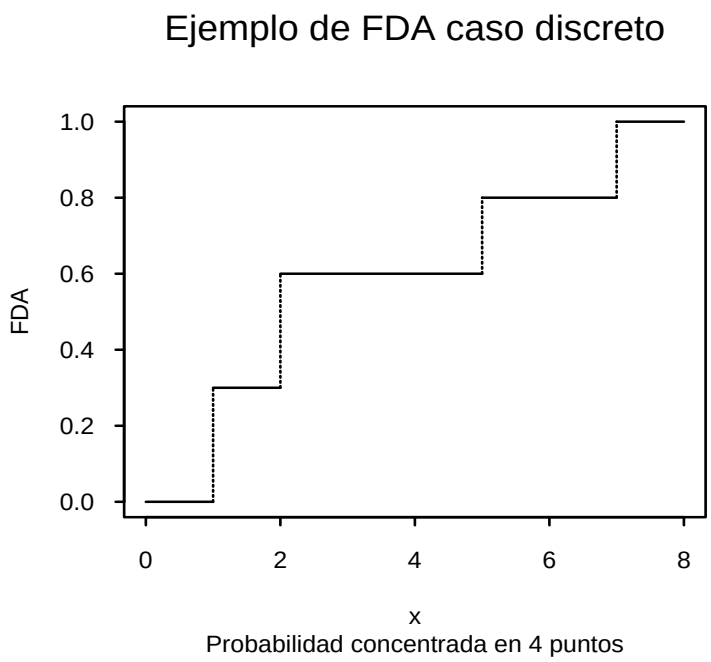


Figura 3.4.5: Ejemplo de Función de Distribución Acumulada para una variable aleatoria discreta.

que se define sobre una clase de subconjuntos de $\mathbb{R}^3$, que representan un cuerpo o sus partes, la densidad de masa es una función que asigna un valor real a cada punto de una región en $\mathbb{R}^3$. Si el cuerpo ocupa una región B en el espacio y la densidad se denota por $\rho(\boldsymbol{x})$, la masa del cuerpo es la integral de la función ρ sobre el conjunto B. Por analogía entre la masa de este cuerpo y la probabilidad $P_X(B)$, es razonable estudiar la posibilidad de expresar $P_X(B)$ como la integral de una cierta función, que naturalmente recibe el nombre de *densidad de probabilidad.* Este segundo enfoque tiene la ventaja de ser inmediatamente generalizable a $\mathbb{R}^k$.

La definición formal de la función densidad es la siguiente:

Definición 3.5.1 La variable aleatoria a valores reales X, o su distribución de probabilidad P_X, se dirá *absolutamente continua*, si existe una función f_X definida sobre $\mathbb{R}$, y con valores no negativos tal que para cualquier suceso $A \subset \mathcal{X}$

$$P_X(A) = P(X \in A) = \int_A f_X(x)dx. \tag{3.5.1}$$

La función f_X se denomina *función densidad de probabilidad*, o, simplemente *densidad* de X.

Las variables aleatorias en el Ejemplo 3.2.4 son absolutamente continuas. Cuando A es un intervalo con límite inferior a y límite superior b, que es el caso más común, (3.5.1)

se escribe en la forma más familiar

$$P_X([a,b]) = P(a \leq X \leq b) = \int_a^b f_X(x)dx. \tag{3.5.2}$$

Cuando $a = -\infty$ o $b = \infty$ la expresión se entiende en el sentido de una integral impropia, es decir, haciendo tender a o b al límite correspondiente. Desde un punto de vista matemático, las propiedades (3.5.1) y (3.5.2) son, de hecho, equivalentes. Geométricamente, (3.5.2) es el área bajo el gráfico de f_X entre a y b.

Tomando $A =]-\infty, x]$ en (3.5.1) se obtiene la importante relación

$$F_X(x) = \int_{-\infty}^x f_X(t)dt, \tag{3.5.3}$$

que liga a la densidad con la distribución acumulada. Por otra parte, a partir de (3.5.3), es inmediato ver que

$$F'_X(x) = f_X(x), \tag{3.5.4}$$

bajo ciertas condiciones de regularidad que mencionamos en la sección 3.5.2. De (3.5.4) se tiene que F_X es una antiderivada o primitiva G de f. Entonces $F_X(x) = G(x)+C$ y la constante C se determina conociendo el valor de $F_X(x)$ en cualquier punto, incluyendo ∞ y $-\infty$. Por ejemplo, si $f_X(x) = e^{-x}, x > 0$ y $f_X(x) = 0$ en otro caso, se tiene que $G(x) = -e^{-x}$ es una primitiva y $F_X(x) = -e^{-x} + C$. De $P(X \leq 0) = 0$ se deduce que $F_X(x) = 0$ y, por tanto, $C = 1$. Lo mismo se obtiene de $1 = F(\infty) = 0+C$.

La definición intuitiva de densidad de masa ρ en un punto $\boldsymbol{x}_0$ dado, es que ella aproxima al cuociente entre la masa de una peque na parte del cuerpo que contiene a $\boldsymbol{x}_0$ y su volumen. El producto de $\rho(\boldsymbol{x}_0)$ y el volumen de la región aproxima entonces la masa de la región. El mismo argumento sugiere que el producto de $f_X(x_0)$ y la longitud de un peque no intervalo que contiene a x_0 aproxima la probabilidad que X tome un valor en dicho intervalo. Si la unidad de medida de x es centímetros, f_X tiene dimensión cm^{-1}; si ella es segundos, la unidad de f_X es seg^{-1}. Esto muestra que no tiene sentido interpretar a $f_X(x)$ como una probabilidad, a diferencia de lo que acontece con $p_X(x)$ en el caso discreto. Por ejemplo, si X mide el peso de una persona en kilogramos, $f_X(68) \times 0{,}2$ aproxima $P(67{,}9 \leq X \leq 68{,}1) = F_X(68{,}1) - F_X(67{,}9)$.

Es instructivo buscar una interpretación directa de (3.5.4), que no descanse en el teorema fundamental del cálculo. Para ello hacemos la analogía con la densidad de masa. Consideremos un intervalo peque no $(x_0 - \frac{\epsilon}{2}, x_0 + \frac{\epsilon}{2}]$, centrado en un punto x_0 de la recta real y aproximemos la densidad en x_0 por el cuociente entre su probabilidad y el largo del intervalo. Entonces,

$$\begin{aligned} f_X(x_0) &\approx \frac{P(x_0 - \frac{\epsilon}{2} < X \leq x_0 + \frac{\epsilon}{2}]}{\text{largo}((x_0 - \epsilon, x_0 + \epsilon])} && (3.5.5)\\ &= \frac{F_X(x_0 + \frac{\epsilon}{2}) - F_X(x_0 - \frac{\epsilon}{2})}{\epsilon} && (3.5.6)\\ &\approx F'_X(x_0) && (3.5.7) \end{aligned}$$

3.5.2. Caracterización de una función densidad de probabilidad

Así como una función F no decreciente, continua por la derecha y que satisface $F(-\infty) = 0$ y $F(\infty) = 1$ se puede considerar como la función de distribución acumulada de cierta variable aleatoria, una función f se puede considerar como la función densidad de probabilidad de cierta variable aleatoria X si ella satisface las condiciones

$$f(x) \geq 0, \text{ para todo } x \in \mathbb{R}. \tag{3.5.8}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{3.5.9}$$

En efecto, basta definir la función de distribución acumulada F_X mediante (3.5.3).

Es más habitual determinar modelos probabilísticos especificando la función densidad de probabilidad que usando la función de distribución acumulada. Por otra parte, suele ser conveniente definir la densidad salvo por una constante de proporcionalidad. Si $f(x) = cg(x)$, donde g es una función definida en $\mathbb{R}$, a valores reales no negativos, y con integral finita, digamos I, (3.5.9) implica $cI = 1$, o sea $f(x) = g(x)I^{-1}$ es efectivamente una densidad. Por ejemplo, sea f definida en $[0, 1]$, como $f(x) = cx^3$, y 0 en todo otro punto. ¿Cuál es el valor de c para que f sea una densidad? Todo lo que se necesita es que

$$\int_{-\infty}^{\infty} f(x)dx = c\int_0^1 x^3 dx = \frac{c}{4} = 1,$$

por lo que se requiere $c = 4$.

3.5.3. Propiedades analíticas y otros tipos de distribuciones

3.5.3.1. Interpretaciones de la densidad

Cuando f_X es continua en la vecindad de x_0 se satisface

$$\frac{P(x_0 - \frac{\epsilon}{2} \leq X \leq x_0 + \frac{\epsilon}{2})}{\epsilon} \to f_X(x_0), \quad \text{cuando } \epsilon \to 0^+,$$

de modo que $f_X(x_0)\epsilon$ es la probabilidad aproximada de un peque no intervalo rodeando a x_0. Haciendo variar x_0, esto describe la forma en que se concentra la distribución de probabilidades de X en torno a x. Reemplazando x_0 por x, ϵ por dx, y el intervalo centrado por uno con límite izquierdo x, la igualdad aproximada toma una forma muy sugerente:

$$P(x \leq X \leq x + dx) \approx f_X(x)dx \tag{3.5.10}$$

El valor exacto del lado izquierdo es $F_X(x + dx) - F_X(x)$. La aproximación (3.5.10) corresponde a una expansión de Taylor de primer orden de F_X en torno a x. Si F_X es diferenciable en x, el error de aproximación en (3.5.10) tiende a 0 más rápido que dx (o sea al dividirlo por el número positivo dx el cuociente converge a 0). Escribimos simbólicamente esto como

$$P(x \leq X \leq x + dx) = f_X(x)dx + o(dx) \tag{3.5.11}$$

Por otra parte, aplicando (3.5.2) al lado izquierdo de (3.5.10) se tiene la aproximación:

$$\int_x^{x+dx} f_X(t)dt \approx f_X(x)dx; \tag{3.5.12}$$

Cuando f_X es continua, el teorema del valor medio para integrales garantiza

$$\int_x^{x+dx} f_X(t)dt \approx f_X(x^*)dx, \quad \text{para algún } x \leq x^* \leq x+dx.$$

El error de aproximación es $|f_X(x^*) - f_X(x)|dx$, que se puede acotar por Mdx, donde M es la máxima variación de la densidad en el intervalo. Si f_X es continua en este intervalo, el número M tiende a 0 cuando dx tiende a 0.

3.5.3.2. Distribuciones absolutamente continuas y no atómicas

Las propiedades de la distribución de probabilidad P_X están vinculadas con propiedades de F_X. Un ejemplo importante es el siguiente

Definición 3.5.2 *Si $P(X = x) = 0$ para todo $x \in \mathbb{R}$ se dice que P_X es* continua *o* no atómica.

El siguiente teorema es inmediato

Teorema 3.5.1 *P_X es no atómica si y sólo si F_X es continua.*

Claramente toda distribución absolutamente continua es no atómica. Por otra parte, todas las distribuciones de probabilidad continuas que se utilizan en la práctica son, de hecho, absolutamente continuas. Los contraejemplos son algo complicados de construir y revisten un interés puramente matemático. Muchos libros utilizan el término variable aleatoria continua para referirse a una variable que admite una función densidad. Con nuestra definición, ambos conceptos no son equivalentes.

La continuidad absoluta de la distribución, es decir la existencia de una función densidad, equivale esencialmente a cualquier de las dos propiedades equivalentes (la demostración de la equivalencia requiere de herramientas matemáticas sofisticadas):

(a) *Si el largo de $A \subseteq \mathbb{R}$ es 0, entonces $P(A) = 0$.*
(b) *Si el largo de $A_n \subseteq \mathbb{R}$ tiende a 0, lo mismo sucede con $P(A_n)$.*

Aplicando (a) a $A = \{x\} = [x, x]$ se deduce que $P(X = x) = 0$; aplicando (b) a $A_n =]x - \frac{1}{n}, x + \frac{1}{n}]$, se deduce que $F_X(x) - F_X(x^-) = \lim n \to \infty (F(x + \frac{1}{n}) - F(x - \frac{1}{n})) = 0$. Esto proporciona dos demostraciones alternativas al hecho que una distribución absolutamente continua es no atómica.

3.5.3.3. Falta de unicidad de la función densidad

Si las funciones f y g satisfacen (3.5.9) y (3.5.3) y difieren sólo en un conjunto finito de puntos, ellas son dos funciones de densidad de una misma distribución. Llamamos a f y g dos *versiones* de la función densidad. Con una definición adecuada de integral, el conjunto finito se puede reemplazar por un conjunto de largo cero. En $\mathbb{R}$ los conjuntos de largo cero que no son numerables resultan ser bastante extra nos (o patológicos, como se dice en lenguaje matemático). En cambio, una curva suave en $\mathbb{R}^2$, e.g. una circunferencia o una línea recta, tiene área cero y no es numerable.

3.5.3.4. Distribuciones mixtas

Existen distribuciones de probabilidad que no son ni discretas ni continuas, a las que se denomina distribuciones *mixtas.* Su función distribución acumulada no es una función puramente de saltos ni una función continua, sino una combinación convexa de ambas. Esto quiere decir que toda distribución mixta F_X se puede escribir como una combinación lineal $\alpha F_D + (1-\alpha)F_C$, donde $0 < \alpha < 1$, y F_D y F_C son las funciones de distribución acumulada de dos variables aleatorias D y C, discreta y continua respectivamente. La variable aleatoria en el Ejemplo 3.2.5 tiene una distribución mixta.

3.6. Familias Paramétricas de Distribuciones de Probabilidad

3.6.1. Propiedades generales

Cuando se cuenta con una distribución de proporciones empíricas, es común tratar de mirarlas como una aproximación a una distribución de probabilidad teórica. Se dispone para ellos de muchos tipos de distribuciones de probabilidad conocidas. Dado un tipo particular de distribuciones, una distribución específica queda determinada por un vector de parámetros, que denotamos por θ. Estos parámetros ajustables se eligen para que las proporciones empíricas se parezcan lo más posible a las probabilidades teóricas correspondientes. Formalmente, tenemos una familia de distribuciones $\{P_\theta,\, \theta \in \Theta\}$. Elegir un miembro de esta familia equivale a elegir un elemento $\theta \in \Theta$. El único caso que consideraremos acá es $\Theta \subset \mathbb{R}^k$, donde k números reales determinan la distribución de manera única. Por simplicidad de lenguaje se suele hablar de la distribución P_θ, aunque θ no esté especificado. Si X sigue la distribución P_θ, lo que escribimos $X \sim P_\theta$, la probabilidad que el valor de X pertenezca a A se denota por $P_\theta(A)$.

Lo más cómodo es representar a P_θ por su función de probabilidad $p(\cdot;\theta)$, o su función de densidad $f(\cdot;\theta)$, según sea la distribución discreta o absolutamente continua. Estas

funciones son no negativas y satisfacen

$$\begin{aligned} \sum_{x \in S} p(x;\theta) &= 1 \\ \int_S f(x;\theta)dx &= 1, \end{aligned}$$

donde S es el soporte de P_θ, o sea, $P(X \in S) = 1$. En la práctica, S es un intervalo de números reales o enteros.

Una función no negativa $g(x,\theta)$ con suma o integral denotada por $I(\theta) < \infty$, genera una función probabilidad o densidad al dividirla por $I(\theta)$. Esto proporciona una fuente ilimitada de familias de distribuciones, siendo el único problema el cálculo de $I(\theta)$. En la práctica $I(\theta)$ es una suma, el valor de una serie, una integral definida o una integral impropia.

Ejemplo 3.6.1 *En la sección 1.6 discutimos especialmente las distribuciones de probabilidad cuyo soporte sea subconjunto de los enteros no negativos. Dada una serie de potencias conocida*

$$G(z) = \sum_{k=0}^{\infty} c_k z^k, \ |z| < r,$$

se obtiene que

$$p(k,\theta) = \frac{c_k \theta^k}{G(\theta)}, \ 0 < \theta < r.$$

es una legítima familia uniparamétrica de funciones de probabilidad, es decir indexadas por el número real θ.

3.6.2. Taxonomía

Los libros de probabilidad suelen entregar una peque na lista de distribuciones de probabilidad, donde se indican algunas de sus principales características. Este libro no es una excepción; la Sección 3.10 entrega tal lista. Cabe se nalar que libros de referencia, como la colección escrita por Johnson y Kotz, contiene muchas más distribuciones e información sobre ellas.

Esencialmente, podemos pensar que disponemos de un diccionario enciclopédico de distribuciones y precisamos estrategias de búsqueda. Los principales elementos para acotar la búsqueda son

- Distribuciones discretas versus continuas.
- El soporte de la distribución.

Al igual que en las tablas de integrales, se reduce mucho el espacio necesario si las distintas expresiones se reducen a un número más peque no de formas estándar o canónicas.

3.6.3. Familias paramétricas discretas

- *Caso degenerado:* Si $\text{card } S = 1$, la variable aleatoria se degenera en una constante.
- *Caso binario:* Si $\text{card } S = 2$, X es una variable aleatoria llamada *binaria*. Si $a < b$ son los dos valores posibles, la variable X se puede expresar como una transformación lineal afín de una variable $Z \sim \text{Bern}(p)$, mediante $Y = a + (b-a)Z$.
- *Caso finito:* Los recuentos constituyen el caso más típico. Otro caso importante es una versión discreta del tiempo. Si el valor mínimo es $m > 0$, la nueva variable $Y = X - m$ toma valores en $\{0, 1, \ldots, m\}$. Las distribuciones Binomial, Hipergeométrica, y Uniforme discreta son los casos más conocidos (ver Sección 3.10).
- *Caso entero no negativo:* Si no hay un número máximo claro, se toma formalmente $n = \infty$, o sea el soporte está constituido por todos los enteros no negativos. Las distribuciones más conocidas son la Geométrica, Poisson y Binomial negativa (ver Sección 3.10). El valor mínimo puede ser $m > 0$, o bien ser eliminado por resta. Por ejemplo el número X de lanzamientos que se requiere para obtener 2 caras tiene distribución $\text{BN}(2, p)$ y el número de de sellos Y tiene distribución $\text{BN0}(2, p)$. Estas variables satisfacen la relación $Y = X - 2$.
- *Reales con número finito de dígitos*. Un intervalo $[a, b] \subseteq \mathbb{R}$ se aproxima por un conjunto finito S de puntos equiespaciados, e.g. truncando los números reales a sólo k dígitos. Por ejemplo, $[2, 3]$ se aproxima por $\{2{,}00, 2{,}02, \ldots, 2{,}99, 3{,}00\}$ para $k = 2$. Un cambio de variables $X = a + hY$ reduce una distribución de probabilidad con soporte S a otra con soporte canónico $\{0, 1, 2, \ldots, n\}$.

3.6.4. Familias paramétricas continuas

Para una distribución continua, es irrelevante si el intervalo contiene o no sus extremos, pues ellos tienen probabilidad nula. Escribimos el soporte como un conjunto cerrado.

3.6.4.1. Reducción a la forma canónica.

- La transformación $x = a + (b-a)y$ reduce el caso de una variable X con soporte $[a, b]$ al de una variable Y con soporte $[0, 1]$.
- La translación $X = a + Y$ reduce el soporte $[a, \infty[$ a $[0, \infty[$.
- $X = b - Y$, que es una combinación de translación con reflección con respecto al origen reduce el soporte $]-\infty, b]$ a $[0, \infty[$.
- Si el soporte es $\mathbb{R}$, él es preservado por toda transformación lineal afín no constante.

De esta forma, es suficiente estudiar familias de distribuciones cuyo soporte es $[0, 1]$, $[0, \infty[$, o $\mathbb{R} =]-\infty, \infty[$.

3.6.4.2. Principales distribuciones.

- *Soporte* $[0,1]$: Distribución Beta $[\alpha,\beta]$, cuyo caso más importante es la distribución uniforme.
- *Soporte* $[0,\infty,[$: Exponencial, Gama, Weibull, Log-normal, valor extremo, Ji-cuadrado, F de Snedecor.
- *Soporte* $\mathbb{R}$. Normal, Student, Logística, Cauchy

3.7. Variables Discretas Asociadas con el Proceso de Bernoulli

3.7.1. Definiciones y notaciones básicas

La definición frecuentista de probabilidad descansa en las repeticiones hipotéticas de un experimento. Con la noción de independencia de variables aleatorias, tal situación se representa por una sucesión de variables aleatorias $Y_1, Y_2, \ldots$, i.i.d., es decir, independientes e idénticamente distribuidas. Consideremos un suceso cualquiera A, que puede o no ocurrir en la i-ésima repetición, y definamos su variable indicatriz X_i por $X_i = 1$ si $Y_i \in A$ y $X_i = 0$ en caso contrario. Entonces, $X_1, X_2, \ldots$ son también i.i.d. y cada variable X_i es binaria, con valores 0 y 1. La distribución de probabilidad de Y_i se denomina Bern (p), donde $p = P(X_i = 1)$ ($p = P(Y_i \in A)$ en nuestro caso). A continuación damos una definición formal, junto con la nomenclatura usual.

Definición 3.7.1 La distribución de probabilidad que asigna probabilidad p al valor 1 y probabilidad $q = 1 = p$ al valor 0, se denomina *Bernoulli con parámetro* p. Un proceso de Bernoulli de parámetro p es una sucesión de variables aleatorias i.i.d. con distribución Bern (p), lo que se escribe $X_1, X_2, \ldots \overset{iid}{\sim}$ Bern (p).

La variable X_i representa el resultado del i-ésimo *ensayo*, interpretándose $X_i = 1$ como un *éxito* y $X_i = 0$ como un *fracaso*. El parámetro p común representa la probabilidad de éxito, $P(X_i = 1)$, denotándose la probabilidad de fracaso por $q = 1 - p$.

Un modelo concreto es la repetición indefinida del lanzamiento de una moneda, con probabilidad p de salir cara y $q = 1 - p$ de salir sello, donde el resultado del i-ésimo lanzamiento es $x_i = 1$ si sale cara y $x_i = 0$ si sale sello. La proporción de éxitos en los primeros n ensayos es

$$p_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

i.e., el promedio de las primeras n variables. Si $X_i = 1$ cuando $Y_i \in A$, esta proporción es la frecuencia relativa con que ocurre el suceso A en n repeticiones del experimento. La Ley de los Grandes Números implica que p_n tiende a $p = P(X_i = 1)$.

Definamos ahora las siguientes variables aleatorias:

N_n : número de éxitos obtenidos en los n primeros ensayos,
es decir, hasta el *instante* n, inclusive.
T_k : instante donde ocurre el k-ésimo éxito, con $T_0 = 0$.
Z_k : número de ensayos que requiere obtener el k-ésimo éxito,
contado a partir del ensayo en que se obtiene el $k-1$-ésimo éxito.
Observe que $Z_1 = T_1$
W_k : número de fracasos consecutivos que precede al k-ésimo éxito.

Es inmediato que $N_n = \sum_{i=1}^{n} X_i$, $Z_k = T_k - T_{k-1}$, $W_k = Z_k - 1$ y $T_k = \inf_{n/N_n=k}$. De aquí se deduce

$$T_k \leq n \Leftrightarrow N_n \geq k.$$

$$T_k = \sum_{i=1}^{k} Z_i, \qquad T_k - k = \sum_{i=1}^{k} W_i.$$

La tabla siguiente ilustra las definiciones para una realización particular de las variables $X_1, X_2, \ldots, X_{20}$.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_n	0	1	0	0	1	1	0	1	0	0	0	0	1	0	1	0	0	0	1	1
N_n	0	1	1	1	2	3	3	4	4	4	4	4	5	5	6	6	6	6	7	8

k	1	2	3	4	5	6	7
T_k	2	5	6	8	13	15	19
Z_k	2	3	1	2	5	2	4
W_k	1	2	0	1	4	1	3

Hacemos notar que las variables aleatorias N_n, T_k, Z_k y W_k han sido definidas sin especificar su distribución de probabilidad. De cualquiera de las 6 filas anteriores se pueden deducir las otras 5 mediante un simple cálculo aritmético. Como el modelo probabilístico subyacente a los resultados de la primera fila está determinado por p, lo mismo sucede con las distribuciones de probabilidad asociadas a las otras filas. La siguiente tabla muestra los nombres asignados a las diversas distribuciones. Posteriormente deduciremos las funciones de probabilidad correspondientes.

Distribución de	nombre	Notación
X_n	Bernoulli de parámetro p.	Bern (p)
N_n	Binomial de parámetros n y p.	Bin (n, p)
Z_k	Geométrica de parámetro p.	Geom(p)
W_k	Geométrica de parámetro p trasladada al origen	Geom0(p)
T_k	Binomial negativa de parámetros k y p.	BN(k, p).
$T_k - k$	Binomial negativa de parámetros k y p trasladada al origen	BN0(k, p).

La tabla indica que las distribuciones de Z_k y de W_k no dependen de k, lo que será demostrado más adelante. Si no se quiere hacer uso de este hecho, usamos $Z_1 = T_1$ y

$W_1 = T_1$ para definir las distribuciones Geom(p) y Geom0(p) respectivamente, las que corresponden a BN$(1,p)$ y BN0$(1,p)$ respectivamente. Fórmulas para sus funciones de probabilidad se encuentran en la Sección 3.10.

3.7.2. Recuentos, camino aleatorio y la distribución Binomial.

La función de probabilidad de N_n se indica en (3.10.7), y ya ha sido deducida, en ejemplos de capítulos anteriores. Un caso especial con $n = 10$, se discute en el Ejemplo 3.2.1. Para mayor facilidad la repetimos acá en el caso general. Observe que $N_n = k$ si y sólo si hay exactamente k unos entre $X_1, \ldots, X_n$ (y, por ende, exactamente $n - k$ ceros). Cada n-tupla de unos y ceros con exactamente k unos tiene probabilidad $p^k(1-p)^{n-k}$. Por otra parte, el número de n-tuplas con exactamente k ceros coincide con el número de formas diferentes de asignar k objetos indistinguibles a n posiciones diferentes, o, equivalentemente, al número de posibles subconjuntos de tama no k de un total de n objetos. Este número es exactamente $\binom{n}{k}$, y así hemos obtenido que

$$p_{N_n}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ para } k \in \{0, 1, \ldots, n\}.$$

Si imaginamos que alguien juega repetidamente, ganando a si $X_i = 1$ y perdiendo b si $X_i = 0$, la ganancia acumulada después de n juegos (que puede ser negativo) es $Y_n = aN_n - b(n - N_n) = -bn + (a + b)N_n$. Si el capital inicial es C_0, el jugador se arruina si $-Y_n$ excede C_0. El caso más importante es $a = b = 1$. El gráfico de Y_n versus n, o la sucesión Y_n se denomina camino aleatorio.

No hay nada especial en los primeros n ensayos. Si representamos al j-ésimo ensayo por el número $j \in \mathcal{N} = \{1, 2, \ldots\}$, el número N_A de éxitos en un conjunto A de ensayos satisface

$$N_A = \sum_{j \in A} X_j \sim \text{ Bin } (\text{ card } A, p) \tag{3.7.1}$$

Para $A = \{1, 2, \ldots, n\}$ N_A se reduce a N_n.

Ejemplo 3.7.1 Demostrar que $Y_j \sim$ Bin (n_j, p), $j = 1, \ldots, k$ e $Y_1, \ldots, Y_k$ independientes, implica

$$\sum_{j=1}^{k} Y_j \sim \text{ Bin } (\sum_{j=1}^{k} n_j, p).$$

Demostración: Aplicando (3.7.1) a cada elemento de una partición ordenada de $S = \{1, 2, \ldots, n\}$, con $n = \sum_{j=1}^{k} n_j$, se tiene

$$N_S = \sum_{j=1}^{k} N_{A_j}.$$

Claramente N_{A_j} e Y_j tienen la misma distribución y los N_{A_j} son independientes. Además $N_S \sim$ Bin $(\sum_{j=1}^{k} n_j, p)$, por definición.

Ejemplo 3.7.2 Demostrar que si $X \sim$ Bin (n, p), entonces se tiene que $Y = n - X \sim$ Bin $(n, 1-p)$.

Demostración: Una demostración directa se obtiene a partir de $p_Y(y) = P(Y = y) = P(n-Y = n-y) = P(X = n-y) = p_X(n-y)$ y aplicando (3.10.7). Una alternativa más interesante consiste en definir $Y_i = 1 - X_i$, verificar que $Y_1, Y_2, \ldots$ es un proceso de Bernoulli con parámetro $1-p$, y utilizar la representación $Y = \sum_1^n Y_i$.

3.7.3. Distribución geométrica

3.7.3.1. Tiempo entre éxitos consecutivos

La función de probabilidad de la distribución geométrica está dada por (3.10.8). El resultado fundamental está contenido en la siguiente proposición.

Proposición 3.7.1 Para un proceso de Bernoulli con probabilidad de éxito p se tiene que los números de ensayos entre éxitos sucesivos, $W_1, W_2, \ldots$ son variables aleatorias i.i.d. con distribución común geométrica de parámetro p trasladada al origen. Las distancias entre éxitos consecutivos $Z_1, Z_2, \ldots$ son i.i.d. con distribución común geométrica de parámetro p.

La demostración general se deja como ejercicio e ilustramos la idea básica mediante un caso particular. De la Tabla

$$\begin{aligned} P(W_1 = 1, W_2 = 2, W_3 = 0, W_4 = 1) &= P(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 0, \\ &\quad X_5 = 1, X_6 = 1, X_7 = 0, X_8 = 1) \\ &= qpqqppqp \\ &= q^1pq^2pq^0pq^1p \\ &= q^{w_1}pq^{w_2}pq^{w_3}pq^{w_4}p \end{aligned}$$

Por el Teorema de Factorización se obtiene la independencia de los sucesos $W_1 = 1$, $W_2 = 2$, $W_3 = 0$, y $W_4 = 1$. Usando este mismo argumento para otros valores de las covariables y comparando con (3.10.9) se completa la demostración.

La variable $W_j = Z_j - 1$ es el número de fracasos que media entre el $(j-1)$-ésimo y el j-ésimo éxito. Su función de probabilidad está dada por (3.10.9).

3.7.3.2. Falta de memoria.

Una propiedad interesante de la distribución geométrica es la llamada *falta de memoria*. En efecto, suponga que, para un proceso de Bernoulli con probabilidad de éxito p,

el instante del primer éxito T_1 (que, como sabemos, tiene distribución geométrica de parámetro p) es posterior al instante actual, digamos, t, esto es, $T_1 > t$. La pregunta que surge entonces, es: ¿Cuál es la probabilidad que tengamos que esperar más de s ensayos para observar el primer éxito? En otras palabras, dado que ya llevamos t ensayos esperando el primer éxito, ¿Cuál es la probabilidad que tengamos que esperar al menos s ensayos más? Lo que se requiere calcular es $P(T_1 > s+t|T_1 > t)$. Ahora,

$$P(T_1 > s+t|T_1 > t) = \frac{P(T_1 > s+t, T_1 > t)}{P(T_1 > t)} = \frac{P(T_1 > s+t)}{P(T_1 > t)}. \qquad (3.7.2)$$

Por otra parte,

$$\begin{aligned} P(T_1 > t) &= \sum_{k=t+1}^{\infty} p_{T_1}(k) = \sum_{k=t+1}^{\infty} p(1-p)^{k-1} \\ &= p(1-p)^t \sum_{k=t+1}^{\infty} (1-p)^{k-t-1} \\ &= p(1-p)^t \sum_{j=0}^{\infty} (1-p)^j \qquad (\text{con } j = k-t-1) \\ &= p(1-p)^t \times \frac{1}{1-(1-p)} = (1-p)^t, \end{aligned}$$

por lo que usando (3.7.2) se obtiene

$$P(T_1 > s+t|T_1 > t) = \frac{(1-p)^{s+t}}{(1-p)^t} = (1-p)^s = P(T_1 > s),$$

y llegamos a la más bien sorprendente conclusión que la probabilidad en cuestión no depende de t. Esta propiedad de la distribución geométrica se llama, precisamente, *falta de memoria*.

3.7.4. Instantes en que ocurre un éxito y la distribución Binomial negativa.

Vamos ahora a demostrar que la distribución de T_k es BN(k,p), cuya función de probabilidad está dada por (3.10.10). Como el k-ésimo éxito no puede obtenerse antes del instante k, y por otra parte, no es posible acotar el número de ensayos requerido para obtenerlo, T_k toma valores en $\{k, k+1, k+2, \ldots\}$. El suceso $\{T_k = n\}$ equivale a observar $k-1$ éxitos en los $n-1$ primeros ensayos (sin especificar en qué posiciones), y un éxito en el n-ésimo ensayo. Por lo tanto, la función de probabilidad de T_k, evaluada en $n \geq k$ coincide con $P(N_{n-1} = k-1, X_n = 1)$. La independencia de los X_i implica que N_{n-1} (que depende de las variables $X_1, \ldots, X_{n-1}$) es independiente de X_n. Entonces:

$$\begin{aligned} P(T_k = n) &= P(\{N_{n-1} = k-1\} \cap \{X_n = 1\}) \\ &= P(N_{n-1} = k-1) \times P(X_n = 1) \\ &= \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \times p. \end{aligned}$$

Una comparación con (3.10.10) concluye la demostración.

Ejercicio: Determine si la distribución binomial negativa posee falta de memoria o no.

3.7.5. Distribución de Poisson

La distribución de Poisson es muy importante por si sola, como modelo probabilístico para recuentos. Por otra parte, ella se puede obtener como límite de la distribución Bin (n, p), para n grande, p peque no, y producto np moderado. Consideremos una serie de n ensayos de Bernoulli, donde la probabilidad de éxito varía con el número de ensayos n, y denotando a esta probabilidad por p_n, imponemos las condiciones

$$\lim_{n\to\infty} p_n = 0, \quad \lim_{n\to\infty} np_n = \lambda > 0.$$

Un ejemplo de esto es la extracción al azar, con reemplazo, de una muestra de tama no n, a partir de una población de tama no N. El número de veces X que aparece en la muestra una ficha predeterminada de la población, sigue una distribución Bin $(n, \frac{1}{N})$. Interesa la aproximación a $P(X = x)$ cuando $N \to \infty$, con $\frac{n}{N} \to \lambda > 0$.

Sea $X \sim$ Bin(n, p_n), con las características antes se naladas. Entonces:

$$\begin{aligned} p_X(k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= (k!)^{-1} n(n-1)\cdots(n-k+1) p_n^k (1-p_n)^{n-k} \\ &= (k!)^{-1} \prod_{i=1}^{k} (n-i+1) p_n \times (1-p_n)^{n-k} \end{aligned}$$

Es fácil ver que para cada $i = 1, \ldots, k$, se tiene que $\lim_{n\to\infty} (n-i+1)p_n = \lambda$, y que $\lim_{n\to\infty} (1-p_n)^{n-k} = e^{-\lambda}$, de modo que $\lim_{n\to\infty} p_X(k) = f(k)$, donde $f(y)$ es la función probabilidad de la distribución de parámetro λ, dada por (3.10.12).

La utilidad de esta aproximación a la distribución Binomial queda de manifiesto si consideramos que para valores grandes de n, el cálculo de probabilidades usando (3.10.7) es computacionalmente complicado, debido a la inestabilidad numérica de la fórmula.

> **Ejemplo 3.7.3** Suponga que sólo 2 de cada 1000 personas expuestas a un cierto virus desarrollan los síntomas que éste provoca. Si un grupo de 2500 personas son expuestas a este virus, ¿Cuál es la probabilidad que 5 o más de ellas desarrollen los síntomas correspondientes?
>
> Si denotamos por X el número total de personas que desarrollan los síntomas, entonces necesitamos $P(X \geq 5)$, o equivalentemente, $1 - \sum_{k=0}^{4} P(X = k)$. Si suponemos que estas personas se comportan independientemente, entonces $X \sim$ Bin$(2500, 0{,}002)$. Usando la aproximación de la distribución

de Poisson para este caso, concluimos que $X \sim$ Poisson(5), aproximadamente. Así,

$$p_X(k) = P(X = k) \approx \frac{5^k e^{-5}}{k!},$$

de modo que $p_X(0) = 0{,}0067$, $p_X(1) = 0{,}0337$, $p_X(2) = 0{,}0842$, $p_X(3) = 0{,}1404$, $p_X(4) = 0{,}1755$, y la probabilidad requerida es $P(X \geq 5) = 0{,}5595$. Considerando que el valor exacto es $0{,}5597$, la aproximación es muy buena.

3.8. Valores Esperados II

3.8.1. Valores Esperados en el Caso Continuo

La extensión a espacios muestrales no numerables o a variables aleatorias más generales, descansa en la idea que cualquier variable se puede aproximar adecuadamente por variables finitas. En el caso continuo, la idea consiste en considerar un intervalo peque no, digamos $[x, x+\Delta x]$, cuya probabilidad aproximada es $f_X(x)\Delta x$. Representando al intervalo por el punto x, el valor esperado de X corresponde a sumar elementos del tipo $x \times f_X(x)\Delta x$. Intuitivamente, la suma se convierte en integral. Esto motiva la siguiente definición:

Definición 3.8.1 Si X es una variable continua con valores en $\mathbb{R}$ y densidad f_X, el valor esperado de X está dado por:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx, \tag{3.8.1}$$

siempre que la integral impropia converja absolutamente, es decir, si

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty.$$

Ejemplo 3.8.1 Si $X \sim \Gamma(\alpha, \lambda)$, entonces

$$\begin{aligned} E(X) &= \int_0^{\infty} t \cdot \frac{t^{\alpha-1}\exp(-t/\lambda)}{\Gamma(\alpha)\lambda^{\alpha}} dt = \int_0^{\infty} \frac{t^{\alpha}\exp(-t/\lambda)}{\Gamma(\alpha)\lambda^{\alpha}} dt \\ &= \frac{\lambda\Gamma(\alpha+1)}{\Gamma(\alpha)} \int_0^{\infty} \frac{t^{(\alpha+1)-1}\exp(-t/\lambda)}{\Gamma(\alpha+1)\lambda^{\alpha+1}} dt = \frac{\lambda\Gamma(\alpha+1)}{\Gamma(\alpha)} \cdot 1 \\ &= \lambda\alpha, \end{aligned}$$

donde usamos que la integral de la densidad de la distribución $\Gamma(\alpha+1, \lambda)$ es 1. Si $\alpha = 1$, llegamos al caso de la distribución exponencial, en el que el valor esperado se reduce a λ.

Ejemplo 3.8.2 Si $X \sim U(a,b)$, entonces

$$E(X) = \int_a^b x \cdot \frac{1}{(b-a)} dx = \frac{b^2-a^2}{2(b-a)} = \frac{(a+b)}{2}.$$

Ejemplo 3.8.3 Si $X \sim \text{Beta}(a,b)$, entonces,

$$\begin{aligned} E(X) &= \int_0^1 x \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} dx = \int_0^1 \frac{x^{a+1-1}(1-x)^{b-1}}{B(a,b)} dx \\ &= \frac{B(a+1,b)}{B(a,b)} = \frac{\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}} = \frac{a\Gamma(a)\Gamma(a+b)}{(a+b)\Gamma(a)\Gamma(a+b)} \\ &= \frac{a}{a+b}. \end{aligned}$$

Ejemplo 3.8.4 Sea X una variable aleatoria con *distribución de Cauchy*, cuya densidad es

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R} \tag{3.8.2}$$

Entonces

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} dx \\ &= \lim_{x\to\infty} \frac{1}{2\pi} \log(1+x^2) - \lim_{y\to-\infty} \frac{1}{2\pi} \log(1+y^2), \end{aligned}$$

expresión que no existe, pues cada límite diverge a $+\infty$. Por lo tanto, X no tiene esperanza.

Ejemplo 3.8.5 Sea g es una función par, no negativa (o sea, $g(-z) = g(z)$) con $\int_0^\infty g(t)dt = 0{,}5$ y $\int_0^\infty tg(t)dt < \infty$. Entonces $f_X(x) = g(x-\theta)$ define una densidad de probabilidad, la distribución de X es simétrica en torno de θ, y $\mu = E(X) = \theta$. Para verificar la verdad de estas aseveraciones, basta plantear las integrales correspondientes, lo que se deja como ejercicio para el lector.

Un ejemplo importante es el de la distribución normal, para la cual

$$g(x) = \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

Así, si $X \sim N(\mu, \sigma^2)$, se sigue que $E(X) = \mu$.

Al igual que en el caso discreto, la esperanza de $Y = g(X)$ se puede calcular a partir de la distribución de X o de la distribución de Y. Si ambas variables son absolutamente continuas, se tiene el siguiente teorema:

Teorema 3.8.1 Sean X e Y variables aleatorias absolutamente continuas, tales que $Y = g(X)$. Entonces

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} g(x) f_X(x) dx, \tag{3.8.3}$$

siempre que una de las integrales converja absolutamente.

Ejemplo 3.8.6 Si $X \sim \text{Exp}(\lambda)$, entonces

$$\begin{aligned} E(X^k) &= \int_0^{\infty} \frac{x^k}{\lambda} \exp(-x/\lambda) dx \\ &= \Gamma(k+1)\lambda^k \int_0^{\infty} \frac{x^{(k+1)-1} \exp(x/\lambda)}{\Gamma(k+1)\lambda^{k+1}} dx \\ &= \Gamma(k+1)\lambda^k = k!\lambda^k \end{aligned}$$

Ejemplo 3.8.7 Si $X \sim N(0, \sigma^2)$, sabemos del Ejemplo 3.8.5 que $E(X) = 0$. Calculemos ahora $E(X^2)$. Se tiene que

$$E(X^2) = \int_{-\infty}^{\infty} \frac{x^2 e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dx.$$

Si $u = x$, y $dv/dx = xe^{-x^2/2\sigma^2}$, entonces usando integración por partes se obtiene:

$$E(X^2) = \sigma^2 \int_{-\infty}^{\infty} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} = \sigma^2.$$

De aquí se deduce que $Var(X) = \sigma^2$. Note también que si $X \sim N(\mu, \sigma^2)$, entonces

$$Var(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} \frac{(x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dx,$$

e introduciendo primero el cambio de variables $y = x - \mu$, se llega a que $Var(X) = \sigma^2$.

3.8.2. Función generadora de momentos

En el cálculo de los momentos de una distribución, la siguiente función, llamada *función generadora de momentos*, juega un importante rol.

Definición 3.8.2 La *función generadora de momentos* de la variable aleatoria X, se define como

$$M_X(t) = E(\exp(tX)), \tag{3.8.4}$$

para $t \in \mathbb{R}$ tal que el valor esperado correspondiente exista.

La importancia de la función generadora de momentos queda establecida en el siguiente resultado.

Teorema 3.8.2

(a) Si $\mu_k(X)$ existe para $k \in \{1, 2, 3, \ldots\}$, y si $\sum_{k=0}^{\infty} \mu_k t^k/k!$ converge absolutamente para $-h < t < h$ con $h > 0$, entonces $M_X(t)$ existe en $-h < t < h$, y

$$\mu_k(X) = M_X^{(k)}(0). \tag{3.8.5}$$

(b) Si $M_X(t)$ es expandible en serie de potencias infinita en una vecindad de $t = 0$, entonces $\mu_k(X)$ existe para todo $k \in \{1, 2, 3, \ldots\}$, y estos momentos se pueden calcular mediante (3.8.5).

Es este resultado el que origina el nombre de $M_X(t)$. Basta con que $M_X(t)$ sea expandible en serie de potencias infinita en una vecindad de $t = 0$, para que los momentos de X existan, caso en el que ellos se obtienen derivando la función y evaluándola en $t = 0$. La demostración de este resultado, se basa en desarrollos de Taylor de $M_X(t)$. De hecho, (3.8.5) dice que $\mu_k(X)$ es simplemente el coeficiente del término t^k en la expansión en serie de Taylor de $M_X(t)$ en torno a $t = 0$.

Veamos a continuación algunos ejemplos.

Ejemplo 3.8.8 Del Ejemplo 3.3.3 se deduce que la función generadora de momentos de una variable aleatoria $X \sim$ Poisson(λ) es

$$\exp(\lambda(\exp(t) - 1)),$$

la que está definida para cualquier real t, por lo que ella caracteriza la distribución Poisson(λ). Con un poco de paciencia, se obtiene que

$$\frac{d}{dt} M_X(t) = \lambda \exp(t) \exp\left(\lambda(\exp(t) - 1)\right)$$

$$\frac{d^2}{dt^2} M_X(t) = \lambda \exp(t) \exp\left(\lambda(\exp(t) - 1)\right) \left(1 + \lambda \exp(t)\right),$$

y aplicando (3.8.5), uno puede obtener que $E(X) = \lambda$ y $E(X^2) = \lambda(1 + \lambda)$, por lo que $Var(X) = \lambda$. Una alternativa es obtener las derivadas en el origen componiendo expansiones de Taylor truncadas. Así $\exp(z) \approx 1 + z + \frac{z^2}{2}$ implica

$$\begin{aligned} \exp(\lambda(\exp(t) - 1)) &\approx 1 + \lambda(t + \frac{t^2}{2}) + \lambda \frac{(t + \frac{t^2}{2})^2}{2} \\ &\approx 1 + \lambda(t + \frac{t^2}{2}) + \lambda^2 \frac{(t + \frac{t^2}{2})^2}{2} \\ &\approx 1 + \lambda t + (\lambda + \lambda^2)\frac{t^2}{2}. \end{aligned}$$

Identificando los coeficientes de t y de $\frac{t^2}{2}$ se obtiene EX, EX^2 y, de acá, $\text{Var } X = \lambda$.

Ejemplo 3.8.9 Sea $X \sim$N(0,1). Se tiene entonces que

$$\begin{aligned} M_X(t) &= E(\exp(tX)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tx) \exp(-x^2/2) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(x^2 - 2tx)) dx \\ &= \frac{\exp(t^2/2)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(x-t)^2) dx \\ &= \exp(t^2/2), \end{aligned}$$

la cual está definida para cualquier $t \in \mathbb{R}$. Puesto que

$$\begin{aligned} M_X(t) &= \exp(t^2/2) = \sum_{k=0}^{\infty} \frac{t^{2k}}{2^k k!} \\ &= 1 + \frac{t^2}{2} + \frac{t^4}{8} + \frac{t^6}{48} + \cdots + \frac{t^{2k}}{2^k k!} + \cdots \end{aligned}$$

no es difícil ver que

$$E(X^k) = M_X^{(k)}(0) = \begin{cases} 0 & \text{si } k \geq 1 \text{ es impar} \\ \frac{k!}{2^{k/2}(k/2)!} & \text{si } k \geq 2 \text{ es par} \end{cases}$$

Ejemplo 3.8.10 Sea $X \sim \Gamma(\alpha, \lambda)$. Tenemos entonces que

$$\begin{aligned} M_X(t) &= \int_0^{\infty} \exp(tx) f_X(x) dx \\ &= \frac{1}{\Gamma(\alpha)\lambda^{\alpha}} \int_0^{\infty} x^{\alpha-1} \exp(-x/(1/\lambda - t)^{-1}) dx \\ &= \frac{(1/\lambda - t)^{-\alpha}}{\lambda^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} \exp(-x/(1/\lambda - t)^{-1})}{\Gamma(\alpha)(1/\lambda - t)^{-\alpha}} dx \\ &= \frac{1}{(1 - t\lambda)^{\alpha}}, \end{aligned}$$

provisto que $t < \lambda^{-1}$. Puesto que

$$\frac{d}{dt} M_X(t) = \frac{\lambda\alpha}{(1 - t\lambda)^{\alpha+1}},$$

$$\frac{d^2}{dt^2} M_X(t) = \frac{\lambda^2\alpha(1+\alpha)}{(1 - t\lambda)^{\alpha+2}},$$

se tiene que $E(X) = \alpha\lambda$, y $E(X^2) = \lambda^2\alpha(1+\alpha)$, de modo que $Var(X) = \alpha\lambda^2$. El caso en que $X \sim$ Exp(λ) se obtiene de imponer $\alpha = 1$, con lo que $Var(X) = \lambda^2$.

Ejemplo 3.8.11 Sea $X \sim \text{Geom}(p)$. Entonces,

$$\begin{aligned} M_X(t) &= \sum_{k=1}^{\infty} \exp(tk)(1-p)^{k-1}p \\ &= p\exp(t)\sum_{k=1}^{\infty}\left((1-p)\exp(t)\right)^{k-1} \\ &= \frac{p\exp(t)}{1-(1-p)\exp(t)}, \end{aligned}$$

siempre que $p\exp(t) < 1$, esto es, $t < -\log(p)$. Por otra parte,

$$\frac{d}{dt}M_X(t) = \frac{p\exp(t)}{(1-(1-p)\exp(t))^2},$$

$$\frac{d^2}{dt^2} = \frac{p(\exp(t)+(1-p)\exp(2t))}{(1-(1-p)\exp(t))^3},$$

con lo que $E(X) = p^{-1}$, $E(X^2) = p^{-2}(2-p)$, y, finalmente, $Var(X) = p^{-2}(1-p)$.

3.8.3. Otras funciones generadoras

Aparte de la función generadora de momentos, existen otras funciones generadoras de interés.

Definición 3.8.3 Sea X una variable aleatoria. Se define, para el rango de valores en que el valor esperado correspondiente exista:

(a) la *función generadora de probabilidades* de X, denotada por $G_X(z)$ mediante

$$G_X(z) = E\left(z^X\right) \tag{3.8.6}$$

(b) la *función generadora de cumulantes* de X, denotada por $K_X(t)$ mediante

$$K_X(t) = \log(M_X(t)) \tag{3.8.7}$$

(c) la *función característica* de X, denotada por $\varphi_X(t)$ mediante

$$\varphi_X(t) = E\left(\exp(itX)\right) = E(\cos(tX)) + iE(\sin(tX)), \tag{3.8.8}$$

donde i es el número complejo $\sqrt{-1}$.

La función generadora de probabilidades se utiliza, casi exclusivamente, cuando la variable aleatoria toma valores enteros no negativos. En este caso, si z es tal que $G_X(z)$ existe, entonces

$$G_X(t) = \sum_{k=0}^{\infty} t^k p_X(k) = p_X(0) + p_X(1)t + p_X(2)t^2 + \cdots + p_X(k)t^k + \cdots, \tag{3.8.9}$$

lo cual coincide con la función $G(z)$ de la Definición 3.8.3. En otras palabras, ambas definiciones son equivalentes.

La ventaja de la función característica de X es que ella está siempre bien definida, cualquiera que sea el real t. La razón de ello es que $|E(\exp(itX))| \leq E|\exp(itX)| = 1$, para todo $t \in \mathbb{R}$, o bien usando el hecho que las funciones seno y coseno son acotadas. Es fácil ver que en la medida que las expresiones involucradas existan, se cumple que

$$\varphi_X(t) = M_X(it) = G_X(\exp(it)). \tag{3.8.10}$$

Por último, la función generadora de cumulantes está definida en el rango de valores para los que la función generadora de momentos existe. Como veremos a continuación, $K_X(t)$ genera los *cumulantes* de la distribución de X, definidos justamente como los coeficientes de la expansión en serie de Taylor de $K_X(t)$ en torno a $t = 0$.

Proposición 3.8.1 (Propiedades de las funciones generadoras)

(a) Sea X una variable aleatoria discreta con $\mathcal{X} \subset \{0, 1, 2, \ldots\}$, y para la cual $G_X(z)$ existe en una vecindad de $z = 0$. Entonces

$$p_X(k) = \frac{1}{k!}\frac{d^k}{dt^k}G_X(0) \tag{3.8.11}$$

(b) Si $K_X(t)$ se puede expandir mediante una serie de potencias infinita en una vecindad de $t = 0$, entonces todos los cumulantes $\kappa_k(X)$ existen y se calculan mediante:

$$\kappa_k(X) = \frac{d^k}{dt^k}K_X(0). \tag{3.8.12}$$

En particular,

$$\kappa_1(X) = E(X) \quad \text{y} \quad \kappa_2(X) = Var(X). \tag{3.8.13}$$

(c) Sean a, b reales cualesquiera. En la medida que las siguientes expresiones existan, se cumple:

1. $M_{a+bX}(t) = \exp(at)M_X(bt)$.
2. $G_{a+bX}(z) = z^a G_X(z^b)$.
3. $K_{a+bX}(t) = at + K_X(bt)$.
4. $\varphi_{a+bX}(t) = \exp(iat)\varphi_X(bt)$.

(d) *(Teorema de Caracterización):* Sean X e Y dos variables aleatorias.

1. Si $M_X(t) = M_Y(t)$ para todo $a < t < b$, entonces $F_X = F_Y$, esto es, X e Y tienen la misma distribución.
2. Si $G_X(z) = G_Y(t)$ para todo $a < z < b$, entonces $F_X = F_Y$.
3. Si $\varphi_X(t) = \varphi_Y(t)$ para todo $t \in \mathbb{R}$, entonces $F_X = F_Y$.

Demostración: La verificación de (a) es inmediata. Para obtener (b), note simplemente que

$$\frac{d}{dt}K_X(t) = \frac{M'_X(t)}{M_X(t)},$$

y que

$$\frac{d^2}{dt^2}K_X(t) = \frac{M''_X(t)M_X(t) - (M'_X(t))^2}{(M_X(t))^2},$$

de donde el resultado sale de evaluar lo anterior en $t = 0$. Por otra parte,

$$\begin{aligned} M_{a+bX}(t) &= E(\exp(t(a+bX))) = E(\exp(at)\exp(btX)) = \\ &= \exp(at)E(\exp(btX)) = \exp(at)M_X(bt), \end{aligned}$$

y las otras tres propiedades se prueban en forma similar. Finalmente, la prueba de (d) será omitida. ■

Ejemplo 3.8.12 Sea $X \sim$N(0,1), y defina $Y = \mu + \sigma X$, donde $\mu \in \mathbb{R}$, y $\sigma \neq 0$. Entonces:

$$M_Y(t) = M_{\mu+\sigma X}(t) = \exp(t\mu + \sigma^2 t^2/2). \tag{3.8.14}$$

Por otra parte, si $\sigma > 0$:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\mu + \sigma X \leq y) = P(X \leq (x-\mu)/\sigma) \\ &= F_X\left(\frac{x-\mu}{\sigma}\right), \end{aligned}$$

de donde, mediante diferenciación se obtiene que

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

para $y \in \mathbb{R}$, y se concluye que $Y \sim$N(μ, σ^2), y su función generadora de momentos está dada por (3.8.14). Note que

$$K_Y(t) = \log(M_Y(t)) = \mu t + \frac{\sigma^2 t^2}{2},$$

de donde se deduce que $E(Y) = K'_Y(0) = \mu$, y $Var(Y) = K''_Y(0) = \sigma^2$. Finalmente,

$$\varphi_Y(t) = \exp(i\mu t - t^2/2)$$

es la función característica de la distribución N(μ, σ^2).

Ejemplo 3.8.13 Sea X con densidad triangular

$$f_X(x) = \begin{cases} 1 - |x| & \text{si } |x| \leq 1 \\ 0 & \text{si no} \end{cases}$$

Es fácil ver que $\mu_k(X)$ debe existir para todo $k \geq 1$, pues X tiene un rango de valores acotado. Así, $M_X(t)$ también existe para cualquier t, y

$$\begin{aligned} M_X(t) &= E(\exp(tX)) = \int_{-1}^{1} \exp(tx)(1-|x|)dx \\ &= \int_{-1}^{0} \exp(tx)(1+x)dx + \int_{0}^{1} \exp(tx)(1-x)dx \\ &= \frac{\exp(t)+\exp(-t)-2}{t^2}. \end{aligned}$$

Note que, de acuerdo a la expresión obtenida, $M_X(t)$ no está definida en $t = 0$. Sin embargo, observe que del desarrollo en serie de Taylor de $\exp(t)$ y $\exp(-t)$ se concluye que

$$\begin{aligned} M_X(t) &= t^{-2}\left(\sum_{k=0}^{\infty}\frac{t^k}{k!} + \sum_{k=0}^{\infty}\frac{(-1)^k t^k}{k!} - 2\right) \\ &= t^{-2}\left(2\frac{t^2}{2!} + 2\frac{t^4}{4!} + 2\frac{t^6}{6!} + \cdots + 2\frac{t^{2k}}{(2k)!}\right) + \cdots \\ &= 1 + \frac{2t^2}{4!} + \frac{2t^4}{6!} + \cdots + \frac{2t^{2k-2}}{(2k)!} + \cdots \end{aligned}$$

por lo que $\mu_{2k-1}(X) = 0$, y $\mu_{2k}(X) = 2(2k+1)^{-1}(2k+2)^{-1}$, es decir,

$$\mu_k(X) = \begin{cases} 0 & \text{si } k \text{ es impar} \\ \frac{2}{(k+1)(k+2)} & \text{si } k >= 2 \text{ es par} \end{cases}$$

Por otra parte, note que $\exp(it) = \cos(t) + i\sin(t)$, y que $\exp(-it) = \cos(t) - i\sin(t)$, por lo que

$$\varphi_X(t) = \frac{2(\cos(t)-1)}{t^2},$$

y la función característica de X es una función a valores reales. No es difícil darse cuenta que este será siempre el caso cuando la distribución de la variable aleatoria en cuestión sea simétrica con respecto al origen. En este caso, $f_X(x) = f_X(-x)$ lo que implica la simetría. La demostración de este resultado se propone como ejercicio.

3.9. Transformaciones de Variables Aleatorias Continuas

3.9.1. El caso biyectivo

El caso discreto es, en general, simple y directo de resolver. Para derivar el resultado en el caso continuo, observe que si g es monótona creciente y diferenciable, entonces, podemos obtener la densidad de $Y = g(X)$ como sigue. La función de distribución acumulada de Y es, por definición, $F_Y(y) = P(Y \leq y)$, y tenemos que:

$$F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)),$$

y entonces

$$\begin{aligned} f_Y(y) &= F_Y'(y) = F_X'(g^{-1}(y)) = f_X(g^{-1}(y))(g^{-1})'(y) \\ &= f_X(g^{-1}(y)) \cdot \frac{1}{g'(g^{-1}(y))}, \end{aligned}$$

donde $\mathcal{Y} = g(\mathcal{X})$.

Cuando g es monótona decreciente, el mismo argumento se puede aplicar, después de ligeras modificaciones. En efecto, el evento $\{g(X) \leq y\}$ equivale ahora al evento $\{X \geq g^{-1}(y)\}$, pues g es decreciente, y entonces $F_Y(y) = 1 - F_X(g^{-1}(y))$. Finalmente, se obtiene que

$$f_Y(y) = -f_X(g^{-1}(y)) \cdot \frac{1}{g'(g^{-1}(y))}.$$

Observe que g' es una función negativa, de modo que el resultado es una función positiva, después de incorporar el signo negativo.

Finalmente, podemos resumir las fórmulas observadas en el siguiente resultado.

Teorema 3.9.1 Sea X una variable aleatoria con densidad f_X, y sea $Y = g(X)$, donde g es monótona y diferenciable. Entonces

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|} \tag{3.9.1}$$

Ejemplo 3.9.1 Suponga que $X \sim N(0,1)$, y sea $Y = \mu + \sigma X$, con $\mu \in \mathbb{R}$ y $\sigma > 0$. En el Ejemplo 3.8.12 se obtuvo que $Y \sim N(\mu, \sigma^2)$ mediante propiedades de funciones generadoras. El mismo resultado se obtiene usando el Teorema 3.9.1 con $g(x) = \mu + \sigma x$, que claramente cumple las hipótesis de dicho resultado. Así, $g^{-1}(y) = (y-\mu)/\sigma$, $g'(x) = \sigma$, y la densidad de Y se obtiene de (3.9.1):

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad y \in \mathbb{R}. \tag{3.9.2}$$

Es interesante notar que de la misma forma se obtiene que si $Y \sim N(\mu, \sigma^2)$, entonces $Z = (Y-\mu)/\sigma \sim N(0,1)$, proceso que recibe el nombre de *estandarización*. Note además que puesto que $P(Y \leq y) = P(\frac{Y-\mu}{\sigma} \leq \frac{y-\mu}{\sigma})$ se concluye que

$$P(Y \leq y) = \Phi\left(\frac{y-\mu}{\sigma}\right), \tag{3.9.3}$$

por lo que probabilidades relativas a una variable aleatoria con distribución normal cualquiera se pueden obtener a partir de la distribución normal estándar. Por ejemplo, si $Y \sim N(3,4)$, entonces

$$\begin{aligned} P(Y > 5) &= 1 - P(Y \leq 5) = 1 - P\left(\frac{Y-3}{2} \leq \frac{5-3}{2}\right) \\ &= 1 - P(Z \leq 1) = 1 - \Phi(1) = 1 - 0{,}841 \\ &= 0{,}159 \end{aligned}$$

Ejemplo 3.9.2 Suponga que X tiene densidad dada por

$$f_X(x) = \begin{cases} 3x^2 & \text{si } 0 < x < 1 \\ 0 & \text{si no,} \end{cases}$$

y considere $g(x) = 2x$. Se tiene que $\mathcal{Y} = (0, 2)$, que g es claramente monótona creciente, con $g^{-1}(y) = y/2$, y $g'(x) = 2$. Usando (3.9.1) es inmediato obtener que si $0 < y < 2$ entonces

$$f_Y(y) = 3 \cdot (y/2)^2 \cdot 1/2 = (3/8)y^2.$$

Finalmente, se obtiene que

$$f_Y(y) = \begin{cases} (3/8)y^2 & \text{si } 0 < y < 2 \\ 0 & \text{si no} \end{cases}$$

Ejemplo 3.9.3 Suponga que $X \sim U(0, 1)$, y considere $g(x) = -\log(x)$, definida sobre los reales positivos. Obtengamos f_Y para $Y = -\log(X)$. Se tiene que $\mathcal{Y} = (0, \infty)$, $g^{-1}(y) = \exp(-y)$, $g'(x) = -x^{-1}$, y g es monótona decreciente. Entonces, usando (3.9.1) se obtiene que, puesto que $f_X(x) = I_{(0,1)}(x)$,

$$f_Y(y) = \exp(-y),$$

y entonces $Y \sim \text{Exp}(1)$.

Ejemplo 3.9.4 Suponga que X tiene densidad triangular en el intervalo $[0, 1]$, esto es, $f_X(x) = c(1 - |1 - 2x|)$, para algún valor adecuado de c. Obtengamos la densidad de $Y = X^2$, con lo que $\mathcal{Y} = [0, 1]$. Primeramente, se debe calcular el valor de c. Note que

$$\int_0^1 (1 - |1 - 2x|)dx = \int_0^{1/2} 2xdx + \int_{1/2}^1 2(1 - x)dx = 1/4 + 1/4 = 1/2,$$

de modo que $c = 2$. Ahora, en [0,1], la función $g(x) = x^2$ es creciente y diferenciable, con $g^{-1}(y) = \sqrt{y}$, $g'(x) = 2x$, y entonces, por (3.9.1):

$$f_Y(y) = \frac{2(1 - |1 - 2\sqrt{y}|)}{2\sqrt{y}} = \frac{1 - |1 - 2\sqrt{y}|}{\sqrt{y}},$$

para $y \in [0, 1]$. Ver Figura 3.9.6.

Ejemplo 3.9.5 Sea $X \sim$N(μ, σ^2), y considere $Y = \exp(X)$. La distribución de Y se conoce como *distribución log-normal*, con parámetros μ y σ^2. La densidad de Y se obtiene de aplicar (3.9.1) con $g(x) = \exp(x)$. Puesto que $g^{-1}(y) = \log(y)$ tenemos que para $y > 0$:

$$f_Y(y) = f_X(g^{-1}(y)) \frac{1}{|g'(g^{-1}(y))|},$$

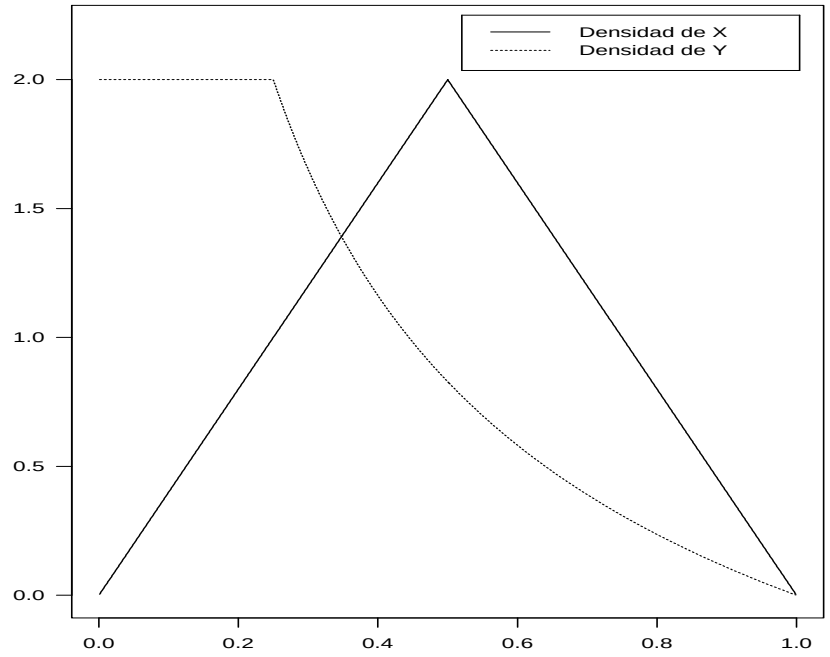


Figura 3.9.6: Densidad triangular f_X y densidad f_Y de la transformación $Y = X^2$.

de donde

$$f_Y(y) = \begin{cases} (y\sqrt{2\pi\sigma^2})^{-1} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right) & \text{si } y > 0 \\ 0 & \text{si no} \end{cases}$$

El cálculo de momentos de Y es complicado si se hace por definición. Sin embargo, note que

$$\begin{aligned} \mu_k(Y) &= E(Y^k) = E(\exp(kX)) = M_X(k) \\ &= \exp\left(k\mu + \frac{k^2\sigma^2}{2}\right). \end{aligned}$$

En particular, $E(Y) = \exp(\mu + \sigma^2/2)$, y $E(Y^2) = \exp(2\mu + 2\sigma^2)$, por lo que $Var(Y) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$.

El resultado de (3.9.1) se puede generalizar de la siguiente manera.

Teorema 3.9.2 Sea X es una variable aleatoria definida sobre $\mathcal{X}$, con densidad f_X, y sea $g : \mathcal{X} \to \mathcal{Y}$ biyectiva, diferenciable y tal que $(g^{-1})'$ es no nulo sobre $\mathcal{Y}$. Entonces la densidad de $Y = g(X)$ está dada por

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1})'(y)|I_{y\in\mathcal{Y}} = f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|} I_{\mathcal{Y}}(y) \tag{3.9.4}$$

3.9.2. El caso continuo no biyectivo

Cuando g no es biyectiva, los resultados recién vistos no tienen validez. Sin embargo, hay veces en que $\mathcal{X}$ se puede particionar de modo que g es biyectiva en cada una de esas porciones. Así, el teorema es válido en cada elemento de dicha partición, y se puede demostrar que la expresión final de la densidad se obtiene de sumar cada una de las densidades restringidas. Este resultado se enuncia a continuación.

Teorema 3.9.3 Sea X una variable aleatoria con densidad f_X definida sobre $\mathcal{X}$, y sea g una función definida sobre $\mathcal{X}$, verificando la propiedad que existe una partición $A_1, A_2, \ldots$ de $\mathcal{X}$ tal que g_i, definida como la restricción de g a A_i es biyectiva y diferenciable. Entonces, $Y = g(X)$ tiene densidad f_Y dada por

$$f_Y(y) = \sum_{i=1}^{\infty} f_X(g_i^{-1}(y))|(g_i^{-1})'(y)|I_{g_i(A_i)}(y) = \sum_{i=1}^{\infty} \frac{f_X(g_i^{-1}(y))}{|g_i'(g_i^{-1}(y))|} \cdot I_{g_i(A_i)}(y) \tag{3.9.5}$$

Ejemplo 3.9.6 Suponga $X \sim N(0,1)$, y calculemos la densidad de $Y = g(X) = X^2$. Aquí $\mathcal{X} = \mathbb{R}$, y es claro que g no es biyectiva en ese dominio. Sin embargo, uno puede descomponer $\mathbb{R}$ en dos partes, los reales no negativos, y los reales negativos. El lugar específico del punto $x = 0$ carece de importancia, y, más aún, se puede eliminar si así uno lo desea. De hecho, este es el procedimiento si la diferenciabilidad de g no se tiene en algún punto particular. Sea, entonces, $A_1 =]-\infty, 0[$ y $A_2 = [0, \infty[$. Es claro que g restringida a cualquiera de A_1 o A_2 es biyectiva y diferenciable, esto es, A_1 y A_2 cumplen las hipótesis del teorema. Por último, nada impide usar sólo dos conjuntos, en vez de una partición numerable pero infinita. Luego, tenemos que $g_1(x) = g_2(x) = x^2$, pero $g_1^{-1}(y) = -\sqrt{y}$, $g_2^{-1}(y) = \sqrt{y}$, y $g_i'(x) = 2x$, para $i = 1, 2$. Por otra parte,

$$f_X(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}.$$

Entonces, el primer término de la suma, correspondiente al aporte de la restricción a números negativos, es

$$\frac{\exp(-y/2)}{\sqrt{2\pi} \cdot 2\sqrt{y}},$$

y es fácil ver que el otro término es idéntico, por lo que al sumar se obtiene:

$$f_Y(y) = \frac{\exp(-y/2)}{\sqrt{2\pi}\sqrt{y}} = \frac{y^{1/2-1}\exp(-y/2)}{\sqrt{\pi}2^{1/2}},$$

y el lector podrá deducir que $Y \sim \text{Gama}(1/2, 2)$, quedando sólo por comprobar el hecho que $\Gamma(1/2) = \sqrt{\pi}$, lo cual queda propuesto como un ejercicio.

Ejemplo 3.9.7 Sea $X \sim \text{Exp}(\lambda)$ con $\lambda > 0$, y considere $Y = \cos(X)$. Es claro que $\mathcal{Y} = [-1, 1]$. Por otra parte, la función $g(x) = \cos(x)$ es claramente no biyectiva en $(0, \infty)$. Considere entonces los conjuntos $A_k = (k\pi, (k+1)\pi)$, para $k = 0, 1, 2, \ldots$.

Entonces g_k, definida como la restricción de g al conjunto A_k es biyectiva y con inversa continuamente diferenciable. Además, $P(X \in \bigcup_{k=0}^{\infty} A_k) = 1$. No es difícil ver que $g_0^{-1}(y) = \arccos(y)$, $g_1^{-1}(y) = 2\pi - \arccos(y)$, $g_2^{-1}(y) = 2\pi + \arccos(y)$, y así sucesivamente. En general, para $-1 < y < 1$ se tiene que

$$g_k^{-1}(y) = \begin{cases} k\pi + \arccos(y) & \text{si } k = 0, 2, 4, 6, \ldots \\ (k+1)\pi - \arccos(y) & \text{si } k = 1, 3, 5, 7, \ldots \end{cases}$$

Así, para cualquier $k = 0, 1, 2, \ldots$ se tiene

$$\frac{1}{g_k'(g_k^{-1}(y))} = (g_k^{-1}(y))' = -\frac{1}{\sqrt{1-y^2}},$$

de modo que aplicando (3.9.4) tenemos que si $-1 < y < 1$:

$$\begin{aligned} f_Y(y) &= \sum_{k=0}^{\infty} f_X(g_k^{-1}(y)) \cdot \frac{1}{|g_k'(g_k^{-1}(y))|} \\ &= \frac{1}{\lambda\sqrt{1-y^2}} \left(\sum_{k=1,3,5,\ldots} \exp(-\lambda^{-1}\{(k+1)\pi - \arccos(y)\}) \right. \\ &\quad \left. + \sum_{k=0,2,4,\ldots} \exp(-\lambda^{-1}\{k\pi + \arccos(y)\}) \right) \\ &= \frac{\exp(-\arccos(y)\lambda^{-1})}{\lambda\sqrt{1-y^2}} \sum_{k=0}^{\infty} \exp(-2\pi\lambda^{-1})^k \\ &\quad + \frac{\exp((\arccos(y) - \pi)\lambda^{-1})}{\lambda\sqrt{1-y^2}} \sum_{k=0}^{\infty} \exp(-2\pi\lambda^{-1})^k \\ &= \frac{\exp(-\lambda^{-1}\arccos(y)) + \exp(\lambda^{-1}(\arccos(y) - \pi))}{\lambda\sqrt{1-y^2}(1 - \exp(-2\pi\lambda^{-1}))}. \end{aligned}$$

Finalmente, es claro que $f_Y(y) = 0$ si $y \notin (-1, 1)$.

3.10. Resumen de Principales Distribuciones Univariadas

Para terminar este capítulo, entregamos a continuación un listado de algunas de las principales familias paramétricas de distribuciones. La mayoría de ellas se usa en diversas partes de este texto.

3.10.1. Algunas funciones de probabilidad discretas

$f(1) = p, f(0) = q = 1 - p$	
Distribución de Bernoulli	Bern (p)

(3.10.6)

$$\boxed{\begin{array}{l} f(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, \ldots, n \\ \\ \text{Distribución Binomial} \qquad \text{Bin } (n, p) \end{array}} \tag{3.10.7}$$

$$\boxed{\begin{array}{l} f(y) = p(1-p)^{y-1} \text{ para } y = 1, 2, \ldots \\ \\ \text{Distribución Geométrica} \qquad \text{Geom}(p) \end{array}} \tag{3.10.8}$$

$$\boxed{\begin{array}{l} f(y) = p(1-p)^{y} \text{ para } y = 0, 1, 2, \ldots \\ \\ \text{Distribución Geométrica} \\ \quad \text{trasladada al origen} \qquad \text{Geom}(p) \end{array}} \tag{3.10.9}$$

$$\boxed{\begin{array}{l} f(y) = \binom{y-1}{k-1} p^k (1-p)^{y-k} \text{ para } y = k, k+1, k+2, \ldots \\ \\ \text{Distribución Binomial negativa} \qquad \text{BN}(k, p). \end{array}} \tag{3.10.10}$$

$$\boxed{\begin{array}{l} f(y) = \frac{k(k+1)\cdots(k+y-1)}{y!} p^k q^y, \ y = 0, 1, 2, \ldots \\ \\ \quad = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} p^k q^y, \ y = 0, 1, 2, \ldots \\ \\ \text{Distribución Binomial negativa} \\ \quad \text{trasladada al origen} \qquad \text{BN0}(k, p). \end{array}} \tag{3.10.11}$$

$$\boxed{\begin{array}{l} f(y) = \frac{\lambda^k e^{-\lambda}}{k!}, \text{para } k = 0, 1, \ldots. \\ \\ \text{Distribución de Poisson} \qquad \text{Poisson}(\lambda) \end{array}} \tag{3.10.12}$$

3.10.2. Algunas funciones densidad continuas

$$\boxed{\begin{array}{l} f(y) = \frac{1}{b-a}, a < y < b \\ \text{Distribución Uniforme en} [a, b]] \qquad U[a, b] \end{array}} \tag{3.10.13}$$

$$\boxed{\begin{array}{l} f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \ y^{\alpha-1}(1-y)^{\beta-1}, \ 0 < y < 1. \\ \\ \text{Distribución Beta} \qquad \text{Beta}(\alpha, \beta) \\ \hline \textbf{Caso especial:} \quad \alpha = \beta = 1 : \text{Distribución} \qquad U[0, 1] \end{array}} \tag{3.10.14}$$

$$f(y) = \lambda(\lambda y)^{\alpha-1} e^{-\lambda y} \frac{1}{\Gamma(\alpha)}, \quad y > 0$$

Distribución Gama (α, λ)

Casos especiales:	
$\alpha = \frac{\nu}{2}, \lambda = \frac{1}{2}$	Distribución Ji cuadrado con ν grados de libertad $\chi^2(\nu)$
$\alpha = 1$	Distribución Exponencial Expo (λ). $f(y) = \lambda e^{-\lambda y}, \quad y > 0$

(3.10.15)

$$f(y) = \lambda\beta y^{\beta-1} e^{-\lambda y^{\beta}}, y > 0$$

Distribución de Weibull Weib(λ, β)

Caso especial $\beta = 1$: Distribución Expo (λ).

(3.10.16)

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \ -\infty < y < \infty$$

Distribución Normal $N(\mu, \sigma^2)$

(3.10.17)

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\frac{1}{2})\sqrt{\nu}} (1 + \frac{x^2}{\nu})^{-\frac{(\nu+1)}{2}} \ -\infty < x < \infty$$

Distribución de Student con ν grados de libertad $t(\nu)$.

(3.10.18)

$$f(y) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2}) \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \frac{y^{\frac{\nu_1}{2}-1}(\nu_1 y + \nu_2)^{\frac{\nu_1+\nu_2}{2}}}{,} \ -\infty < y < \infty$$

Distribución F con ν_1 y ν_2 grados de libertad $F(\nu_1, \nu_2)$.

(3.10.19)

3.11. Problemas

1. En un experimento se observa la temperatura $X = x$ en grados Celsius. Esta temperatura sigue una distribución de probabilidad con función de distribución acumulada F y densidad f. Suponga ahora que se cambia la escala del instrumento, de modo que el resultado y queda expresado en grados Fahrenheit. Denotemos la distribución acumulada de esta temperatura por G, y su densidad de probabilidad por g.

   (a) Demuestre que $G(y) = F(\frac{y-32}{1,8})$.

   (b) Demuestre que $g(y) = \frac{1}{1,8} f(\frac{y-32}{1,8})$.

   (c) Aplique los resultados anteriores al caso $F(x) = 1 - e^{-\lambda x}$, con $x > 0$.

2. Un dado equilibrado se lanza cuatro veces. Sea X el mínimo número que se obtiene.

   (a) Encuentre la distribución de X.

   (b) Calcule $E(X)$ y $\mathrm{Var}(X)$.

3. Calcule el número esperado de tréboles que se obtienen en una mano de poker, consistente en 5 cartas escogidas al azar de un total de 52.

4. Sea X el número de aciertos en una cartilla de LOTO. Calcule $E(X)$ y $Var(X)$.

5. En una secuencia de ensayos de Bernoulli, sea X el número necesario de intentos requeridos para obtener al menos un éxito y un fracaso.

   (a) Calcule $E(X)$ y $Var(X)$.

   (b) Calcule la función generadora de momentos de X, y repita (a) usando dicha función.

6. Un dado no equilibrado asigna a la cara con el número x probabilidades dadas por $p(x) = c \times 0{,}7^x \times 0{,}3^{6-x}$, $x = 1, 2, 3, 4, 5, 6$.

   (a) Calcule el valor de c.

   (b) Haga una tabla con los valores de la función de distribución F.

   (c) Utilice la tabla para calcular la probabilidad que

      (i) El número esté entre 2 y 4.

      (ii) El número sea mayor que 2.

7. El tiempo entre dos terremotos consecutivos tiene densidad

$$f_k(x) = cx^k e^{-x},\ x > 0.$$

   (a) Demuestre que $c = k!$.

   (b) Obtenga la función de distribución acumulada F_k. (Integre por partes y use inducción).

   (c) Haga una tabla con los valores de F_3, evaluándola para múltiplos de 0.5 entre 0 y 8.

   (d) Utilice la tabla obtenida en (c) para calcular la probabilidad que el tiempo x : (i) sea inferior a 4 a nos. (ii) esté comprendido entre 2.5 y 3.5 a nos. (iii) Exceda los 5 a nos.

(e) De todos los intervalos de la forma $[0{,}5j, 0{,}5j + 0{,}5]$, $j = 0, 1, \ldots 15$, encuentre aquél que tiene la máxima probabilidad.

8. La probabilidad que el número de personas en una fila sea k está dada por el coeficiente de z^k en el desarrollo en serie de Taylor de $(q + pz)^{-2}$.

   (a) Demuestre que para que este modelo probabilístico tenga sentido es necesario que $q + p = 1$.

   (b) Obtenga la función de distribución acumulada F.

   (c) Construya una tabla para $p = \frac{1}{2}$.

9. La proporción de calcio en un mineral es altamente variable. La probabilidad que esta proporción esté entre a y b es $\int_a^b f(x)dx$, con

$$f(x) = c_k x^k (1 - x)^k, \quad 0 < x < 1,$$

   y con $k = 0, 1, 2$.

   (a) Encuentre c_k.

   (b) Calcule la función de distribución acumulada $F_k(x)$.

   (c) Evalúe la probabilidad π_k que la proporción esté entre 0.25 y 0.75.

   (d) Conjeture el comportamiento de π_k a medida que k crece.

10. Sea X una variable aleatoria continua con función de densidad $f > 0$. Si F es la función de distribución de X pruebe que la variable $Y = F(X)$ tiene distribución uniforme en [0,1].

11. Si $X \sim U(0, 1)$ encontrar la función densidad de $Y = e^X$.

    $Resp.: f_Y(t) = \frac{1}{t}$ si $1 < t < e$

12. Si $Y \sim U(0, 5)$, ¿cuál es la probabilidad que las raíces de la ecuación $4x^2 + 4xY + Y + 2 = 0$ sean ambas reales?.

    $Resp.: \frac{3}{5}$

13. Si un proyectil se lanza en un ángulo $\theta \sim U(0, \frac{\pi}{4})$ de la tierra con una velocidad v, éste caerá al suelo a una distancia R que puede ser expresada por $R = (\frac{v^2}{g})(\sin 2\theta)$, donde g es la aceleración de gravedad. Encontrar la función de distribución de R.

    $Resp.: F(x) = \frac{2}{\pi} \arcsin \frac{gx}{v^2}$ para $0 \leq x \leq \frac{v^2}{g}$

14. Un entero positivo I es seleccionado con $P(I = n) = \frac{1}{2^n}$ para $n = 1, 2, ...$ Si el entero es n, se lanza una moneda al aire en que la probabilidad de obtener una cara es e^{-n}. ¿Cuál es la probabilidad que al lanzar la moneda obtengamos una cara?.

    $Resp: \frac{1}{2e-1}$

15. Se lanza una moneda en que la probabilidad de obtener una cara es $p = \frac{1}{2}$, y suponga que la moneda se lanza repetidamente. Sea X_n el número total de caras que han sido obtenidas en los primeros n lanzamientos y sea $Y_n = n - X_n$. Supongamos que paramos los lanzamientos

cuando se obtiene el primer n tal que $X_n = Y_n + 3$ o $Y_n = X_n + 3$. Determine la probabilidad que $X_n = Y_n + 3$ cuando se detienen los lanzamientos.

$Resp$: $\frac{1}{2}$

16. Considere un elevador que comienza en el subterráneo de un edificio y viaja hacia arriba. Sea N_i el número de personas que suben al elevador en el piso i. Suponga que los N_i son independientes y que $N_i \sim$ Poisson(λ_i). Cada persona que sube en i, independiente del resto sale en j con probabilidad p_{ij}. Sea N_{ij} el número de personas que suben al elevador en el piso i y bajan en el j. Calcule $P(N_{ij} = k)$.

    $Resp$: $N_{ij} \sim$ Poisson$(\lambda_i p_{ij})$.

17. Suponga que $N_1 \sim$ Poisson(λ_1), $N_2 \sim$ Poisson(λ_2) donde N_1 y N_2 son independientes. Pruebe que $N_1 + N_2 \sim$ Poisson$(\lambda_1 + \lambda_2)$, y además calcule $P(N_1 = 1 | N_1 + N_2 = 1)$.

    $Resp$: $\frac{\lambda_1}{\lambda_1 + \lambda_2}$

18. (a) Sean $X_1, \ldots, X_n$ variables aleatorias independientes, y defina las nuevas variables aleatorias Y y Z mediante $Y = \min(X_1, \ldots, X_n)$, $Z = \max(X_1, \ldots, X_n)$. Argumente que las siguientes relaciones son verdaderas:

$$\begin{aligned} P(Y > y) &= P(X_1 > y) \cdots P(X_n > y) \\ P(Z \leq z) &= P(X_1 \leq z) \cdots P(X_n \leq z). \end{aligned}$$

    (b) Asuma que los tiempos de falla de un sistema de n componentes son $T_1, \ldots, T_n$, los que se suponen independientes. Lo que nos interesa calcular es la distribución del tiempo de falla T del sistema completo en términos de las distribuciones de $T_1, \ldots, T_n$. Aplique la parte (a) a lo siguiente:

        (i) Si de 10 componentes cada una tiene probabilidad 0.99 de durar al menos 100 horas, y éstas se encuentran en serie, ¿cuál es la probabilidad que el sistema no fallará en 100 horas?.

        (ii) ¿Cuál es la probabilidad, si ahora las componentes están en paralelo?. $Resp$: $1 - 10^{-20}$.

19. Se lanzan dos dados perfectos. Sea X igual al producto de los valores obtenidos en los dados. Determine $\mathcal{X}$, y calcule $P(X = x)$ para $x \in \mathcal{X}$.

20. Suponga que un dado se lanza dos veces. ¿Cuáles son los posibles valores que pueden tomar las siguientes variables aleatorias?

    (a) El máximo valor en los dos lanzamientos.

    (b) El mínimo valor en los dos lanzamientos.

    (c) La suma de los dos lanzamientos.

    (d) El valor del primer lanzamiento menos el valor del segundo lanzamiento.

21. Calcule el valor esperado y varianza en cada una de las partes del Problema 20.

22. Compare la aproximación de Poisson con la probabilidad Binomial correcta para los siguientes casos:

    (a) $P(X = 2)$ cuando $n = 8$, $p = 0{,}1$.
    (b) $P(X = 9)$ cuando $n = 10$, $p = 0{,}95$.
    (c) $P(X = 0)$ cuando $n = 10$, $p = 0{,}1$.
    (d) $P(X = 4)$ cuando $n = 9$, $p = 0{,}2$.

23. El número de suicidios en cierto estado es de 1 por cada 100.000 habitantes en un mes.

    (a) Encontrar la probabilidad que en una ciudad de 400.000 habitantes del mismo estado, se produzcan por lo menos ocho suicidios.
    (b) ¿Cuál es la probabilidad que durante dos meses del a no ocurran ocho o más suicidios?.
    (c) Contando el presente mes como el mes número uno, ¿cuál es la probabilidad que en el mes i ocurran ocho o más suicidios?.

    ¿Que supuestos se deben hacer?.

24. Cada caja de una cierta marca de cereal contiene un animalito de plástico en su interior. Hay un total de N posibles animalitos disponibles, y suponga que es igualmente probable encontrar uno cualquiera de ellos en una caja dada. Determine el número esperado de cajas que se debe comprar para obtener la colección completa de animalitos.

25. Una urna contiene n bolas numeradas $1, 2, \ldots, n$. Una persona extrae al azar una bola de la urna y la devuelve, saca otra y la devuelve, continuando hasta sacar una misma bola *por segunda vez*. Sea X el número de intentos necesarios para obtener dicha repetición.

    (a) Obtenga la distribución de X. (Indicación: calcule $P(X > k)$)
    (b) Demuestre que

$$E(X) = 2 + \left(1 - \frac{1}{n}\right) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots \prod_{i=1}^{n-1}\left(1 - \frac{i}{n}\right).$$

26. Pruebe que si X es una variable aleatoria cualquiera tal que $P(X \in [a, b])$, entonces $a \leq E(X) \leq b$ y $Var(X) \leq (b - a)^2/4$. (Indicación: haga primero el caso $a = 0$, $b = 1$). Encuentre una variable aleatoria que alcance la máxima varianza.

27. Sea X una variable aleatoria con distribución $U(0, 1)$, y defina $Y = \min\{X, c\}$, donde $0 < c < 1$. Calcule $E(Y)$ y $Var(Y)$.
    Nota:

$$Y(\omega) = \begin{cases} X(\omega) & \text{si } X(\omega) \leq c \\ c & \text{si no} \end{cases}$$

28. El tiempo de vida en horas de un tubo fluorescente, es una variable aleatoria que tiene una densidad de probabilidad dada por:

$$f(x) = \alpha^2 x e^{-\alpha x} \qquad x \geq 0.$$

    Calcule el tiempo de vida esperado del tubo.

29. Sea X una variable aleatoria con densidad

$$f_X(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}, \quad x \in \mathcal{R}.$$

(a) Pruebe que la distribución de X es simétrica en torno de 0.

(b) Determine si $E(X)$ existe, y calcule su valor en caso afirmativo.

(c) Obtenga la densidad de $Y = \exp(X)$, y determine $E(Y)$.

(d) Calcule $E(Y)$ esta vez sin utilizar su densidad.

30. Se dice que X tiene distribución de Weibull si

$$f_X(x) = \begin{cases} \lambda\alpha x^{\alpha-1}\exp(-\lambda x^\alpha) & \text{si } x > 0 \\ 0 & \text{si no.} \end{cases}$$

Se asume que $\alpha > 0$ y $\lambda > 0$. Determine $E(X)$. ¿Cuál es la distribución de $Y = X^\alpha$?

31. Encuentre la función generadora de momentos de una variable aleatoria $X \sim U(a,b)$. Use este resultado para calcular $E(X)$ y $Var(X)$.

32. Sea X una variable aleatoria absolutamente continua con valores en los reales positivos, y defina $S_X(x) = 1 - F_X(x) = P(X > x)$ para un real positivo x cualquiera.

(a) Pruebe que si $E(X^2)$ existe, entonces

$$E(X) = \int_0^\infty S_X(x)dx \qquad \text{y} \qquad E(X^2) = 2\int_0^\infty xS_X(x)dx.$$

(b) Aplique lo anterior al caso de la distribución exponencial, y al caso de la distribución de Weibull.

33. Una urna contiene a bolas blancas y b bolas negras. Si sacamos una bola a la vez hasta obtener la primera bola blanca, encontrar el número esperado de bolas negras sacadas de la urna.

$Resp: \frac{b}{a+1}$.

34. Una caja contiene inicialmente 3 bolitas rojas, 4 azules y 6 verdes, las que se retiran una a una y sin reemplazo, hasta que todas las bolitas rojas han sido retiradas. Sea X el número de bolitas que se han retirado hasta ese momento.

(a) Calcule $P(X \leq 9)$

(b) Calcule $P(X = 9)$.

(b) Calcule $E(X)$.

35. Sea X una variable aleatoria que sigue una de las siguientes distribuciones.

(a) Bin(n,p).

(b) Poisson(λ).

(c) Geométrica con parámetro p.

(d) Uniforme en los enteros entre m y n, con $m < n$.

Para cada distribución calcule

a) $E(X)$.
b) $E(X(X-1))$.
c) $E(X^2)$.
d) $Var(X)$
e) $E(z^X)$, donde z es un número real.

36. Sea X una variable aleatoria con valores en $\{0, 1, \ldots, n\}$, función de probabilidad f y función de distribución F. Demuestre que

$$EX = \sum_{x=0}^{n} (1 - F(x)).$$

Muestre que esta relación es también válida para $n = \infty$. Aplíquela para calcular la media de la distribución geométrica.

37. Un equipo tiene 5 componentes, de las cuales 2 son defectuosas. Se inspeccionan las componentes en un orden aleatorio.

(a) Si X es el número de componentes que deben examinarse antes de encontrar una defectuosa, calcule $E(X)$.

(b) Si Y es el número de componentes que deben examinarse para encontrar las dos defectuosas, calcule $E(Y)$.

38. Si X es una variable aleatoria con esperanza finita μ y varianza σ^2, y si $g(\cdot)$ es una función dos veces diferenciable, demuestre que:

$$E[g(X)] \approx g(\mu) + \frac{g''(\mu)}{2}\sigma^2.$$

Hint: usar la expansión de Taylor en torno a μ para $g(\cdot)$. Use sólo los primeros tres términos.

39. Se realizan ensayos independientes, donde en el i-ésimo ensayo se obtiene un éxito con probabilidad p_i. Encuentre el número esperado y la varianza del número de éxitos que ocurren en los primeros n ensayos.

40. Un hombre dispara a un blanco. Diez de estos tiros caen a una pulgada del blanco, cinco entre una y tres pulgadas del blanco, y tres entre tres y cinco pulgadas del blanco. Encontrar el número esperado de tiros acertados si:

(a) Los tiros del hombre se distribuyen uniformemente en el círculo de radio ocho pulgadas con el blanco como centro.

(b) Las distancias verticales y horizontales de los tiros del hombre al blanco son (medidas en pulgadas) variables aleatorias independientes e idénticamente distribuidas $N(0, 4)$.

41. La duración T de cierto tipo de llamada telefónica satisface la relación:

$$P(T \geq t) = ae^{-\lambda t} + (1-a)e^{-\mu t}, \qquad t \geq 0,$$

donde $0 \leq a \leq 1$, $\lambda \geq 0$ y $\mu \geq 0$ son constantes determinadas estadísticamente. Encontrar la media y la varianza de T.

42. Una variable aleatoria X puede tomar cada uno de los siete valores $-3, -2, -1, 0, 1, 2, 3$ con la misma probabilidad. Determinar $f_Y(y)$, en donde $Y = X^2 - X$.

43. Suponga que X es una variable aleatoria cuya densidad es f y que $Y = aX + b$ $(a \neq 0)$. Demuestre que la densidad de Y es la siguiente:

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right), \qquad -\infty \leq y \leq \infty.$$

44. Suponga que X tiene función densidad:

$$g(x) = \begin{cases} ce^{-cx} & \text{si } x \geq 0 \\ 0 & \text{si no.} \end{cases}$$

(a) Demostrar que $\frac{X}{1+X}$ tiene función densidad:

$$g(x) = \begin{cases} \frac{e^{-\frac{x}{1-x}}}{(1-x)^2} & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

(b) Demostrar que $X + c$ tiene función densidad:

$$g(x) = \begin{cases} e^{-(x-c)} & \text{si } c \leq x \\ 0 & \text{si } x \leq c. \end{cases}$$

45. Sea X una variable aleatoria continua con función densidad f y función distribución F. Pruebe que la distribución de $Y = F(x)$ es $U(0,1)$.

46. Supongamos que una calculadora posee cuatro circuitos. Si ésta se envía a reparación, las probabilidades que necesite 1,2,3 o 4 circuitos nuevos son $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, y $\frac{1}{8}$. La empresa que realiza las reparaciones mantiene un stock de 18875 circuitos anuales. Si en un a no recibe 10000 calculadoras para reparar, ¿cuál es la probabilidad que no se cubra la demanda?.

$Resp$: 0,117

47. Observando que, en promedio, el 12 % de los pasajes reservados no se ocupan, una compa nía aérea decide aceptar reservas por un 10 % más de su capacidad en aviones de 450 pasajeros. Calcular la proporción de vuelos en que algún pasajero con reserva no tiene cabida.

$Resp$: 0,02

48. Suponga que $X \sim U(0,1)$. Determine los valores de $t \in \mathbb{R}$ tales que $E(X^t)$ existe.

49. (a) Un dado se lanza hasta obtener un dos. Si X es el número de lanzamientos requeridos, demostrar que la función generadora de momentos de X es $\frac{t}{6-5t}$.

(b) Un dado se lanza hasta obtener un dos o un tres. Demostrar que la función generadora de momentos del número de lanzamientos requeridos es $\frac{t}{3-2t}$.

50. Suponiendo que X tenga la siguiente función densidad :

$$f(x) = \lambda e^{-\lambda(x-a)} \qquad X \leq a$$

(a) Encontrar $M_X(t)$.

(b) Calcular $E(X)$ y $Var(X)$.

51. Si X es una variable aleatoria continua no negativa, pruebe que

$$E(X^n) = \int_0^\infty nx^{n-1}(1 - F(x))dx,$$

donde F es la función de distribución de X.

52. Demostrar que si $X_i,\ i = 1, \ldots, k$ representa el número de éxitos en k repeticiones de un experimento para el que $P(\text{éxito}) = p\ \forall i$, entonces $X_1 + \cdots + X_k$ tiene una distribución Binomial.

Capítulo 4

Vectores Aleatorios

4.1. Motivación

En el capítulo anterior hemos estudiado el importante concepto de variable aleatoria, con énfasis en el caso en que ésta es univariada. Es usual, sin embargo, el caso en que el objeto aleatorio natural para modelar una situación dada es un *vector aleatorio* de n componentes, es decir, se observa $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, en que cada X_i es una variable aleatoria unidimensional, ya sea discreta, absolutamente continua o mixta.

Este es el caso del Ejemplo 3.2.4 del Capítulo 3, en que el resultado de escoger un punto al azar en el círculo unitario se describe por un vector aleatorio bidimensional $\boldsymbol{X} = (X_1, X_2)$, y en donde $\boldsymbol{\mathcal{X}} = \{(x_1, x_2) \mid x_1^2 + x_2^2 \leq 1\}$. Note que en este mismo ejemplo el resultado puede también ser descrito en términos de coordenadas polares $\boldsymbol{Y} = (R, \Theta)$, donde $\boldsymbol{\mathcal{Y}} = \{(r, \theta) : 0 \leq r \leq 1, -\pi \leq \theta \leq \pi\}$. Observe, sin embargo, que si este experimento se cambia por escoger un punto en la *circunsferencia* unitaria $\{(x_1, x_2) \mid x_1^2 + x_2^2 = 1\}$, el vector correspondiente es en realidad un objeto unidimensional, lo que se puede modelar empleando las técnicas del Capítulo 3. Concretamente, en coordenadas polares, escogemos $R = 1$ y $\Theta \sim U(0, 2\pi)$. Se propone como ejercicio al lector obtener la correspondiente distribución de las coordenadas cartesianas X_1 y X_2.

Otra situación es cuando una cierta medición se lleva a cabo en varios individuos. Por ejemplo, suponga que interesa medir la estatura de cada uno de los 6 integrantes de una cierta familia. El resultado de este experimento se puede representar mediante un vector aleatorio de dimensión 6, en que cada componente representa la estatura de uno de los miembros de esta familia. Esta clase de ejemplo es muy frecuente en problemas estadísticos de la vida real. Sin entrar en mayores detalles por ahora, es conveniente distinguir el vector aleatorio obtenido de esta forma con aquel que uno obtendría si se midiera la estatura de uno de los miembros de esta familia 6 veces, y aún con el caso en que a este mismo individuo se le miden 6 características diferentes (por ejemplo, estatura, peso, etc.). Como veremos más adelante estas tres situaciones requieren de modelos probabilísticos radicalmente distintos, aún cuando se trata de vectores aleatorios de la misma dimensión, y obtenidos en situaciones "similares".

4.2. Definiciones y Conceptos Básicos

4.2.1. Definiciones

Veremos que muchas de las ideas del caso unidimensional tienen una extensión natural al caso multidimensional. Por esta razón, no nos detendremos mayormente en revisar algunos aspectos cubiertos en el Capítulo 3. Más bien, enfatizaremos los cambios específicos que involucra el salto desde dimensión 1 a n.

Definición 4.2.1 Un vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ se dice *vector aleatorio* si cada uno de los X_i, $i = 1, \ldots, n$ es una variable aleatoria, siendo todas ellas definidas sobre un espacio muestral común Ω. La notación $\boldsymbol{X} \in \mathbb{R}^n$ indicará que $\boldsymbol{X}$ tiene n coordenadas.

En otras palabras, si cada X_i es una variable aleatoria, $i = 1, \ldots, n$, entonces se tiene que $\boldsymbol{X} = (X_1, \ldots, X_n)$ es un vector aleatorio de dimensión n. La restricción que las n variables aleatorias estén definidas sobre un mismo Ω obedece a razones técnicas, y en la práctica, uno puede asumir dicha condición sin pérdida de generalidad. Al igual que en el caso unidimensional, Ω suele ser un conjunto difícil de especificar (¡en algunos caso es incluso difícil de imaginar!), por lo que los modelos probabilísticos – expresados en términos de distribuciones – se postulan usualmente sobre $\boldsymbol{\mathcal{X}}$, el conjunto de posibles valores de $\boldsymbol{X}$.

Definición 4.2.2 La *función de distribución conjunta* de un vector aleatorio $\boldsymbol{X}$ se define para un vector dado $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ mediante:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1,\cdots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n). \tag{4.2.1}$$

Nótese que en este caso se habla de función de distribución *conjunta* de $\boldsymbol{X}$, denotando el hecho que $\boldsymbol{X}$ posee más de una coordenada, estableciéndose así una distinción explícita con el caso unidimensional.

4.2.2. Propiedades de la función de distribución conjunta

La función de distribución conjunta tiene las siguientes propiedades:

1. $F_{\boldsymbol{X}}$ es no decreciente en cada coordenada, esto es, si $x_{i1} < x_{i2}$, entonces

$$F_{\boldsymbol{X}}(x_1, \ldots, x_{i1}, \ldots, x_n) \leq F_{\boldsymbol{X}}(x_1, \ldots, x_{i2}, \ldots, x_n).$$

2. $F_{\boldsymbol{X}}$ es continua por la derecha en cada coordenada, esto es,

$$\lim_{x_i \to x_{i0}^+} F_{\boldsymbol{X}}(x_1, \ldots, x_i, \ldots, x_n) = F_{\boldsymbol{X}}(x_1, \ldots, x_{i0}, \ldots, x_n).$$

   Análogamente, los límites por la izquierda en cada coordenada existen (aunque no necesariamente coinciden con los valores de $F_{\boldsymbol{X}}$ en los puntos en cuestión).

3. Para cualquier i se tiene
$$\lim_{x_i \to -\infty} F_{\boldsymbol{X}}(x_1, \ldots, x_n) = 0,$$
y
$$\lim_{x_1, \ldots, x_n \to \infty} F_{\boldsymbol{X}}(x_1, \ldots, x_n) = 1,$$
donde este último límite significa que *todas* las coordenadas tienden simultáneamente a ∞.

4. Para $g : \mathbb{R}^n \to \mathbb{R}$ sea
$$\Delta^k_{(a_k, b_k]} g(x_1, \ldots, x_n) = g(x_1, \ldots, b_k, \ldots, x_n) - g(x_1, \ldots, a_k, \ldots, x_n).$$
Entonces,
$$\Delta^1_{(a_1, b_1]} \cdots \Delta^n_{(a_n, b_n]} F_{\boldsymbol{X}}(x_1, \ldots, x_n) \geq 0,$$
cualesquiera que sean $a_i < b_i$, $i = 1, \ldots, n$.

Se puede probar que estas cuatro propiedades caracterizan completamente la función de distribución conjunta, en el sentido que una función F satisfaciéndolas coincide con $F_{\boldsymbol{X}}$ para algún vector aleatorio $\boldsymbol{X}$. Resulta entonces natural asignar el nombre *función de distribución n-dimensional o conjunta* a cualquier función F satisfaciendo 1-4 arriba.

La propiedad 4 es quizás la más novedosa entre ellas. Para visualizar lo que sucede, consideremos el caso $n = 2$, y el siguiente ejemplo.

Ejemplo 4.2.1 Sea $F : \mathbb{R}^2 \to \mathbb{R}$ definida por
$$F(x, y) = \begin{cases} 1 & \text{si } x \geq 0,\, y \geq 0,\, x + y \geq 1 \\ 0 & \text{si no.} \end{cases}$$
Si F fuera la función de distribución de algún vector aleatorio (X, Y), entonces, anotando $F_{X,Y}(x, y) = F(x, y)$ se tiene
$$\begin{aligned} P(0 < X \leq 1, 0 < Y \leq 1) &= F_{X,Y}(1,1) - F_{X,Y}(1,0) \\ &\quad -F_{X,Y}(0,1) + F_{X,Y}(0,0) \\ &= F(1,1) - F(1,0) - F(0,1) \\ &\quad +F(0,0) \\ &= 1 - 1 - 1 + 0 = -1, \end{aligned}$$
que es claramente una contradicción.

Es claro, entonces, que la cuarta propiedad (que simplemente establece que probabilidades calculadas a partir de F deben ser no negativas) resulta relevante, y no puede ser omitida. Dicha propiedad se puede visualizar como la extensión multivariada de aquella establecida en la Sección 3.4.1, y que se traduce en el hecho que las funciones de distribución univariadas son no decrecientes.

Como en el caso unidimensional, es posible clasificar vectores aleatorios como *discretos* y *continuos*.

Definición 4.2.3 El vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)$ se dice

- *discreto* si $\boldsymbol{X}$ toma valores sobre un conjunto finito o infinito numerable. En este caso, si $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, la función
$$p_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 = x_1, \ldots, X_n = x_n)$$
recibe el nombre de *función de probabilidad conjunta discreta*.

- *absolutamente continuo* si existe una función $f_{\boldsymbol{X}} : \mathbb{R}^n \to \mathbb{R}^+ \cup \{0\}$ tal que para cualquier $\boldsymbol{x} \in \mathbb{R}^n$ se cumple
$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(t_1, \ldots, t_n) dt_1 \cdots dt_n.$$
En este caso, $f_{\boldsymbol{X}}$ se llama *función densidad* del vector aleatorio $\boldsymbol{X}$, o *función densidad conjunta* de las variables aleatorias $X_1, \ldots, X_n$.

Como en el caso unidimensional, $F_{\boldsymbol{X}}$ suele tener poca importancia práctica cuando $\boldsymbol{X}$ es discreto, y uno trabaja usualmente con $p_{\boldsymbol{X}}$. En el caso continuo, la probabilidad que el vector aleatorio $\boldsymbol{X}$ tome valores en el rectángulo n-dimensional

$$\boldsymbol{R}_n = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

se obtiene mediante integración:

$$P(\boldsymbol{X} \in \boldsymbol{R}_n) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f_{\boldsymbol{X}}(t_1, \ldots, t_n) dt_1 \cdots dt_n. \tag{4.2.2}$$

Si el vector aleatorio $\boldsymbol{X}$ está definido en Ω, sobre el cual se ha definido una medida de probabilidad P, el vector $\boldsymbol{X}$ induce una nueva medida de probabilidad, $P_{\boldsymbol{X}}$, esta vez sobre $\boldsymbol{\mathcal{X}}$, y dada mediante la fórmula:

$$P_{\boldsymbol{X}}(B) = P(\boldsymbol{X} \in B) = P(X^{-1}(B)) \text{ para } B \in \boldsymbol{\mathcal{X}}. \tag{4.2.3}$$

Al igual que en el caso univariado, $P_{\boldsymbol{X}}$ recibe el nombre de *medida de probabilidad inducida* por $\boldsymbol{X}$, o *distribución* de $\boldsymbol{X}$.

Otras propiedades de las funciones de probabilidad se verán a continuación:

1. Si la función de probabilidad discreta de $\boldsymbol{X} = (X_1, \ldots, X_n)$ es $p_{\boldsymbol{X}}$, entonces la función de probabilidad conjunta de un subconjunto cualquiera de $(X_1, \ldots, X_n)$ se obtiene simplemente de sumar $p_{\boldsymbol{X}}$ sobre las coordenadas correspondientes a las variables no incluidas en dicho subconjunto, y atendiendo a las restricciones que los puntos en $\boldsymbol{\mathcal{X}}$ poseen. En particular:
$$p_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \sum_{S(k+1,\ldots,n)} p_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n), \tag{4.2.4}$$
en donde
$$S(k+1, \ldots, n) = \{(x_{k+1}, \ldots, x_n) : (x_1, \ldots, x_n) \in \boldsymbol{\mathcal{X}}\}.$$

En efecto, el conjunto de posibles valores para $(X_1, \ldots, X_k)$ se obtiene de "proyectar" $\boldsymbol{\mathcal{X}}$ sobre las primeras k coordenadas. Pero en este proceso, la probabilidad de un punto cualquiera se obtiene como la suma de todos los puntos de $\boldsymbol{\mathcal{X}}$ cuyas primeras k coordenadas coinciden (esto es, aplicamos probabilidades totales), que es exactamente lo que se establece en (4.2.4). Una notación alternativa para (4.2.4) es

$$p_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = p_{\boldsymbol{X}}(x_1, \ldots, x_k, +, \ldots, +). \tag{4.2.5}$$

2. Si $\boldsymbol{X} = (X_1, \ldots, X_n)$ tiene densidad conjunta $f_{\boldsymbol{X}}$, entonces, y en analogía con el caso discreto, la densidad conjunta de un subconjunto de ellas se obtiene de integrar las coordenadas que no pertenecen a dicho subconjunto:

$$f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_{k+1} \cdots dx_n. \tag{4.2.6}$$

3. Si $F_{\boldsymbol{X}}$ es la función de distribución conjunta de $\boldsymbol{X} = (X_1, \ldots, X_n)$, entonces la función de distribución conjunta de algún subconjunto de ellas se obtiene de tomar el límite de cada coordenada no involucrada en el subconjunto, cuando esta coordenada tiende a ∞. En particular:

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \lim_{x_{k+1},\ldots,x_n \to \infty} F_{X_1,\ldots,X_n}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n). \tag{4.2.7}$$

Intuitivamente, al tomar límite a infinito en alguna coordenada, digamos, para fijar ideas, la última, se reemplaza el evento $\{X_n \leq x_n\}$ por $\{X_n \leq \infty\}$, el cual tiene probabilidad 1, y por lo tanto, este evento no altera la probabilidad de los otros eventos que definen $F_{\boldsymbol{X}}$:

$$\begin{aligned} \lim_{x_n \to \infty} F_{\boldsymbol{X}}(x_1, ..., x_n) &= \lim_{x_n \to \infty} P(X_1 \leq x_1, \ldots, X_{n-1} \leq x_{n-1}, X_n \leq x_n) \\ &= P(X_1 \leq x_1, \ldots, X_{n-1} \leq x_{n-1}, X_n \leq \infty) \\ &= P(X_1 \leq x_1, \ldots, X_{n-1} \leq x_{n-1}) \\ &= F_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1}). \end{aligned}$$

El argumento en el caso del límite de dos o más coordenadas en forma simultánea es esencialmente idéntico.

4.2.3. Ejemplos

Veamos a continuación algunos ejemplos.

Ejemplo 4.2.2 Sea $X_1, X_2, \ldots$ un proceso de Bernoulli con probabilidad de éxito p. En este caso se tiene que $\boldsymbol{X} = (X_1, \ldots, X_n)$ es un vector aleatorio discreto n-dimensional, cualquiera que sea n. Por las propiedades ya estudiadas para este caso, $\boldsymbol{\mathcal{X}} = \{0, 1\}^n$, y para cualquier $\boldsymbol{x} = (x_1, \ldots, x_n) \in \boldsymbol{\mathcal{X}}$ se cumple:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}.$$

Veamos ahora cómo deducir esta fórmula. Para fijar ideas, supongamos que $\boldsymbol{x}$ consta de k unos seguidos de $n-k$ ceros. Por la independencia de $X_1, \ldots, X_n$, es claro que la probabilidad de tal configuración es $p^k(1-p)^{n-k}$. Más aun, cualquier configuración con k unos y $n-k$ ceros tiene exactamente la misma probabilidad. Pero el número de unos coincide con $\sum_{i=1}^{n} X_i$, de donde se obtiene el resultado.

Ejemplo 4.2.3 Sean X e Y variables aleatorias discretas con función de probabilidad conjunta dada por la siguiente tabla:

	y		
x	0	1	2
0	0.15	0.15	0.25
1	0.10	0.15	0.20

Para obtener la función de probabilidad discreta de X e Y, y usando (4.2.4), sólo debemos sumar por filas o columnas, respectivamente. Así, se obtiene que $p_X(0) = 0{,}55$, $p_X(1) = 0{,}45$, y $p_Y(0) = 0{,}25$, $p_Y(1) = 0{,}3$, $p_Y(2) = 0{,}45$.

Ejemplo 4.2.4 Considere (X, Y) con densidad conjunta

$$f_{X,Y}(x,y) = \begin{cases} c(|x|+|y|) & \text{si } |x|+|y| \leq 1 \\ 0 & \text{si no.} \end{cases}$$

En este ejemplo, el primer paso consiste en calcular el valor de $c > 0$ para que efectivamente se tenga una densidad. Se debe cumplir

$$1 = \iint\limits_{|x|+|y|\leq 1} c(|x|+|y|)dxdy.$$

Por la simetría de ambos, el dominio de integración y la función en cuestión, la integral sobre cada cuadrante es la misma, y

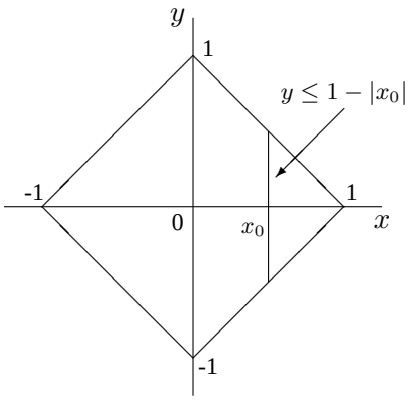
$$\begin{aligned} 1 &= 4c \iint\limits_{\{x+y\leq 1,\ x\geq 0,\ y\geq 0\}} (x+y)dxdy = 4c\int_0^1 \int_0^{1-y} (x+y)dxdy \\ &= 4c\int_0^1 [(1-y)^2/2 + y(1-y)]dy = 4c\int_0^1 [1/2 - y^2/2]dy \\ &= 4c(1/2 - 1/6) = 4c/3, \end{aligned}$$

de donde se concluye que $c = 3/4$.

Calculemos ahora las correspondientes densidades marginales. A partir de (4.2.6) se tiene que

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy.$$

Es necesario ser muy cuidadoso con los límites de integración. En primer lugar, $\mathcal{X}$ se obtiene de proyectar el dominio sobre el eje x, obteniéndose $\mathcal{X} = [0,1]$, de modo que $f_X(x) = 0$ si $x \notin [0,1]$. Para $x_0 \in [0,1]$ fijo, el rango de posibles valores de y se obtiene de la desigualdad $|x_0| + |y| \leq 1$, de donde se sigue que $-(1-|x_0|) \leq y \leq 1 - |x_0|$, como se muestra en el siguiente diagrama.



Se tiene entonces que

$$\begin{aligned} f_X(x) &= \frac{3}{4}\int_{-(1-|x|)}^{1-|x|} (|x|+|y|)dy \\ &= \frac{3}{2}|x|(1-|x|) + \frac{3}{4}\int_{-(1-|x|)}^{1-|x|} |y|dy \\ &= \frac{3}{2}|x|(1-|x|) + \frac{3}{2}\int_{0}^{1-|x|} ydy \\ &= \frac{3}{2}|x|(1-|x|) + \frac{3}{4}(1-|x|)^2 = \frac{3}{4}(1-x^2). \end{aligned}$$

Análogamente, y por la simetría del problema,

$$f_Y(y) = \begin{cases} \frac{3}{4}(1-y^2) & \text{si } y \in [-1,1] \\ 0 & \text{si no.} \end{cases}$$

4.2.4. El caso mixto

Hay aún un caso que discutir, y que corresponde a cuando parte de las variables en el vector aleatorio $\boldsymbol{X}$ son discretas, y el resto absolutamente continuas, caso en el que hablamos de vector aleatorio mixto. Para simplificar la exposición, supongamos un vector bidimensional (X,Y), donde X es discreta, e Y es absolutamente continua, y denotemos por $\boldsymbol{\mathcal{D}}$ al conjunto de posibles valores para este vector. Notemos que la distribución de (X,Y) asigna probabilidades positivas a algunos subconjuntos de $\mathbb{R}^2$ de la forma $\{x\} \times [a,b]$. Sin pérdida de generalidad, podemos descartar aquellos subconjuntos tales que $P(X = x) = 0$ *marginalmente*. Surge entonces el problema de cómo definir una "función densidad" que permita realizar los cálculos como lo hemos estado haciendo hasta ahora. Es claro que al operar con esta función densidad, se requerirá una combinación de sumas e

integrales, correspondientes a la parte discreta y continua respectivamente. Para ello, introduzcamos primero la siguiente notación. Sea $A \subset \boldsymbol{\mathcal{D}}$ un evento de interés, y defina:

$$\begin{aligned} A_x &= \{y \in \mathbb{R} : (x,y) \in A\} \\ A_y &= \{x \in \mathbb{R} : (x,y) \in A\} \\ A(X) &= \bigcup_{y \in \mathcal{Y}} A_y \\ A(Y) &= \bigcup_{x \in \mathcal{X}} A_x. \end{aligned}$$

Los conjuntos A_x y A_y reciben el nombre de *secciones* de A. Así, A_x contiene todos los puntos $y \in \mathcal{Y}$ para los que el segmento paralelo al eje y y que pasa por x está contenido en A. Por su parte, $A(X)$ contiene todos los posibles valores x tal que $(x,y) \in A$ para algún $y \in A(Y)$. Luego, $A(X)$ puede verse como la proyección de A sobre el eje x correspondiente a la primera coordenada. Una interpretación análoga vale para A_y y $A(Y)$. Observe que, en general, $A \subset A(X) \times A(Y)$, pudiendo la inclusión ser estricta. Por otra parte, el soporte de X es simplemente $\boldsymbol{\mathcal{D}}(X)$, y el de Y es $\boldsymbol{\mathcal{D}}(Y)$.

Con esta notación, es posible probar que para un vector aleatorio mixto, existe una función densidad mixta $p_{X,Y}(x,y)$ tal que

$$P((X,Y) \in A) = \sum_{x \in A(X)} \left\{ \int_{y \in A_x} p_{X,Y}(x,y)dy \right\} = \int_{y \in A(Y)} \left\{ \sum_{x \in A_y} p_{X,Y}(x,y) \right\} dy. \qquad (4.2.8)$$

Además, la función de probabilidad discreta marginal de X se obtiene mediante

$$p_X(x) = \int_{y \in \boldsymbol{\mathcal{D}}_x} p_{X,Y}(x,y)dy, \qquad x \in \boldsymbol{\mathcal{D}}(X), \qquad (4.2.9)$$

mientras que la densidad marginal de Y se obtiene mediante

$$f_Y(y) = \sum_{x \in \boldsymbol{\mathcal{D}}_y} p_{X,Y}(x,y), \qquad x \in \boldsymbol{\mathcal{D}}(Y). \qquad (4.2.10)$$

Finalmente, el procedimiento se extiende en forma análoga al caso $n > 2$, en que algunas de las coordenadas son variables discretas, y las otras poseen una función densidad conjunta.

Ejemplo 4.2.5 Considere un vector aleatorio de tipo mixto (X,Y) para el que $\mathcal{X} = \{0,1,\ldots,n\}$, e $\mathcal{Y} = (0,1)$, y con

$$p_{X,Y}(x,y) = \begin{cases} \frac{\binom{n}{x}}{B(a,b)} y^{x+a-1}(1-y)^{n-x+b-1} & \text{si } (x,y) \in \mathcal{X} \times \mathcal{Y} \\ 0 & \text{si no,} \end{cases}$$

donde a y b son reales positivos, $n \geq 1$ es un entero cualquiera, y $B(a,b)$ es la función Beta, definida por

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Para calcular p_X se procede como a continuación, usando (4.2.9):

$$\begin{aligned} p_X(x) &= \int_0^1 p_{X,Y}(x,y)dy \\ &= \binom{n}{x}\text{Beta}(a,b)^{-1}\int_0^1 y^{x+a-1}(1-y)^{n-x+b-1}dy \\ &= \frac{\binom{n}{x}\text{Beta}(x+a,n-x+b)}{\text{Beta}(a,b)}, \end{aligned}$$

para $x = 0, 1, \ldots, n$. Por otra parte, de (4.2.10)

$$\begin{aligned} f_Y(y) &= \sum_{x=0}^{n} p_{X,Y}(x,y) \\ &= \text{Beta}(a,b)^{-1}y^{a-1}(1-y)^{b-1}\sum_{x=0}^{n}\binom{n}{x}y^x(1-y)^{n-x} \\ &= \frac{y^{a-1}(1-y)^{b-1}}{\text{Beta}(a,b)}, \end{aligned}$$

y hemos visto así que $Y \sim \text{Beta}(a,b)$. Se recomienda al lector establecer el paralelo entre la distribución aquí considerada y un experimento consistente en escoger un número $0 < Y < 1$ de acuerdo a la distribución $\text{Beta}(a,b)$, para luego lanzar una moneda con probabilidad Y de dar cara n veces en forma independiente, anotando el número de caras X que se obtienen.

Ejemplo 4.2.6 Considere un vector aleatorio mixto (X,Y) para el que

$$p_{X,Y}(x,y) = \begin{cases} \frac{y^x}{x!}\exp(-2y) & \text{si } (x,y) \in \{0,1,2,\ldots\}\times(0,\infty) \\ 0 & \text{si no,} \end{cases}$$

Usando (4.2.9) se obtiene, después de algunos cálculos directos,

$$p_X(x) = \int_0^\infty \frac{\exp(-2y)y^x}{x!}dy = \frac{1}{2^{x+1}}, \qquad x = 0,1,2,\ldots,$$

y de (4.2.10)

$$f_Y(y) = \sum_{x=0}^{\infty}\frac{\exp(-2y)y^x}{x!} = \exp(-y), \qquad y > 0,$$

y entonces Y tiene distribución exponencial con parámetro 1.

4.3. Independencia de Variables Aleatorias

Retomamos aquí el concepto de independencia de variables aleatorias introducido anteriormente, dándole un tratamiento más general, y estudiando sus consecuencias en términos de las componentes de un vector aleatorio.

Para comenzar, recordemos la definición, que se aplica a variables aleatorias de cualquier tipo.

Definición 4.3.1 Las variables aleatorias $X_1, \ldots, X_n$, definidas en el mismo espacio muestral, se dicen *independientes* si para cualquier colección de n eventos $A_1, \ldots, A_n$ se tiene

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i) \tag{4.3.1}$$

Como en el caso discreto, independencia de $X_1, \ldots, X_n$ significa que eventos relacionados a subconjuntos disjuntos de estas variables son independientes, es decir, la ocurrencia de uno de ellos no da información respecto de la probabilidad de ocurrencia de los otros.

Es posible obtener diversas caracterizaciones de independencia. La más general de ellas dice relación con la factorización de la función de distribución conjunta.

Proposición 4.3.1

(a) Si $X_1, \ldots, X_n$ son independientes, entonces

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i),$$

para cualquier $x_1, \ldots, x_n$.

(b) A la inversa, si existen funciones $F_1, \ldots, F_n$ tales que

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i),$$

y si $\lim_{x_i \to \infty} F_i(x_i) = 1$, para $i = 1, \ldots, n$, entonces $X_1, \ldots, X_n$ son independientes, y $F_{X_i}(x_i) = F_i(x_i)$ para $i = 1, \ldots, n$.

En otras palabras, si las variables en cuestión son independientes, entonces la función de distribución conjunta de ellas factoriza como el producto de las funciones de distribución univariadas involucradas. La parte (b) establece un resultado recíproco, pero esta vez, es necesario verificar que para $i = 1, \ldots, n$ se tiene $\lim_{x_i \to \infty} F_i(x_i) = 1$. Note que *no es necesario* verificar que cada F_i es una función de distribución. Por otro lado, si cada X_i es absolutamente continua (el caso discreto ya fue anteriormente tratado en la Sección 2.8.1), podemos dar una versión de este resultado basado sólo en densidades.

Proposición 4.3.2

(a) Si $X_1, \ldots, X_n$ son independientes, entonces para cualquier $x_1, \ldots, x_n$ se cumple:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

(b) Recíprocamente, si existen funciones densidad $f_1, \ldots, f_n$ tales que

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_i(x_i),$$

entonces $X_1, \ldots, X_n$ son independientes, y X_i tiene densidad f_i, $i = 1, \ldots, n$.

Veamos a continuación algunos ejemplos.

Ejemplo 4.3.1 Sea (X, Y) un vector aleatorio con distribución uniforme en el círculo unitario, esto es,

$$f_{X,Y}(x,y) = \begin{cases} \pi^{-1} & \text{si } x^2 + y^2 \leq 1 \\ 0 & \text{si no} \end{cases}$$

Tenemos que $\mathcal{X} = [-1, 1]$, y para $-1 \leq x \leq 1$ se cumple:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy \\ &= \frac{2\sqrt{1-x^2}}{\pi}, \end{aligned}$$

y $f_X(x) = 0$ si no. En completa analogía, $f_Y(y) = f_X(y)$, pero es claro que $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$, por lo que X e Y no son independientes.

Se propone como ejercicio verificar que si (X, Y) está distribuido uniformemente en el cuadrado unitario $[0, 1] \times [0, 1]$, entonces X e Y son independientes, cada una con distribución U(0,1).

Ejemplo 4.3.2 Considere el vector aleatorio (X, Y) con densidad conjunta

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

donde $(x, y) \in \mathbb{R}^2$, y en donde $-1 < \rho < 1$. Se tiene que:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(y-\rho x)^2 - \frac{x^2}{2}\right) dy \\ &= \frac{\exp\left(-\frac{x^2}{2}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(y-\rho x)^2\right) dy \\ &= \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}, \end{aligned}$$

y $X \sim N(0, 1)$. Análogamente, $Y \sim N(0, 1)$, pero es claro que X e Y no son independientes, *a menos que* $\rho = 0$, caso en que la densidad conjunta sí factoriza. Este ejemplo será nuevamente discutido más adelante.

4.4. Transformaciones de Vectores Aleatorios

4.4.1. Enfoque intuitivo

En muchos casos la información obtenida viene en la forma de un vector aleatorio n-dimensional con distribución conjunta conocida (ya sea mediante consideraciones propias al experimento, o como parte de un cierto modelo probabilístico), pero lo que realmente interesa es determinar probabilidades que digan relación con una variable aleatoria definida como una función del vector aleatorio en cuestión, digamos, $Y = g(X_1, \ldots, X_n)$. Ejemplos típicos de esta situación son sumas, promedios, productos, cambios de unidades de medida, etcétera. Concretamente, ya hemos visto el caso en que la variable de interés sea el número de éxitos obtenidos hasta el n-ésimo ensayo en un proceso de Bernoulli, que simplemente corresponde a sumar $X_1, \ldots, X_n$.

Note que

$$F_Y(y) = P(g(X_1, \ldots, X_n) \leq y),$$

de modo que, en teoría el problema ya está resuelto. En la práctica, sin embargo, son pocos los casos en que este cálculo se puede hacer directamente. Veamos un par de ejemplos simples.

Ejemplo 4.4.1 Sea $X \sim N(0,1)$, y sea $Y = X^2$. Se tiene $\mathcal{Y} = \mathbb{R}^+$, de modo que para $y > 0$:

$$\begin{aligned} F_Y(y) &= P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx = F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Luego, la densidad de Y se obtiene como $F'_Y(y)$:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \\ &= \frac{\exp(-\frac{y}{2})}{2\sqrt{y\pi}} = \frac{y^{\frac{1}{2}-1} \exp(-\frac{y}{2})}{2^{\frac{1}{2}}\sqrt{\pi}}, \end{aligned}$$

que corresponde a la distribución Gamma$(\frac{1}{2}, 2)$. Esta distribución recibe también el nombre de *Chi-cuadrado* con 1 grado de libertad, como se verá más adelante.

Ejemplo 4.4.2 Sean X_1 y X_2 i.i.d. con distribución exponencial de parámetro $\lambda > 0$. Calculemos la densidad de $Y = X_1 + X_2$. Es inmediato ver que $\mathcal{Y} = \mathbb{R}^+$, y que

$$f_{X_1,X_2}(x_1, x_2) = \lambda^{-2} \exp(-(x_1 + x_2)/\lambda).$$

Entonces, para $y > 0$:

$$\begin{aligned} F_Y(y) &= P(X_1 + X_2 \leq y) \\ &= \int\limits_{\{x_1+x_2\leq y,\, x_1\geq 0, x_2\geq 0\}} \lambda^{-2} \exp(-(x_1+x_2)/\lambda) dx_1 dx_2 \\ &= \int_0^y \int_0^{y-x_2} \lambda^{-(x_1+x_2)} \exp(-(x_1+x_2)/\lambda) dx_1 dx_2 \\ &= \int_0^y \lambda^{-1} \exp(-x_2/\lambda) \left\{ \int_0^{y-x_2} \lambda^{-1} \exp(-x_1/\lambda) dx_1 \right\} dx_2 \\ &= \int_0^y \lambda^{-1} (\exp(-x_2/\lambda) - \exp(-y/\lambda)) dx_2 \\ &= 1 - \exp(-y/\lambda) - \lambda^{-1} y \exp(-y/\lambda), \end{aligned}$$

y de aquí se concluye, mediante diferenciación, que

$$f_Y(y) = \lambda^{-2} y \exp(-y/\lambda),$$

y por lo tanto $Y \sim \Gamma(2, \lambda)$.

4.4.2. El Teorema del cambio de variables: caso biyectivo

Cuando la transformación involucra funciones más complicadas, este método "directo" se torna difícil de emplear. Afortunadamente, es posible recurrir al *Teorema del cambio de variables* para obtener el siguiente e importante resultado.

Teorema 4.4.1 Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ un vector aleatorio n-dimensional con valores en $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^n$, y con densidad conjunta $f_{\boldsymbol{X}}$. Sea $\boldsymbol{Y} = g(\boldsymbol{X})$ una función para la que $g : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}} = g(\boldsymbol{\mathcal{X}}) \subset \mathbb{R}^n$ es biyectiva y tal que g^{-1} es continuamente diferenciable, y en donde $\boldsymbol{\mathcal{X}}$ e $\boldsymbol{\mathcal{Y}}$ son regiones abiertas de $\mathbb{R}^n$. Entonces $\boldsymbol{Y}$ es también absolutamente continua, con densidad conjunta dada por

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \begin{cases} f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y})) |\det(Jg^{-1}(\boldsymbol{y}))| & \text{si } \boldsymbol{y} \in \boldsymbol{\mathcal{Y}} \\ 0 & \text{si no,} \end{cases} \tag{4.4.1}$$

y en donde $Jg^{-1}(\boldsymbol{y})$ es la matriz Jacobiana de la transformación inversa $g^{-1} = (g_1^{-1}, \ldots, g_n^{-1}) : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$, dada por

$$Jg^{-1}(\boldsymbol{y}) = \begin{pmatrix} \frac{\partial g_1^{-1}(\boldsymbol{y})}{\partial y_1} & \cdots & \frac{\partial g_1^{-1}(\boldsymbol{y})}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial g_n^{-1}(\boldsymbol{y})}{\partial y_1} & \cdots & \frac{\partial g_n^{-1}(\boldsymbol{y})}{\partial y_n} \end{pmatrix}.$$

El teorema del cambio de variables es una herramienta bastante útil en el cálculo de distribuciones de transformaciones de vectores aleatorios en el caso absolutamente continuo. Note que si $n = 1$, el resultado se reduce a lo ya visto en el Teorema 3.9.1.

Veamos a continuación algunas aplicaciones.

Ejemplo 4.4.3 Sean X e Y variables aleatorias i.i.d. con distribución común U(0,1). Sean $R = \sqrt{2\log(1/(1-X))}$ y $\Theta = \pi(2Y-1)$. Vamos a probar que $Z = R\cos(\Theta)$ y $W = R\sin(\Theta)$ son independientes e idénticamente distribuidos, con distribución común N(0,1).

Primero, observe que Θ tiene distribución uniforme en $[-\pi,\pi]$. En efecto, si $g(y) = \pi(2y-1)$ entonces $g([0,1]) = [-\pi,\pi]$, y es claro que g es biyectiva y continuamente diferenciable. Por otra parte,

$$f_Y(y) = \begin{cases} 1 & \text{si } y \in [0,1] \\ 0 & \text{si no.} \end{cases}$$

Además $g^{-1}(\theta) = 2^{-1}(1+\theta\pi^{-1})$, por lo que $|\det(Jg^{-1}(\theta))| = (2\pi)^{-1}$, y de (4.4.1) se obtiene

$$f_\Theta(\theta) = \begin{cases} (2\pi)^{-1} & \text{si } \theta \in [-\pi,\pi] \\ 0 & \text{si no,} \end{cases}$$

de modo que $\Theta \sim U(-\pi,\pi)$. En segundo lugar, calculemos la densidad de R. Es fácil ver que R toma valores en $]0,\infty[$. Ahora, si definimos $g(x) = \sqrt{2\log(1/(1-X))}$, entonces g es también biyectiva y continuamente diferenciable. Además, se tiene que $g^{-1}(r) = 1-\exp(-r^2/2)$, y $|\det(Jg^{-1}(r))| = r\exp(-r^2/2)$. De aquí se sigue que

$$f_R(r) = \begin{cases} r\exp(-r^2/2) & \text{si } r > 0 \\ 0 & \text{si no.} \end{cases}$$

Veamos ahora cómo obtener el resultado. Puesto que X e Y son independientes, R y Θ también lo son, de modo que

$$f_{R,\Theta}(r,\theta) = \begin{cases} (2\pi)^{-1} r\exp(-r^2/2) & \text{si } r>0,\ -\pi<\theta<\pi \\ 0 & \text{si no.} \end{cases}$$

Defina ahora las nuevas variables

$$(z,w) = g(r,\theta) = (r\cos(\theta), r\sin(\theta))$$

sobre $\{(r,\theta) : r>0, -\pi<\theta<\pi\}$. Es claro que (Z,W) toma valores en todo $\mathbb{R}^2$, que g es biyectiva y continuamente diferenciable, y que g^{-1} está dada por $g^{-1}(z,w) = (\sqrt{z^2+w^2}, \arctan(w/z))$. La matriz Jacobiana de la transformación inversa está dada por

$$Jg^{-1}(z,w) = \begin{pmatrix} \frac{z}{\sqrt{z^2+w^2}} & \frac{w}{\sqrt{z^2+w^2}} \\ \frac{-w}{z^2+w^2} & \frac{z}{z^2+w^2} \end{pmatrix},$$

y de aquí $|\det(Jg^{-1}(z,w))| = 1/\sqrt{z^2+w^2}$. Por (4.4.1), la densidad conjunta de (Z,W) está dada por

$$\begin{aligned} f_{Z,W}(z,w) &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}(z^2+w^2)\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \times \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w^2}{2}\right), \end{aligned}$$

que factoriza como el producto de dos funciones densidad N(0,1), y esto es exactamente lo que queríamos probar. Por último, note que se puede tomar un único cambio de variables, que no requiera calcular previamente las densidades de R y Θ. Los detalles de este procedimiento se proponen como ejercicio.

Ejemplo 4.4.4 Suponga que X e Y son independientes con $X \sim \Gamma(a,\lambda)$, e $Y \sim \Gamma(b,\lambda)$. Calculemos la densidad de $Z = X/(X+Y)$.

Aún cuando este problema se puede hacer sin usar el Teorema del cambio de variables (es un buen ejercicio), preferimos utilizar aquí dicho resultado. La idea es construir un cambio de variables en $\mathbb{R}^2$ que tenga a $X/(X+Y)$ en alguna coordenada, y alguna transformación simple en la otra. Una vez obtenida la densidad conjunta, se procede a calcular la densidad marginal de la variable de interés. Este método suele aplicarse muy a menudo en problemas de esta índole.

Consideremos $(Z,W) = g(X,Y) = (X/(X+Y), Y)$, donde es claro que este nuevo vector toma valores en $]0,1[\times]0,\infty[$. Se tiene que $g^{-1}(z,w) = (zw/(1-z), w)$. Note que la matriz Jacobiana es triangular, pues el elemento (2,1) de esta matriz es $\partial w/\partial z = 0$, de modo que el determinante correspondiente es el producto de los elementos en la diagonal de la matriz, y así no se necesita calcular el elemento (1,2). Luego:

$$|\det(Jg^{-1}(z,w))| = \left|\frac{w}{(1-z)^2} \times 1\right| = \frac{w}{(1-z)^2}.$$

Por otra parte, debido a la independencia,

$$f_{X,Y}(x,y) = \begin{cases} \frac{x^{a-1}y^{b-1}\exp(-\frac{x+y}{\lambda})}{\Gamma(a)\Gamma(b)\lambda^{a+b}} & \text{si } x,y>0 \\ 0 & \text{si no.} \end{cases}$$

Por (4.4.1), la densidad conjunta de (Z,W) está dada por

$$f_{Z,W}(z,w) = \begin{cases} \frac{z^{a-1}w^{(a+b)-1}}{\Gamma(a)\Gamma(b)\lambda^{a+b}}\exp\left(-\frac{w}{\lambda(1-z)}\right) & \text{si } 0<z<1, w>0 \\ 0 & \text{si no.} \end{cases}$$

Para obtener f_Z, usamos (4.2.6):

$$\begin{aligned} f_Z(z) &= \int_0^\infty f_{Z,W}(z,w)dw \\ &= \frac{z^{a-1}}{(1-z)^{a+1}\Gamma(a)\Gamma(b)}\int_0^\infty \frac{w^{(a+b)-1}}{\lambda^{a+b}}\exp\left(-\frac{w}{\lambda(1-z)}\right)dw \\ &= \frac{z^{a-1}(1-z)^{b-1}}{B(a,b)}\int_0^\infty \frac{w^{(a+b)-1}\exp\left(-\frac{w}{\lambda(1-z)}\right)}{\Gamma(a+b)(\lambda(1-z))^{a+b}}dw \\ &= \frac{z^{a-1}(1-z)^{b-1}}{B(a,b)}, \end{aligned}$$

y así hemos probado que $Z \sim \text{Beta}(a,b)$.

Ejemplo 4.4.5 Sea (X, Y) un vector aleatorio con valores en $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^2$ y densidad conjunta $f_{X,Y}$. Sea $Z = X + Y$. Podemos calcular la densidad de Z mediante aplicación del cambio de variables $(Z, W) = g(X, Y) = (X + Y, Y)$. Es claro que g cumple las hipótesis del Teorema 4.4.1, y que $(x, y) = g^{-1}(z, w) = (z - w, w)$, por lo que es fácil obtener que $|\det(Jg^{-1}(z, w))| = 1$. Se tiene, entonces, que

$$f_{Z,W}(z, w) = \begin{cases} f_{X,Y}(z - w, w) & \text{si } (z - w, w) \in \boldsymbol{\mathcal{X}} \\ 0 & \text{si no,} \end{cases}$$

por lo que

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(z - w, w)dw. \tag{4.4.2}$$

En el caso particular que X e Y son *independientes*, la ecuación (4.4.2) toma la forma especial de *convolución* de f_X y f_Y:

$$f_Z(z) = f_X \star f_Y(z) = \int_{-\infty}^{\infty} f_X(z - w) f_Y(w)dw, \tag{4.4.3}$$

es decir, si X e Y son independientes con densidades respectivas f_X y f_Y, su suma tiene densidad dada por (4.4.3).

A modo de aplicación, consideremos el caso en que $X \sim \Gamma(a, \lambda)$, e $Y \sim \Gamma(b, \lambda)$. La densidad de $Z = X + Y$ se obtiene de (4.4.3) mediante

$$\begin{aligned} f_Z(z) &= \int_0^z \left\{ \frac{(z - w)^{a-1} e^{-(z-w)/\lambda}}{\Gamma(a)\lambda^a} \times \frac{w^{b-1} e^{-w/\lambda}}{\Gamma(b)\lambda^b} \right\} dw \\ &\quad \text{(note que se debe cumplir } z - w > 0) \\ &= \frac{\exp(-z/\lambda)}{\Gamma(a)\Gamma(b)\lambda^{a+b}} \int_0^z (z - w)^{a-1} w^{b-1} dw \\ &= \frac{\exp(-z/\lambda)}{\Gamma(a)\Gamma(b)\lambda^{a+b}} \times z^{(a+b)-1} \int_0^1 (1 - x)^{a-1} x^{b-1} dx \\ &\quad \text{(cambio de variable } x = w/z) \\ &= \frac{z^{(a+b)-1} \exp(-z/\lambda) B(a, b)}{\Gamma(a)\Gamma(b)\lambda^{a+b}} = \frac{z^{(a+b)-1} \exp(-z/\lambda)}{\Gamma(a + b)\lambda^{a+b}}, \end{aligned}$$

de donde se tiene $Z = X + Y \sim \Gamma(a + b, \lambda)$.

4.4.3. El teorema del cambio de variables: caso no biyectivo

Consideremos ahora el caso en que la función g no es biyectiva o diferenciable en todo el conjunto $\boldsymbol{\mathcal{X}}$. En este caso, y al igual que en el caso unidimensional, hay una versión del Teorema del cambio de variables basado en la existencia de subconjuntos $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2, \ldots$ tales que la restricción de g a $\boldsymbol{\mathcal{X}}_i$ verifique las hipótesis del Teorema 4.4.1. Este resultado se enuncia a continuación.

Teorema 4.4.2 Sea $\boldsymbol{X}$ un vector aleatorio n-dimensional con valores en $\boldsymbol{\mathcal{X}}$. Suponga que existen subconjuntos de $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2, \ldots$ de $\boldsymbol{\mathcal{X}}$ tales que $P(\boldsymbol{X} \in \bigcup_{i=1}^{\infty} \boldsymbol{\mathcal{X}}_i) = 1$. Sea $g : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}} = g(\boldsymbol{\mathcal{X}})$ una

función tal que h_i, definida como la restricción de g a $\boldsymbol{\mathcal{X}}_i$, verifica las hipótesis del Teorema 4.4.1. Entonces $\boldsymbol{Y} = g(\boldsymbol{X})$ tiene densidad conjunta dada por

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \begin{cases} \sum_{i=1}^{\infty} f_{\boldsymbol{X}}(h_i^{-1}(\boldsymbol{y}))|\text{det}(Jh_i^{-1}(\boldsymbol{y}))| & \text{si } \boldsymbol{y} \in \boldsymbol{\mathcal{Y}} \\ 0 & \text{si no.} \end{cases} \tag{4.4.4}$$

Veamos una aplicación de este resultado.

Ejemplo 4.4.6 Sean X_1 y X_2 variables aleatorias i.i.d. con distribución común N(0,1). Mostremos que $Y_1 = X_1^2 + X_2^2$ e $Y_2 = X_1/X_2$ son independientes. El candidato natural para función g es en este caso $g(x_1, x_2) = (x_1^2 + x_2^2, x_1/x_2)$. Es claro, sin embargo, que esta función no es biyectiva. Por ejemplo, $g(1,1) = g(-1,-1)$. Además, no está definida para $(x_1, 0)$, cualquiera que sea $x_1 \in \mathbb{R}$. Claramente $\boldsymbol{\mathcal{X}} = \mathbb{R}^2$, y $g(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{Y}} = \mathbb{R}^2$. Consideremos ahora $\boldsymbol{\mathcal{X}}_1 = \{(x_1, x_2) : x_1 < 0\}$, y $\boldsymbol{\mathcal{X}}_2 = \{(x_1, x_2) : x_1 > 0\}$. Puesto que $P(\boldsymbol{X} \in \{(x_1, x_2) : x_1 = 0\}) = 0$ ($\boldsymbol{X}$ es absolutamente continua) se tiene que $P(\boldsymbol{X} \in \boldsymbol{\mathcal{X}}_1 \cup \boldsymbol{\mathcal{X}}_2) = 1$. Además, h_1 y h_2, las restricciones de g a $\boldsymbol{\mathcal{X}}_1$ y $\boldsymbol{\mathcal{X}}_2$ respectivamente, son claramente biyectivas y satisfacen las hipótesis del Teorema 4.4.2.

Hay otra faceta interesante de este problema, y que consiste en que h_i^{-1} no necesita ser determinado explícitamente. Note que

$$Jh_1^{-1}(y_1, y_2) = \left(Jh_1(h_1^{-1}(y_1, y_2))\right)^{-1},$$

de modo que

$$|\text{det}(Jh_1^{-1}(y_1, y_2))| = |\text{det}(Jh_1(h_1^{-1}(y_1, y_2)))|^{-1}.$$

Además,

$$Jh_1(x_1, x_2) = \begin{pmatrix} 2x_1 & 2x_2 \\ 1/x_2 & -x_1/x_2^2 \end{pmatrix},$$

y $|\text{det}(Jh_1(x_1, x_2))| = -2(x_1^2/x_2^2 + 1)$ y por lo tanto

$$|\text{det}(Jh_1(h_1^{-1}(y_1, y_2)))| = \frac{1}{2(y_2^2 + 1)}.$$

Análogamente,

$$|\text{det}(Jh_2(h_2^{-1}(y_1, y_2)))| = \frac{1}{2(y_2^2 + 1)},$$

y la densidad conjunta de (Y_1, Y_2) se obtiene de aplicar (4.4.4):

$$f_{X_1,X_2}(x_1, x_2) = \frac{\exp(-(x_1^2 + x_2^2)/2)}{2\pi},$$

y se tiene finalmente que

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{Y}}(h_1^{-1}(\boldsymbol{y}))|\text{det}(Jh_1^{-1}(\boldsymbol{y}))| + f_{\boldsymbol{Y}}(h_2^{-1}(\boldsymbol{y}))|\text{det}(Jh_2^{-1}(\boldsymbol{y}))| \\ &= \frac{\exp(-y_1/2)}{4\pi(1 + y_2^2)} + \frac{\exp(-y_1/2)}{4\pi(1 + y_2^2)} = \frac{\exp(-y_1/2)}{2\pi(1 + y_2^2)} \\ &= \frac{\exp(-y_1/2)}{2} \times \frac{1}{\pi(1 + y_2^2)}, \end{aligned}$$

y puesto que esta densidad conjunta factoriza como el producto de la densidad exponencial de parámetro 2, y de la densidad de Cauchy – definida en (3.8.2) –, concluimos que $Y_1 \sim \text{Exp}(2)$, e Y_2 tiene distribución de Cauchy, siendo ellas además, independientes.

4.4.4. Aplicación: Estadísticos de orden

Para finalizar esta sección, estudiaremos los *estadísticos de orden* asociados a una secuencia de variables aleatorias i.i.d. $X_1, \ldots, X_n$, definidas como sigue:

Definición 4.4.1 Considere $X_1, \ldots, X_n$ variables aleatorias i.i.d. con $X_i \sim F_X$. Los *estadísticos de orden* de esta muestra se definen como las variables aleatorias $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, donde $X_{(1)}(\omega), \ldots, X_{(n)}(\omega)$ se obtienen de ordenar $X_1(\omega), \ldots, X_n(\omega)$ de menor a mayor. En consecuencia, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, con $X_{(1)} = \text{mín}\{X_1, \ldots, X_n\}$, y $X_{(n)} = \text{máx}\{X_1, \ldots, X_n\}$.

Proposición 4.4.1 Supongamos que $X_1, \ldots, X_n$ son variables aleatorias i.i.d. con densidad común f_X y función de distribución común F_X, y con valores en $\mathcal{X}$. Entonces, la densidad conjunta de los estadísticos de orden está dada por

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^{n} f_X(x_i) & \text{si } x_1 < x_2 < \cdots < x_n \\ 0 & \text{si no.} \end{cases} \tag{4.4.5}$$

Demostración: Considere la función $g : \mathcal{X}^n \to \mathcal{X}^n$ dada por

$$g(x_1, x_2, \ldots, x_n) = \boldsymbol{x}_\pi = (x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n}),$$

donde $\pi = (\pi_1, \ldots, \pi_n)$ es una permutación que deja los elementos $x_1, \ldots, x_n$ ordenados *ascendentemente*, esto es, $x_{\pi_1} \leq x_{\pi_2} \leq \cdots \leq x_{\pi_n}$. Note que hay $n!$ permutaciones de $x_1, \ldots, x_n$. Por otra parte, los casos en que $x_i = x_j$ para algún $i \neq j$ pueden descartarse, pues tiene probabilidad 0. Así, si $\mathcal{P}$ es el conjunto de estas $n!$ permutaciones de $\{1, 2, \ldots, n\}$, tenemos que a $\pi \in \mathcal{P}$ se le asocia un subconjunto $\boldsymbol{\mathcal{X}}_\pi$ tal que si $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_\pi$ se cumple $x_{\pi_1} \leq \cdots \leq x_{\pi_n}$. Se tiene entonces que la función h_π definida como la restricción de g a $\boldsymbol{\mathcal{X}}_\pi$ es biyectiva y diferenciable. Más aún, la matriz Jacobiana de h_π es una permutación de las filas de la matriz identidad, y por lo tanto su determinante es ya sea 1 ó -1, y se tiene que $|\det(Jh_\pi(h_\pi^{-1}(x_\pi)))| = 1$ para todo $\pi \in \mathcal{P}$. Finalmente, se cumple que $P(X \in \bigcup_{\pi \in \mathcal{P}}) = 1$, y el resultado se tiene entonces como consecuencia inmediata de (4.4.4). ■

Veamos ahora algunas consecuencias de este resultado.

1. La densidad de $X_{(k)}$ está dada por

$$f_{X_{(k)}}(x_k) = \begin{cases} n\binom{n-1}{k-1} F_X(x_k)^{k-1}(1 - F_X(x_k))^{n-k} f_X(x_k) & \text{si } x \in \mathcal{X} \\ 0 & \text{si no.} \end{cases} \tag{4.4.6}$$

En efecto, la densidad conjunta (4.4.5) se puede integrar con respecto a $x_1, \ldots, x_{k-1}$ y a $x_{k+1}, \ldots, x_n$. Así, integrando x_n, con $x_1 < \cdots < x_{n-1}$, y con $x_i \in \mathcal{X}$ se tiene

$$f_{X_{(1)},\ldots,X_{(n-1)}}(x_1, \ldots, x_{n-1}) = n! \left\{ \prod_{i=1}^{n-1} f_X(x_i) \right\} (1 - F_X(x_{n-1})).$$

Integrando respecto de x_{n-1} se tiene para $x_1 < \cdots < x_{n-2}$:

$$f_{X_{(1)},\ldots,X_{(n-2)}}(x_1, \ldots, x_{n-2}) = \frac{n!}{2!} \left\{ \prod_{i=1}^{n-2} f_X(x_i) \right\} (1 - F_X(x_{n-2}))^2.$$

Por inducción, para $x_1 < \cdots < x_k$ se tiene:

$$f_{X_{(1)},\ldots,X_{(k)}}(x_1, \ldots, x_k) = \frac{n!}{(n-k)!} \left\{ \prod_{i=1}^{k} f_X(x_i) \right\} (1 - F_X(x_k))^{n-k}.$$

Ahora, integrando con respecto a x_1 se obtiene que para $x_2 < \cdots < x_k$:

$$f_{X_{(2)},\ldots,X_{(k)}}(x_2, \ldots, x_k) = \frac{n!}{(n-k)!} \left\{ \prod_{i=2}^{k} f_X(x_i) \right\} F_X(x_2)(1 - F_X(x_k))^{n-k}.$$

Integrando respecto de x_2 se encuentra que para $x_3 < \cdots < x_k$:

$$f_{X_{(3)},\ldots,X_{(k)}}(x_3, ..., x_k) = \frac{n!}{2!(n-k)!} \left\{ \prod_{i=3}^{k} f_X(x_i) \right\} F_X(x_3)^2(1 - F_X(x_k))^{n-k}$$

y finalmente, por inducción se obtiene el resultado.

2. El caso particular $k = 1$ corresponde al mínimo entre $\{X_1, \ldots, X_n\}$. En este caso, la densidad se obtiene de (4.4.6) con $k = 1$:

$$f_{X_{(1)}}(x_1) = \begin{cases} n(1 - F_X(x_1))^{n-1} f_X(x_1) & \text{si } x_1 \in \mathcal{X} \\ 0 & \text{si no.} \end{cases} \tag{4.4.7}$$

Análogamente, el caso $k = n$ corresponde al máximo entre $X_1, \ldots, X_n$. Por (4.4.6)

$$f_{X_{(n)}}(x_n) = \begin{cases} nF_X(x_n)^{n-1} f_X(x_n) & \text{si } x_1 \in \mathcal{X} \\ 0 & \text{si no.} \end{cases} \tag{4.4.8}$$

3. Una forma alternativa de derivar los resultados del punto anterior es la siguiente.

$$\begin{aligned} P(X_{(n)} \leq x) &= P(\text{máx}\{X_1, \ldots, X_n\} \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) \\ &= \prod_{i=1}^{n} P(X_i \leq x) \quad \text{(por independencia de } X_1, \ldots, X_n) \\ &= F_X(x)^n. \end{aligned}$$

Así,

$$\begin{aligned} f_{X_{(n)}}(x) &= \frac{d}{dx}F_{X_{(n)}}(x) = \frac{d}{dx}F_X(x)^n \\ &= nF_X(x)^{n-1}f_X(x) \text{ para } x \in \mathcal{X}. \end{aligned}$$

Por otra parte,

$$\begin{aligned} P(X_{(1)} > x) &= P(\text{mín}\{X_1, \ldots, X_n\} > x) = P(X_1 > x, \ldots, X_n > x) \\ &= \prod_{i=1}^{n} P(X_i > x) \quad \text{(por independencia de } X_1, \ldots, X_n\text{)} \\ &= (1 - F_X(x))^n. \end{aligned}$$

Así, $F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n$, y

$$\begin{aligned} f_{X_{(1)}}(x) &= \frac{d}{dx}F_{X_{(1)}}(x) = \frac{d}{dx}\left(1 - (1 - F_X(x))\right)^n \\ &= n(1 - F_X(x))^{n-1}f_X(x) \quad \text{para } x \in \mathcal{X}. \end{aligned}$$

4. La densidad conjunta de $X_{(1)}$ y $X_{(n)}$ se puede obtener de (4.4.5), mediante integrar las variables $x_2, \ldots, x_{n-1}$. Alternativamente, considere el siguiente razonamiento. El evento $(X_{(1)} > x_1, X_{(n)} < x_n)$ equivale a

$$\text{mín}\{X_1, \ldots, X_n\} > x_1, \;\; \text{máx}\{X_1, \ldots, X_n\} \leq x_n,$$

y por lo tanto

$$\begin{aligned} P(X_{(1)} > x_1, X_{(n)} \leq x_n) &= P(x_1 < X_1 \leq x_1, \ldots, x_n < X_n \leq x_n) \\ &= \prod_{i=1}^{n} P(x_1 < X_i \leq x_n) \\ &= (F_X(x_n) - F_X(x_1))^n. \end{aligned}$$

Note que

$$P(X_{(n)} \leq x_n) = P(X_{(1)} \leq x_1, X_{(n)} \leq x_n) + P(X_{(1)} > x_1, X_{(n)} \leq x_n),$$

de donde se obtiene que

$$\begin{aligned} F_{X_{(1)},X_{(n)}}(x_1, x_n) &= P(X_{(1)} \leq x_1, X_{(n)} \leq x_n) \\ &= P(X_{(n)} \leq x_n) - P(X_{(1)} > x_1, X_{(n)} \leq x_n) \\ &= F_X(x_n)^n - (F_X(x_n) - F_X(x_1))^n. \end{aligned}$$

Finalmente, la densidad conjunta en cuestión se obtiene de derivar parcialmente con respecto a cada argumento esta última expresión:

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = n(n-1)(F_X(x_n) - F_X(x_1))^{n-2}f_X(x_1)f_X(x_n), \tag{4.4.9}$$

para $x_1 < x_n$, y con $x_1, x_n \in \mathcal{X}$, y es claro que $f_{X_{(1)},X_{(n)}}(x_1, x_n)$ vale cero en caso contrario.

Veamos algunos ejemplos.

Ejemplo 4.4.7 Si $X_1, \ldots, X_n$ son i.i.d con distribución exponencial de parámetro $\lambda > 0$, entonces por (4.4.7), y recordando que $F_X(x) = 1-\exp(-x/\lambda)$, la densidad de $X_{(1)}$ está dada por

$$f_{X_{(1)}}(x) = \begin{cases} \frac{n}{\lambda}\exp(-nx/\lambda) & \text{si } x > 0 \\ 0 & \text{si no,} \end{cases}$$

y se tiene que $X_{(1)} \sim \text{Exp}(\lambda/n)$.

Ejemplo 4.4.8 Sean $X_1, \ldots, X_n$ variables aleatorias i.i.d. U(0,1), y sean $U = X_{(1)}$, y $V = X_{(n)}$. Por lo hecho anteriormente, se tiene que

$$f_{U,V}(u,v) = \begin{cases} n(n-1)(v-u)^{n-2} & \text{si } 0 \leq u < v \leq 1 \\ 0 & \text{si no.} \end{cases} \tag{4.4.10}$$

Ejemplo 4.4.9 Calculemos ahora la densidad de $X = V - U$ en el Ejemplo 4.4.8. Sea $(x,y) = g(u,v) = (v-u, v)$. El Jacobiano de esta transformación tiene determinante 1, y además $(u,v) = g^{-1}(x,y) = (x+y, y)$. Luego,

$$f_{X,Y}(x,y) = \begin{cases} n(n-1)x^{n-2} & \text{si } 0 \leq x \leq y \leq 1 \\ 0 & \text{si no,} \end{cases}$$

por lo que

$$f_X(x) = \int_x^1 n(n-1)x^{n-2}dy = n(n-1)x^{n-2}(1-x),$$

si $0 \leq x \leq 1$, y 0 si no. Se tiene entonces que $X \sim$ Beta$(n-1, 2)$. La variable X aquí considerada suele llamarse en Estadística el *rango* de las observaciones $X_1, \ldots, X_n$.

4.5. Valor Esperado de Vectores Aleatorios

4.5.1. Definición

Corresponde ahora definir el valor esperado de un vector aleatorio, y la correspondiente generalización del concepto de varianza.

Definición 4.5.1 Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ un vector aleatorio n-dimensional. El *vector de valores esperados* o *esperanza* de $\boldsymbol{X}$ se define mediante

$$E(\boldsymbol{X}) = E\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}, \tag{4.5.1}$$

provisto que todos los valores esperados en cuestión existan.

Se tiene entonces que la esperanza del vector aleatorio $\boldsymbol{X}$ es simplemente el vector de los valores esperados de cada componente.

4.5.2. Valor esperado de funciones de un vector aleatorio

El caso del valor esperado de una función del vector aleatorio $\boldsymbol{X}$ se trata a continuación.

Teorema 4.5.1 Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ un vector aleatorio n-dimensional, y sea $g : \mathbb{R}^n \to \mathbb{R}^m$ una función dada por

$$g(x_1, \ldots, x_n) = \begin{pmatrix} g_1(x_1, \ldots, x_n) \\ g_2(x_1, \ldots, x_n) \\ \vdots \\ g_m(x_1, \ldots, x_n) \end{pmatrix},$$

donde $g_1, \ldots, g_m$ son m funciones definidas en $\mathbb{R}^n$ y a valores reales. Entonces:

(a) Si $m = 1$, el valor esperado de $g(\boldsymbol{X})$ está dado por

$$E(g(\boldsymbol{X})) = \begin{cases} \sum\limits_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}} g(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) & \text{si } \boldsymbol{X} \text{ es discreto} \\ \int \cdots \int g(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) & \text{si } \boldsymbol{X} \text{ es continuo,} \end{cases} \tag{4.5.2}$$

provisto que la suma o integral múltiple converja absolutamente.

(b) Si $m \geq 2$ entonces

$$E(g(\boldsymbol{X})) = E(g(X_1, \ldots, X_n)) = \begin{pmatrix} E(g_1(x_1, \ldots, x_n)) \\ E(g_2(x_1, \ldots, x_n)) \\ \vdots \\ E(g_m(x_1, \ldots, x_n)) \end{pmatrix}, \tag{4.5.3}$$

provisto que todas los valores esperados en cuestión existan.

El resultado del Teorema 4.5.1 es simplemente la correspondiente generalización multivariada del Teorema 3.8.1 del capítulo anterior.

En el caso particular $m = 1$ y g definida por la suma de las coordenadas de $\boldsymbol{x}$, esto es, $g(x_1, \ldots, x_n) = \sum\limits_{i=1}^{n} x_i$ se puede probar que el Teorema 4.5.1 establece que si $E(X_i)$ existe para todo $i = 1, \ldots, n$, entonces $\sum\limits_{i=1}^{n} X_i$ también posee valor esperado y

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i). \tag{4.5.4}$$

Esto simplemente establece que la esperanza es lineal.

4.5.3. Valor esperado de productos de variables aleatorias independientes

Consideremos nuevamente el caso especial $m = 1$, y donde g está dada ahora por $g(x_1, \ldots, x_n) = x_1 x_2 \cdots x_n = \prod_{i=1}^{n} x_i$, esto es, el producto de las n coordenadas. Si $X_1, \ldots, X_n$ son además independientes, entonces, en el caso absolutamente continuo se tiene:

$$\begin{aligned} E(g(X_1, \ldots, X_n)) &= E(X_1 \cdots X_n) \\ &= \int \cdots \int x_1 \cdots x_n f_{X_1}(x_1) \cdots f_{X_n}(x_n) dx_1 \cdots dx_n \\ &= \int \cdots \int \left\{ \prod_{i=1}^{n} x_i f_{X_i}(x_i) \right\} dx_1 \cdots dx_n \\ &= \prod_{i=1}^{n} \left\{ \int_{\mathcal{X}} x_i f_{X_i}(x_i) dx_i \right\} = \prod_{i=1}^{n} E(X_i). \end{aligned}$$

Se puede probar que este resultado vale no sólo en el caso continuo, y así tenemos:

Proposición 4.5.1 Sean $X_1, \ldots, X_n$ variables aleatorias independientes cada una con valor esperado finito $E(X_i)$. Entonces $E(\prod_{i=1}^{n} X_i)$ también existe y

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i). \tag{4.5.5}$$

Nota: Es posible probar que una condición suficiente para asegurar la existencia de $E(XY)$, es que ambos X e Y posean segundos momentos, esto es, $E(X^2) < \infty$ y $E(Y^2) < \infty$.

Juntando los resultados de (4.5.4) y (4.5.5) podemos establecer lo siguiente:

Proposición 4.5.2 Sean $X_1, \ldots, X_n$ variables aleatorias independientes con segundos momentos finitos. Entonces

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i). \tag{4.5.6}$$

Demostración: Basta probar el caso $n = 2$, del que el resultado se obtiene por inducción sobre n. Por definición se cumple:

$$\begin{aligned} Var(X+Y) &= E(X+Y)^2 - (E(X+Y))^2 \\ &= E(X^2 + Y^2 + 2XY) - (E(X) + E(Y))^2 \\ &= Var(X) + Var(Y) + 2E(X)E(Y) - 2E(X)E(Y) \\ &= Var(X) + Var(Y). \quad \blacksquare \end{aligned}$$

Veamos a continuación algunos ejemplos.

Ejemplo 4.5.1 Sean X_1 y X_2 variables aleatorias i.i.d. U(0,1). Calculemos $E(X_{(1)})$ de dos formas diferentes. Primero, por (4.4.7) se tiene que

$$\begin{aligned} E(X_{(1)}) &= \int_0^1 x \cdot 2(1-x)dx = 2\int_0^1 x(1-x)dx \\ &= \frac{2\Gamma(2)\Gamma(2)}{\Gamma(2+2)} = \frac{2 \cdot 1!}{3!} = \frac{1}{3}. \end{aligned}$$

Nótese que este cálculo es inmediato debido a que (4.4.7) se había obtenido previamente. Por otra parte,

$$\begin{aligned} E(X_{(1)}) &= \int_0^1 \int_0^1 \text{mín}\{x_1, x_2\} \cdot 1 dx_1 dx_2 \\ &= \int_0^1 \int_0^{x_2} x_1 dx_1 dx_2 + \int_0^1 \int_{x_2}^1 x_2 dx_1 dx_2 \\ &= \int_0^1 \frac{x_2^2}{2} dx_2 + \int_0^1 x_2(1-x_2)dx_2 \\ &= \frac{1}{6} + \frac{1}{2} - \frac{1}{3} = \frac{1}{3} \end{aligned}$$

Ejemplo 4.5.2 Sean X, Y, Z i.i.d. U(0,1), y defina $W = (X+Y)Z$. Calculemos $E(W)$ y $Var(W)$. Tenemos que $E(X) = 1/2$, y $E(X^2) = 1/3$.

$$\begin{aligned} E(W) &= E(XZ + YZ) = E(XZ) + E(YZ) \\ &= E(X)E(Z) + E(Y)E(Z) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

Por otra parte,

$$\begin{aligned} E(W^2) &= E((X+Y)^2 Z^2) = E(Z^2)E(X^2 + Y^2 + 2XY) \\ &= \frac{1}{3}\left(E(X^2) + E(Y^2) + 2E(X)E(Y)\right) \\ &= \frac{1}{3}\left(\frac{1}{3} + \frac{1}{3} + \frac{1}{2}\right) \\ &= \frac{7}{18}, \end{aligned}$$

de donde se sigue que

$$Var(W) = E(W^2) - E(W)^2 = \frac{7}{18} - \frac{1}{4} = \frac{5}{36} \approx 0{,}1389$$

Ejemplo 4.5.3 Considere un punto (X, Y) distribuido uniformemente en el círculo unitario centrado en el origen. ¿Cuál es la distancia media de este punto al origen?. ¿Cuál es la varianza?.

Tenemos que

$$f_{X,Y}(x,y) = \begin{cases} \pi^{-1} & \text{si } x^2+y^2 \leq 1 \\ 0 & \text{si no.} \end{cases}$$

La distancia desde (X,Y) al origen está dada por $R = \sqrt{X^2+Y^2}$, de modo que necesitamos $E(R)$. Por (4.5.2):

$$E(R) = \pi^{-1} \iint\limits_{\{(x,y):x^2+y^2\leq 1\}} \sqrt{x^2+y^2}dxdy.$$

Aún cuando esta integral se puede calcular directamente, es conveniente cambiar las variables de integración a coordenadas polares. Así,

$$E(R) = \pi^{-1} \int_0^1 \int_{-\pi}^{\pi} r \cdot rdrd\theta = \pi^{-1} \cdot 2\pi \int_0^1 r^2 dr = \frac{2}{3}.$$

Similarmente,

$$E(R^2) = \pi^{-1} \int_0^1 \int_{-\pi}^{\pi} r^2 \cdot rdrd\theta = \pi^{-1} \cdot 2\pi \int_0^1 r^3 dr = \frac{1}{2},$$

de donde se obtiene que $Var(R) = 1/18$.

Ejemplo 4.5.4 Consideremos $X_1, X_2, \ldots$ variables aleatorias i.i.d. con distribución Bernoulli(p), y sea N independiente de éstas, con distribución Poisson(λ), donde $\lambda > 0$. Considere la variable aleatoria

$$S_N = \sum_{i=1}^{N} X_i,$$

la cual se puede interpretar como determinar un número aleatorio de variables aleatorias con distribución Bernoulli(p) de acuerdo a la distribución de N, y luego sumarlas. Para efectuar el cálculo de $E(S_N)$, es conveniente considerar la distribución conjunta de (S_N, N). Así, note que para $k \leq n$:

$$\begin{aligned} P(S_N = k, N = n) &= P(S_N = k \mid N = n)P(N = n) \\ &= P(S_n = k \mid N = n)P(N = n) = P(S_n = k)P(N = n) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{\lambda^n \exp(-\lambda)}{n!} = \frac{p^k(1-p)^{n-k}\lambda^n \exp(-\lambda)}{k!(n-k)!}. \end{aligned}$$

Note que $P(S_n = k \mid N = n) = P(S_n = k)$, pues, una vez que el número de variables a sumar se fija, la dependencia en N se elimina. La razón es que $X_1, X_2, \ldots$

es independiente de N. Luego,

$$\begin{aligned}
E(S_N) &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} k \cdot \frac{p^k(1-p)^{n-k}\lambda^n \exp(-\lambda)}{k!(n-k)!} \\
&= \sum_{n=1}^{\infty}\sum_{k=1}^{n} \frac{p^k(1-p)^{n-k}\lambda^n \exp(-\lambda)}{(k-1)!(n-k)!} \\
&= p\sum_{n=1}^{\infty}\lambda^n \exp(-\lambda)\sum_{j=0}^{n-1}\frac{p^j(1-p)^{n-1-j}}{j!(n-1-j)!} \qquad (j=k-1) \\
&= p\lambda\sum_{n=1}^{\infty}\frac{\lambda^{n-1}\exp(-\lambda)}{(n-1)!}\left(\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j}\right) \\
&= p\lambda\sum_{n=1}^{\infty}\frac{\lambda^{n-1}\exp(-\lambda)}{(n-1)!} \\
&= p\lambda.
\end{aligned}$$

Note que el resultado obtenido coincide con $E(X_1)E(N)$, lo cual, lejos de ser una coincidencia, tiene una justificación que se verá más adelante. Para calcular $Var(S_N)$, obtengamos primero $E(S_N(S_N-1))$. Se tiene:

$$\begin{aligned}
E(S_N(S_N-1)) &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} k(k-1) \cdot \frac{p^k(1-p)^{n-k}\lambda^n \exp(-\lambda)}{k!(n-k)!} \\
&= \sum_{n=2}^{\infty}\sum_{k=2}^{n} \frac{p^k(1-p)^{n-k}\lambda^n \exp(-\lambda)}{(k-2)!(n-k)!} \\
&= p^2\sum_{n=2}^{\infty}\lambda^n \exp(-\lambda)\sum_{j=0}^{n-2}\frac{p^j(1-p)^{n-2-j}}{j!(n-2-j)!} \\
&\qquad \text{(Note el cambio } j=k-2\text{)} \\
&= p^2\sum_{n=2}^{\infty}\frac{\lambda^n\exp(-\lambda)}{(n-2)!}\sum_{j=0}^{n-2}\binom{n-2}{j}p^j(1-p)^{n-2-j} \\
&= p^2\lambda^2\sum_{n=2}^{\infty}\frac{\lambda^{n-2}\exp(-\lambda)}{(n-2)!} \\
&= p^2\lambda^2.
\end{aligned}$$

Luego,

$$E(S_N^2) = E(S_N(S_N-1)) + E(S_N) = p^2\lambda^2 + p\lambda,$$

de donde, finalmente:

$$Var(S_N) = E(S_N^2) - E(S_N)^2 = p\lambda.$$

Una forma alternativa de derivar este resultado consiste en calcular directamente la distribución de S_N. Se propone como ejercicio demostrar que en este caso $S_N \sim$ Poisson$(p\lambda)$, de donde el resultado se sigue inmediatamente.

4.5.4. Covarianza y coeficiente de correlación

En el caso univariado, vimos que la varianza proporciona una idea de la *dispersión* de la distribución de la variable aleatoria considerada. Cuando se trabaja con un vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)$, la varianza de cada uno de los X_i no proporciona una visión completa de la dispersión de la distribución conjunta, ni da una idea del grado de dependencia que pueda haber entre las variables. Recurrimos entonces a la versión multivariada de varianza, llamada *matriz de varianza-covarianza* de $\boldsymbol{X}$.

Definición 4.5.2 La *matriz de varianza-covarianza*, o simplemente *matriz de covarianza* de $\boldsymbol{X}$ se define mediante

$$V(\boldsymbol{X}) = E\left((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))'\right), \tag{4.5.7}$$

provisto que todos los valores esperados en cuestión existan.

La matriz de covarianza de $\boldsymbol{X}$ tiene una estructura novedosa. El elemento $V(\boldsymbol{X})_{i,j}$ con $i, j = 1, \ldots, n$ corresponde a

$$\begin{aligned} V(\boldsymbol{X})_{i,j} &= E\left((X_i - E(X_i))(X_j - E(X_j))\right) \\ &= E(X_iX_j) - E\left(E(X_i)X_j\right) - E\left(X_iE(X_j)\right) + E\left(E(X_i)E(X_j)\right) \\ &= E(X_iX_j) - E(X_i)E(X_j) - E(X_i)E(X_j) + E(X_i)E(X_j) \\ &= E(X_iX_j) - E(X_i)E(X_j), \end{aligned}$$

asumiendo que todos estos valores esperados existen. En el caso que $i = j$, esto se reduce simplemente a la *varianza* de X_i. En el caso $i \neq j$, nos referiremos a esta cantidad como la *covarianza* entre X_i y X_j, de acuerdo a la siguiente definición formal.

Definición 4.5.3

1. La *covarianza* entre las variables aleatorias X e Y se define como

$$Cov(X, Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(X), \tag{4.5.8}$$

   provisto que los valores esperados en cuestión existan. Es inmediato ver que en este caso se tiene $Cov(X, Y) = Cov(Y, X)$, esto es, la covarianza, vista como una función de dos variables aleatorias, es *simétrica*.

2. El *coeficiente de correlación* entre X e Y se define como

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}, \tag{4.5.9}$$

   provisto que todas las cantidades en cuestión existan.

El concepto de covarianza se puede también extender a vectores aleatorios.

Definición 4.5.4 Si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son vectores aleatorios de dimensión n y m respectivamente, se define la matriz de covarianza entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ mediante

$$Cov(\boldsymbol{X}, \boldsymbol{Y}) = E((\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{X} - E(\boldsymbol{X}))'). \tag{4.5.10}$$

Así, $Cov(\boldsymbol{X}, \boldsymbol{Y})$ es una matriz de $n \times m$ cuyo elemento (i, j) es $Cov(X_i, Y_j)$. Note que $Cov(\boldsymbol{X}, \boldsymbol{Y}) = Cov(\boldsymbol{Y}, \boldsymbol{X})'$, y que $Cov(\boldsymbol{X}, \boldsymbol{X})$ es simplemente la matriz de varianza-covarianza de $\boldsymbol{X}$.

Veamos ahora algunas propiedades relacionadas a estos conceptos.

1. Si X e Y son independientes, entonces

   $$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0,$$

   y en consecuencia $\rho(X, Y) = 0$. En general, si $Cov(X, Y) = 0$, diremos que X e Y son *no correlacionadas*.

2. Se tiene que para todo a, b, c, d, números reales, y puesto que $E(a + bX) = a + bE(X)$, $E(c + dY) = c + dE(Y)$, entonces:

   $$\begin{aligned} Cov(a + bX, c + dY) &= E\left\{b(X - E(X))d(Y - E(Y))\right\} \\ &= bdE((X - E(X))(Y - E(Y))) \\ &= bdCov(X, Y), \end{aligned}$$

   y puesto que $Var(a + bX) = b^2 Var(X)$, y $Var(c + dY) = d^2 Var(Y)$, entonces si además $b \neq 0$ y $d \neq 0$ se cumple:

   $$\rho(a + bX, c + dY) = \text{sign}(b)\text{sign}(d)\rho(X, Y).$$

   En particular, si $b, d > 0$, $\rho(a + bX, c + dY) = \rho(X, Y)$.

   En otras palabras, el coeficiente de correlación es invariante bajo cambios de escala y localización.

3. Si X e Y son no correlacionadas, ello no implica que sean independientes, como lo muestra el siguiente ejemplo. Sea $X \sim U(-1, 1)$, e $Y = X^2$. Es claro que $E(X) = 0$, y que

   $$E(XY) = E(X^3) = \int_{-1}^{1} \frac{x^3}{2} dx = 0,$$

   de modo que $Cov(X, Y) = 0$, pero es claro que X e Y no pueden ser independientes.

4. Si $E(X^2) < \infty$ entonces $Cov(X, X) = E(X^2) - (E(X))^2 = Var(X)$.

5. Si las expresiones involucradas existen, entonces

$$\begin{aligned} Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) &= E(\sum_{i=1}^{n} X_i \sum_{j=1}^{m} Y_j) - E(\sum_{i=1}^{n} X_i)E(\sum_{j=1}^{m} Y_j) \\ &= E(\sum_{i=1}^{n}\sum_{j=1}^{m} X_iY_j) - \sum_{i=1}^{n} E(X_i) \sum_{j=1}^{m} E(Y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} E(X_iY_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} E(X_i)E(Y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} (E(X_iY_j) - E(X_i)E(Y_j)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j), \end{aligned}$$

   lo que muestra que la covarianza, visto como una función de dos variables aleatorias es *bilinieal*.

6. **Desigualdad de Cauchy-Schwartz:**

$$|Cov(X,Y)| \leq \sqrt{Var(X)}\sqrt{Var(Y)}.$$

   En particular, se tiene que

$$-1 \leq \rho(X,Y) \leq 1,$$

   cualesquiera que sean X e Y, y asumiendo que las cantidades involucradas existen.

7. Si $Y = a + bX$, con $b \neq 0$ entonces:

$$\begin{aligned} Cov(X,Y) &= Cov(a+bX, X) = E((a+bX)X) - E(a+bX)E(X) \\ &= aE(X) + bE(X^2) - aE(X) - bE(X)^2 \\ &= b(E(X^2) - E(X)^2) = bVar(X), \end{aligned}$$

   de donde se sigue que

$$|\rho(X,Y)| = 1.$$

   Es decir, si Y se obtiene de una transformación lineal afín de X, entonces el coeficiente de correlación entre X e Y es 1 o -1, dependiendo del signo de b. Esto muestra que $\rho(X,Y)$ mide el grado de dependencia lineal que existe entre X e Y, correspondiendo el caso extremo (esto es, $|\rho(X,Y)| = 1$) a la dependencia lineal perfecta.

8. Sea $\boldsymbol{X} = (X_1, \ldots, X_n)$ y $\boldsymbol{A} = (a_{i,j})$ una matriz de $n \times n$. Defina $\boldsymbol{Y} = \boldsymbol{AX}$, donde los vectores son interpretados como columnas. Suponga que $E(\boldsymbol{X}) = \boldsymbol{\mu}$ y que $V(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Puesto que $Y_k = \sum_{j=1}^{n} a_{k,j}X_j$, se tiene que $E(Y_k) = \sum_{j=1}^{n} a_{k,j}E(X_j)$, y de aquí se obtiene que

$E(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\mu}$. Por otra parte,

$$\begin{aligned} Cov(Y_k, Y_l) &= Cov(\sum_{j=1}^{n} a_{k,j}X_j, \sum_{m=1}^{n} a_{l,m}X_m) \\ &= \sum_{j=1}^{n}\sum_{m=1}^{n} a_{k,j}Cov(X_j, X_m)a_{m,l}, \end{aligned}$$

y hemos así probado las fórmulas

$$E(\boldsymbol{Y}) = \boldsymbol{A}E(\boldsymbol{X}) \qquad \text{y} \qquad V(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'. \tag{4.5.11}$$

Es directo ver que estas propiedades también valen en el caso en que $\boldsymbol{A}$ es una matriz cualquiera, no necesariamente cuadrada.

9. Sean $\boldsymbol{X} \in \mathbb{R}^n$, $\boldsymbol{Y} \in \mathbb{R}^m$ vectores aleatorios con matriz de covarianza $\boldsymbol{C} = Cov(\boldsymbol{X}, \boldsymbol{Y})$. Considere matrices $\boldsymbol{A}$ de $k \times n$ y $\boldsymbol{B}$ de $l \times m$. La i-ésima coordenada de $\boldsymbol{AX}$ es $\sum_{s=1}^{n} A_{is}X_s$, la j-ésima coordenada de $\boldsymbol{BY}$ es $\sum_{t=1}^{m} B_{jt}Y_t$, y la covarianza entre estas coordenadas es

$$Cov(\sum_{s=1}^{n} A_{is}X_s, \sum_{t=1}^{m} B_{jt}Y_t) = \sum_{s=1}^{n}\sum_{t=1}^{m} A_{is}Cov(X_s, Y_t)B_{jt},$$

de donde se puede obtener directamente que

$$Cov(\boldsymbol{AX}, \boldsymbol{BY}) = \boldsymbol{ACB}'. \tag{4.5.12}$$

Note que la segunda ecuación en (4.5.11) se puede obtener como caso particular de (4.5.12).

10. Para vectores aleatorios $\boldsymbol{X}$ e $\boldsymbol{Y}$, y para vectores y matrices $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{A}$, $\boldsymbol{B}$ con dimensiones apropiadas, se tiene que

$$Cov(\boldsymbol{a} + \boldsymbol{AX}, \boldsymbol{b} + \boldsymbol{BY}) = \boldsymbol{A}Cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}',$$

que es una ligera generalización de (4.5.12).

11. Se propone como ejercicio mostrar que

$$E(\boldsymbol{XY}') = Cov(\boldsymbol{X}, \boldsymbol{Y}) + E(\boldsymbol{X})(E(\boldsymbol{Y}))',$$

y que si además $\boldsymbol{X}$ e $\boldsymbol{Y}$ tienen las mismas dimensiones, entonces

$$V(\boldsymbol{X} + \boldsymbol{Y}) = V(\boldsymbol{X}) + V(\boldsymbol{Y}) + Cov(\boldsymbol{X}, \boldsymbol{Y}) + Cov(\boldsymbol{X}, \boldsymbol{Y})'.$$

Veamos a continuación algunos ejemplos.

Ejemplo 4.5.5 Sean $X_1, \ldots, X_n$ i.i.d. U(0,1). Calculemos el coeficiente de correlación entre $X_{(1)}$ y $X_{(n)}$. Por (4.4.10), se tiene que

$$\begin{aligned}
E(X_{(1)}X_{(n)}) &= n(n-1)\int_0^1\int_0^v uv(v-u)^{n-2}dudv \quad (w=u/v)\\
&= n(n-1)\int_0^1 v\left(\int_0^1 wv^2(v-wv)^{n-2}dw\right)dv\\
&= n(n-1)\int_0^1 v^{n+1}dv\cdot\int_0^1 w(1-w)^{n-2}dw\\
&= n(n-1)\cdot\frac{1}{n+2}\cdot B(2,n-1)\\
&= \frac{n(n-1)}{n+2}\cdot\frac{1!(n-2)!}{n!}\\
&= \frac{1}{n+2}.
\end{aligned}$$

Por otra parte,

$$\begin{aligned}
E(X_{(1)}) &= n(n-1)\int_0^1\int_0^v u(v-u)^{n-2}dudv \quad (\text{ tome } w=u/v)\\
&= n(n-1)\int_0^1 1\left(\int_0^1 wv^2(v-wv)^{n-2}dw\right)dv\\
&= n(n-1)\int_0^1 v^n dv\cdot\int_0^1 w(1-w)^{n-2}dw\\
&= n(n-1)\cdot\frac{1}{n+1}\cdot B(2,n-1) = \frac{n(n-1)}{n+1}\cdot\frac{1!(n-2)!}{n!}\\
&= \frac{1}{n+1}.
\end{aligned}$$

Además:

$$\begin{aligned}
E(X_{(n)}) &= n(n-1)\int_0^1\int_0^v v(v-u)^{n-2}dudv \quad (\text{ tome } w=u/v)\\
&= n(n-1)\int_0^1 v\left(\int_0^1 v(v-wv)^{n-2}dw\right)dv\\
&= n(n-1)\int_0^1 v^n dv\cdot\int_0^1 (1-w)^{n-2}dw\\
&= n(n-1)\cdot\frac{1}{n+1}\cdot\frac{1}{n-1}\\
&= \frac{n}{n+1},
\end{aligned}$$

por lo que

$$Cov(X_{(1)},X_{(n)}) = \frac{1}{n+2}-\frac{1}{n+1}\cdot\frac{n}{n+1} = \frac{1}{(n+1)^2(n+2)}.$$

En forma análoga se prueba que

$$E(X_{(1)}^2) = \frac{2}{(n+1)(n+2)} \quad \text{y} \quad E(X_{(n)}^2) = \frac{n}{n+2},$$

de donde

$$Var(X_{(1)}) = Var(X_{(n)}) = \frac{n}{(n+1)^2(n+2)},$$

y, finalmente,

$$\rho(X_{(1)}, X_{(n)}) = \frac{1}{n}.$$

Ejemplo 4.5.6 Sea (X, Y) con densidad conjunta dada por

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

donde $(x, y) \in \mathbb{R}^2$, y en donde $-1 < \rho < 1$. Esta densidad corresponde a una forma de la distribución normal bivariada, como ya ha sido mencionado en el Ejemplo 4.3.1. Calculemos ahora $Cov(X, Y)$. Puesto que sabemos que marginalmente ambos X e Y tienen distribución $N(0, 1)$, sólo necesitamos calcular $E(XY)$. Se tiene:

$$\begin{aligned}
E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{xy}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)}\right) dxdy \\
&= \int_{-\infty}^{\infty} \frac{y \exp\left(-\frac{y^2}{2}\right)}{2\pi\sqrt{1-\rho^2}} \left(\int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx\right) dy \\
&= \int_{-\infty}^{\infty} \frac{y \exp\left(-\frac{y^2}{2}\right)}{2\pi\sqrt{1-\rho^2}} \rho y \sqrt{2\pi}\sqrt{1-\rho^2} dy \\
&= \frac{\rho}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \exp\left(-\frac{y^2}{2}\right) dy = \rho E(Y^2) \\
&= \rho.
\end{aligned}$$

Es fácil ver que de aquí uno puede concluir que $Cov(X, Y) = \rho$, y finalmente,

$$\rho(X, Y) = \rho.$$

Por otra parte, en el Ejemplo 4.3.1 habíamos ya probado que X e Y son independientes sí y sólo si $\rho = 0$, lo cual se traduce en que X e Y con distribución normal bivariada son independientes sí y sólo si ellas son no correlacionadas.

4.6. Funciones Generadoras Revisitadas

4.6.1. Funciones Generadoras e Independencia

Es muy frecuente – en la práctica – encontrar aplicaciones en que el resultado de un experimento corresponde a la suma de ciertas variables aleatorias independientes. El caso más típico es

el promedio de un número de variables aleatorias i.i.d. A continuación veremos una propiedad muy simple de las funciones generadoras que dice relación con esta situación.

Proposición 4.6.1 Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes, y sea $S_n = \sum_{i=1}^{n} X_i$. En la medida que las expresiones siguientes existan, se tiene:

1. $M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$.
2. $\Phi_{S_n}(t) = \prod_{i=1}^{n} \Phi_{X_i}(t)$.
3. $G_{S_n}(t) = \prod_{i=1}^{n} G_{X_i}(t)$.
4. $K_{S_n}(t) = \sum_{i=1}^{n} K_{X_i}(t)$.

La demostración de este resultado es muy simple, y está basada en el hecho que si X e Y son independientes, entonces $E(XY) = E(X)E(Y)$. Queda ésta propuesta como ejercicio.

Un caso particularmente importante es cuando $X_1, X_2, \ldots, X_n$ son i.i.d. En este caso, tenemos que $M_{X_i}(t) = M_{X_1}(t)$ para $i = 2, 3, \ldots, n$, y entonces los resultados de la Proposición 4.6.1 se reducen a:

1. $M_{S_n}(t) = (M_{X_1}(t))^n$.
2. $\Phi_{S_n}(t) = (\Phi_{X_1}(t))^n$.
3. $G_{S_n}(t) = (G_{X_1}(t))^n$.
4. $K_{S_n}(t) = nK_{X_1}(t)$.

Veamos algunas aplicaciones de estos resultados.

Ejemplo 4.6.1 De acuerdo a lo visto en el Ejemplo 3.3.3, se concluye que si $X \sim$ Poisson(λ), entonces $M_X(t) = \exp(\lambda(\exp(t) - 1))$, con $t \in \mathbb{R}$. Si $X_1, \ldots, X_n$ son independientes, con $X_i \sim$ Poisson(λ_i), se tiene que

$$M_{S_n}(t) = \prod_{i=1}^{n} \exp(\lambda_i(\exp(t) - 1)) = \exp\left((\sum_{i=1}^{n} \lambda_i)(\exp(t) - 1)\right),$$

de donde se sigue que $S_n \sim$ Poisson($\sum_{i=1}^{n} \lambda_i$).

Ejemplo 4.6.2 Sean $X_1, \ldots, X_n$ independientes, y tales que $X_i \sim N(\mu_i, \sigma_i^2)$. Entonces por lo hecho en el Ejemplo 3.8.12 se tiene que $M_{X_i}(t) = \exp(t\mu_i + \sigma_i^2 t^2/2)$, y entonces

$$M_{S_n}(t) = \prod_{i=1}^{n} \exp(t\mu_i + \sigma_i^2 t^2/2) = \exp\left(t \sum_{i=1}^{n} \mu_i + (t^2/2) \sum_{i=1}^{n} \sigma_i^2 \right),$$

por lo que $S_n \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$. Si definimos $\overline{X}_n$ como el promedio de $X_1, \ldots, X_n$, entonces $\overline{X}_n = n^{-1} S_n$, y por el resultado de la Proposición 3.8.1(c) se tiene que

$$M_{\overline{X}_n}(t) = \exp\left(t n^{-1} \sum_{i=1}^{n} \mu_i + (t^2/2) n^{-2} \sum_{i=1}^{n} \sigma_i^2 \right),$$

de donde se sigue que $\overline{X}_n \sim N(\frac{1}{n} \sum_{i=1}^{n} \mu_i, \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2)$. En el caso particular que las variables son i.i.d., entonces $\mu_1 = \cdots = \mu_n = \mu$, y $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$, y es fácil ver que $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

Ejemplo 4.6.3 Sean $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$, y defina la variable aleatoria $Y = \sum_{j=1}^{n} X_j^2$. Entonces, si $t < 1/2$:

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = \prod_{j=1}^{n} M_{X_j^2}(t) = \left(M_{X_1^2}(t) \right)^n \\ &= \left(\int_{-\infty}^{\infty} \frac{e^{-x^2(1/2-t)}}{\sqrt{2\pi}} dx \right)^n \\ &= \left(\frac{1}{(1/2-t)^{1/2}} \right)^n = \frac{1}{(1/2-t)^{n/2}}, \end{aligned}$$

de donde se sigue que $Y \sim \Gamma(n/2, 2)$. A pesar de ser un caso particular de distribución Gama, la distribución de Y recibe también el nombre de distribución chi-cuadrado con n grados de libertad, lo que se denota $Y \sim \chi^2(n)$, y como se mostró en este ejemplo, corresponde a la suma de los cuadrados de n variables aleatorias i.i.d. con distribución $N(0,1)$ (ver Ejemplo 4.4.1). Como consecuencia de las propiedades de la distribución Gama, se tiene que $E(Y) = n$ y $Var(Y) = 2n$.

Ejemplo 4.6.4 Sea $X \sim \text{BN}(k, p)$. Por lo visto en la Sección 3.7, la distribución binomial negativa es la distribución de T_k, el instante del k-ésimo éxito en una secuencia de ensayos de Bernoulli. Por otra parte, también se vio que las variables $T_1, T_2 - T_1, T_3 - T_2, \ldots, T_k - T_{k-1}, \ldots$ son i.i.d. con distribución geométrica de parámetro p. Pero

$$T_k = T_1 + (T_2 - T_1) + (T_3 - T_2) + \cdots + (T_k - T_{k-1}), \tag{4.6.1}$$

de modo que T_k es simplemente la suma de k variables aleatorias i.i.d. con distribución Geom(p). Por lo hecho en el Ejemplo 3.8.11, se tiene que

$$M_{T_1}(t) = \frac{p\exp(t)}{1-(1-p)\exp(t)},$$

provisto que $t < -\log(p)$. Es entonces inmediato concluir que

$$M_X(t) = M_{T_k}(t) = (M_{T_1}(t))^k = \left(\frac{p\exp(t)}{1-(1-p)\exp(t)}\right)^k,$$

la cual corresponde a la función generadora de momentos para esta distribución. Para obtener esperanza y varianza de X (o, lo que es lo mismo, de T_k), hay varias alternativas. Primero, se puede aplicar directamente el resultado de (3.8.5), lo cual se propone como ejercicio. Por otra parte, recordemos que

$$E(T_1) = \frac{1}{p} \quad \text{y} \quad Var(T_1) = \frac{1-p}{p^2},$$

de modo que de (4.6.1) se sigue inmediatamente que

$$E(T_k) = E(T_1) + E(T_2 - T_1) + \cdots E(T_k - T_{k-1}) = \frac{k}{p},$$

y que

$$Var(T_k) = Var(T_1) + Var(T_2 - T_1) + \cdots Var(T_k - T_{k-1}) = \frac{k(1-p)}{p^2},$$

donde usamos el hecho que $T_1, T_2 - T_1, \ldots, T_k - T_{k-1}$ son i.i.d., y (4.5.6).

4.6.2. Funciones Generadoras Multivariadas

Definimos a continuación la contraparte multivariada de las funciones generadora de momentos y característica, vistas en las subsecciones 3.8.2 y 3.8.3.

Definición 4.6.1 Sea $\boldsymbol{X} \in \mathbb{R}^n$ un vector aleatorio. En la medida que las expresiones involucradas existan, se define:

(a) La *función generadora de momentos multivariada* de $\boldsymbol{X}$ mediante

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E(e^{\boldsymbol{t}'\boldsymbol{X}}), \qquad \boldsymbol{t} \in \mathbb{R}^n, \tag{4.6.2}$$

donde $\boldsymbol{t}'\boldsymbol{X} = t_1X_1 + \cdots + t_nX_n$.

(b) La *función característica multivariada* de $\boldsymbol{X}$ mediante

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}'\boldsymbol{X}}), \qquad \boldsymbol{t} \in \mathbb{R}^n, \tag{4.6.3}$$

donde, como antes, i es el número complejo $\sqrt{-1}$.

Resumimos a continuación las propiedades más importantes de estas funciones.

1. Al igual que en el caso $n = 1$, la función característica multivariada está siempre bien definida, cualquiera que sea $\boldsymbol{t} \in \mathbb{R}^n$. No ocurre lo mismo con la función generadora de momentos multivariada, pues su existencia depende, en general, de $\boldsymbol{t}$.

2. Si la función generadora de momentos existe en una vecindad de $\boldsymbol{t} = \mathbf{0}$, entonces para enteros $k_1, \ldots, k_n$ no todos nulos se tiene

$$\frac{\partial^{k_1+\cdots+k_n}}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} M_{\boldsymbol{X}}(\boldsymbol{t}) \mid_{\boldsymbol{t}=\mathbf{0}} = E(X_1^{k_1} \cdots X_n^{k_n}). \tag{4.6.4}$$

   Análogamente, se tiene que si el valor esperado en cuestión existe, entonces

$$\frac{\partial^{k_1+\cdots+k_n}}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} \varphi_{\boldsymbol{X}}(\boldsymbol{t}) \mid_{\boldsymbol{t}=\mathbf{0}} = i^{k_1+\cdots+k_n} E(X_1^{k_1} \cdots X_n^{k_n}). \tag{4.6.5}$$

3. **Teorema de Caracterización:** Si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son vectores aleatorios tales que $\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{Y}}(\boldsymbol{t})$ para todo $\boldsymbol{t} \in \mathbb{R}^n$, entonces $F_{\boldsymbol{X}}$ y $F_{\boldsymbol{Y}}$ coinciden, es decir, tienen la misma distribución. Puesto que la recíproca es obviamente cierta, se tiene entonces una relación uno a uno entre la distribución y la función característica de vectores aleatorios.

4. Para obtener la función característica o generadora de momentos (univariada o multivariada) marginal de una parte del vector aleatorio, basta con tomar como cero las coordenadas correspondientes a la parte no deseada. Por ejemplo, $\varphi_{X_1}(t_1) = \varphi_{(X_1,X_2,\ldots,X_n)}(t_1, 0, \ldots, 0)$.

5. Sean $\boldsymbol{X} = (X_1, \ldots, X_n)$ e $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ vectores aleatorios, y defina el vector aleatorio $(n+m)$-dimensional $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$. Entonces $\boldsymbol{X}$ e $\boldsymbol{Y}$ son independientes si y sólo si cualquiera que sean los números reales $t_1, \ldots, t_n, t_{n+1}, \ldots, t_{n+m}$ se cumple

$$\varphi_{\boldsymbol{Z}}(t_1, \ldots, t_n, t_{n+1}, \ldots, t_{n+m}) = \varphi_{\boldsymbol{X}}(t_1, \ldots, t_n) \varphi_{\boldsymbol{Y}}(t_{n+1}, \ldots, t_{n+m}).$$

   Esta propiedad establece que independencia de dos vectores aleatorios es equivalente a poder factorizar la función característica conjunta de ambos vectores. El resultado se puede generalizar a tres o más vectores sin mayor dificultad. En particular, se tiene que las variables aleatorias $X_1, \ldots, X_n$ son independientes si y sólo si para cualquier $t_1, \ldots, t_n \in \mathbb{R}$ se tiene

$$\varphi_{(X_1,\ldots,X_n)}(t_1, \ldots, t_n) = \varphi_{X_1}(t_1) \cdots \varphi_{X_n}(t_n).$$

Veamos a continuación algunos ejemplos.

Ejemplo 4.6.5 Sea $\boldsymbol{X} = (X_1, \ldots, X_m)$ con distribución *multinomial*, cuya función de probabilidad está dada por

$$p_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{n!}{x_1! x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m},$$

en donde $p_1, \ldots, p_m$ son números reales no negativos con $\sum_{j=1}^{m} p_j = 1$, n es un entero positivo, y $x_1, \ldots, x_m$ son enteros no negativos tales que $\sum_{j=1}^{m} x_j = n$. Dado $\boldsymbol{t} \in \mathbb{R}^m$ se tiene que

$$\begin{aligned} E(e^{\boldsymbol{t}'\boldsymbol{X}}) &= E(e^{t_1X_1+\cdots+t_mX_m}) \\ &= \sum_{x_1,\ldots,x_m} \frac{n!}{x_1!\cdots x_m!}(p_1e^{t_1})^{x_1}\cdots(p_me^{t_m})^{x_m} \\ &= (p_1e^{t_1}+\cdots+p_me^{t_m})^n, \end{aligned}$$

lo que nos da una expresión para $M_{\boldsymbol{X}}(\boldsymbol{t})$. Observe que mediante el expediente de tomar $t_j = 0$ para $j \neq k$, se obtiene

$$M_{X_k}(t_k) = (1 - p_k + p_ke^{t_k})^n, \qquad t_k \in \mathbb{R},$$

de modo que $X_k \sim \text{Bin}(n, p_k)$ para cualquier $k = 1, \ldots, m$. Calculemos ahora $\rho(X_1, X_2)$. Se tiene que, por las propiedades de la distribución binomial, $E(X_k) = np_k$ y $Var(X_k) = np_k(1 - p_k)$, Por otra parte.

$$M_{(X_1,X_2)}(t_1, t_2) = (p_1e^{t_1} + p_2e^{t_2} + 1 - p_1 - p_2)^n,$$

de modo que

$$\begin{aligned} E(X_1X_2) &= \frac{\partial^2}{\partial t_1 \partial t_2} M_{(X_1,X_2)}\,\big|_{(0,0)} \\ &= n(n-1)p_1p_2, \end{aligned}$$

de modo que $Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = -np_1p_2$, y finalmente,

$$\rho(X_1, X_2) = \sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}}.$$

En forma análoga se obtienen la correlación para otro par dado de componentes de $\boldsymbol{X}$.

Ejemplo 4.6.6 Sea $\boldsymbol{X} \in \mathbb{R}^n$ un vector aleatorio, $\boldsymbol{A}$ una matriz de $n \times n$, y defina $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, en donde $\boldsymbol{X}$ e $\boldsymbol{Y}$ se interpretan aquí como vectores columna, o matrices de $n \times 1$. Entonces

$$\begin{aligned} \varphi_{\boldsymbol{Y}}(\boldsymbol{t}) &= E(e^{i\boldsymbol{t}'\boldsymbol{Y}}) = E(e^{i\boldsymbol{t}'\boldsymbol{A}\boldsymbol{X}}) \\ &= \varphi_{\boldsymbol{X}}(\boldsymbol{A}'\boldsymbol{t}). \end{aligned} \tag{4.6.6}$$

A modo de aplicación, considere el caso $n = 2$, donde las componentes de $\boldsymbol{X}$ son variables aleatorias $X_1, X_2 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, y sea

$$\boldsymbol{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Se tiene entonces que

$$\boldsymbol{A}'\boldsymbol{t} = \frac{1}{\sqrt{2}} \begin{pmatrix} t_1 + t_2 \\ t_1 - t_2 \end{pmatrix}.$$

Puesto que X_1 y X_2 son independientes, se tiene

$$\varphi_{(X_1,X_2)}(t_1, t_2) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2) = e^{-t_1^2/2} e^{-t_2^2/2},$$

de modo que la función característica conjunta del vector $\boldsymbol{Y}$ es

$$\begin{aligned} \varphi_{(Y_1,Y_2)}(t_1, t_2) &= \varphi_{(X_1,X_2)}(\boldsymbol{A}'\boldsymbol{t}) \\ &= \exp\left\{ -\frac{1}{2}\left(\frac{t_1 + t_2}{\sqrt{2}} \right)^2 - \frac{1}{2}\left(\frac{t_1 - t_2}{\sqrt{2}} \right)^2 \right\} \\ &= e^{-t_1^2/2} e^{-t_2^2/2}, \end{aligned}$$

de donde se concluye que (Y_1, Y_2) tiene componentes i.i.d., cada una con distribución $N(0, 1)$. En otras palabras, hemos mostrado que

$$\frac{X_1 + X_2}{\sqrt{2}} \qquad \text{y} \qquad \frac{X_1 - X_2}{\sqrt{2}}$$

son i.i.d. con distribución $N(0, 1)$.

4.7. La Distribución Normal Multivariada

Estudiaremos a continuación una distribución que corresponde a la extensión a varias dimensiones de la densidad definida en (3.9.2). Primero daremos una definición general, que es conveniente para ciertos aspectos de manejo teórico, y posteriormente daremos una versión un tanto más restringida, pero de mayor utilidad práctica. Es además conveniente utilizar la convención que cualquier vector en $\mathbb{R}^n$ se entiende como un vector columna, o equivalentemente, como una matriz con n filas y 1 columna. Por razones también teóricas, es conveniente introducir el concepto de *distribución normal degenerada*. En la fórmula (3.9.2) se requiere que la varianza σ^2 sea positiva, pues en caso contrario dicha densidad no está definida. Permitiremos que σ^2 tome el valor 0, caso en el cual se dice que la distribución normal es *degenerada*, lo que corresponde a decir que $X \sim N(\mu, 0)$ si X es constante e igual a μ. Ciertamente, esto corresponde a una variable aleatoria discreta, y no existe densidad.

Definición 4.7.1 Diremos que el vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ tiene distribución *normal multivariada*, si para cualquier $\boldsymbol{A} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ no nulo se tiene $\boldsymbol{A}'\boldsymbol{X} = \sum_{i=1}^{n} a_i X_i$ tiene distribución normal univariada.

Notemos que esta definición no hace referencia a densidad alguna. Sin embargo, si $\boldsymbol{e}_i$ es el i-ésimo vector de la base canónica de $\mathbb{R}^n$, se tiene $\boldsymbol{e}_i'\boldsymbol{X} = X_i$, y se concluye que si $\boldsymbol{X}$ tiene distribución normal multivariada, entonces cada una de sus coordenadas tiene distribución normal

univariada. Por lo tanto, y puesto que $E(X_i^2)$ es finito para cada $i = 1, \ldots, n$, también existe la matriz de varianza-covarianza (ver Definición 4.5.2).

Sea ahora $\boldsymbol{t} \in \mathbb{R}^n$. Puesto que $\boldsymbol{t}'\boldsymbol{X}$ tiene distribución normal univariada, se concluye que

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \varphi_{\boldsymbol{t}'\boldsymbol{X}}(1) = e^{i\mu(\boldsymbol{t}) - \sigma^2(\boldsymbol{t})/2},$$

donde $\mu(\boldsymbol{t}) = E(\boldsymbol{t}'\boldsymbol{X})$ y $\sigma^2(\boldsymbol{t}) = Var(\boldsymbol{t}'\boldsymbol{X})$. Denotando $\boldsymbol{\mu} = E(\boldsymbol{X})$ y $\boldsymbol{\Sigma} = V(\boldsymbol{X})$, tenemos que por (4.5.11), $\mu(\boldsymbol{t}) = \boldsymbol{t}'\boldsymbol{\mu}$ y que $\sigma^2(\boldsymbol{t}) = \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}$, de modo que la función característica multivariada de $\boldsymbol{X}$ es

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}'\boldsymbol{\mu} - \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2}, \qquad \boldsymbol{t} \in \mathbb{R}^n. \tag{4.7.7}$$

Puesto que la función característica de $\boldsymbol{X}$ determina su distribución, vemos que basta con conocer el vector de medias, y la matriz de varianza-covarianza de $\boldsymbol{X}$ para conocer su distribución. La notación usual para una vector aleatorio n-dimensional $\boldsymbol{X}$ con distribución normal multivariada y tal que $E(\boldsymbol{X}) = \boldsymbol{\mu}$ y $V(\boldsymbol{X}) = \boldsymbol{\Sigma}$ es $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. En el caso univariado $n = 1$, el subíndice n suele omitirse.

Ejemplo 4.7.1 Consideremos variables aleatorias independientes $Y_i \sim N(\mu_i, \sigma_i^2)$, con $i = 1, 2$, y defina $\boldsymbol{X} = (Y_1, Y_2)$, visto como un vector columna. Sea $\boldsymbol{A} = (a_1, a_2) \neq (0, 0)$ un vector en $\mathbb{R}^2$. Usando funciones generadoras, tal como en el Ejemplo 4.6.2, es fácil ver que $\boldsymbol{A}'\boldsymbol{X} = a_1X_1 + a_2X_2 \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$, y por la Definición 4.7.1 se concluye que $\boldsymbol{X}$ tiene distribución normal multivariada. El vector de medias, y la matriz de varianza-covarianza correspondientes están respectivamente dados por

$$E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \text{y} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Observe que $\boldsymbol{\Sigma}$ es una matriz invertible, provisto que $\sigma_i^2 > 0$ para $i = 1, 2$. Defina ahora $\boldsymbol{Y} = (Y_1, -2Y_1)$. Observe que $\boldsymbol{A}'\boldsymbol{Y} = (a_1 - 2a_2)Y_1$, el cual tiene distribución normal univariada cualquiera que sean a_1 y a_2, incluso en el caso en que $a_1 = 2a_2$ en el que $\boldsymbol{A}'\boldsymbol{Y} = 0$, lo que corresponde a la distribución degenerada $N(0, 0)$. Note además que la matriz de covarianza es ahora

$$\begin{pmatrix} \sigma_1^2 & -2\sigma_1^2 \\ -2\sigma_1^2 & 4\sigma_1^2 \end{pmatrix}.$$

Es fácil ver que cualquiera que sea σ_1^2, esta matriz es no invertible.

El Ejemplo 4.7.1 motiva establecer una distinción entre vectores aleatorios con distribución normal multivariada. En el caso que la matriz de covarianza $\boldsymbol{\Sigma}$ de $\boldsymbol{X}$ sea no invertible, diremos que $\boldsymbol{X}$ tiene distribución normal multivariada *degenerada*, y esto corresponde a la extensión a varias dimensiones del concepto anteriormente introducido para variables con distribución normal univariada. Intuitivamente, esto corresponde al caso en que alguna de las componentes de $\boldsymbol{X}$ se puede escribir como una combinación lineal de las otras. En otras palabras, cuando el vector aleatorio $\boldsymbol{X}$ toma valores en un conjunto cuya dimensión es inferior a la dimensión de $\boldsymbol{X}$, tal como aconteció en el Ejemplo 4.7.1.

En el caso en que $\mathbf{\Sigma}$ es invertible, definimos la siguiente forma cuadrática:

$$Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \qquad \boldsymbol{x} \in \mathbb{R}^n \tag{4.7.8}$$

en donde $\boldsymbol{\mu} \in \mathbb{R}^n$ es un vector cualquiera. Observe que los valores de Q son siempre números reales, y el hecho que $\mathbf{\Sigma}$ sea invertible garantiza que ésta es además definida positiva, por lo que se concluye que $Q(\boldsymbol{x}) \geq 0$ para cualquier $\boldsymbol{x}$, y con igualdad sólo si $\boldsymbol{x} = \boldsymbol{\mu}$.

El siguiente resultado nos da una expresión para la densidad de $\boldsymbol{X}$ cuando $\mathbf{\Sigma}$ es invertible.

Proposición 4.7.1 Sea $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$, donde $\mathbf{\Sigma}$ es una matriz invertible. Entonces, $\boldsymbol{X}$ tiene densidad conjunta dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{e^{-\frac{1}{2}Q(\boldsymbol{x})}}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\mathbf{\Sigma})}} \tag{4.7.9}$$

y en donde $Q(\boldsymbol{x})$ fue definido en (4.7.8).

El lector podrá fácilmente convencerse que para el caso $n = 1$, (4.7.9) se reduce a (3.9.2). Consideremos ahora el caso particular en que $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I}_n$, es decir, cuando la matriz de varianza-covarianza adopta la forma especial de una matriz diagonal, donde cada elemento no nulo es igual a σ^2. Es claro que las componentes de $\boldsymbol{X} = (X_1, \ldots, X_n)$ son no correlacionadas, pues para $i \neq j$, $Cov(X_i, X_j) = \mathbf{\Sigma}_{i,j} = 0$, y además, $Var(X_i) = \sigma^2$ para $i = 1, 2, \ldots, n$. Por otra parte, observe que la forma cuadrática (4.7.8) adopta la forma especial de

$$Q(\boldsymbol{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu_i)^2,$$

de modo que la densidad conjunta de $\boldsymbol{X}$ está dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n}\frac{e^{-\frac{(x_i-\mu_i)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}},$$

y se concluye que $X_1, \ldots, X_n$ son **independientes**. Este resultado se puede generalizar en forma directa, para obtener:

Proposición 4.7.2 Si $\boldsymbol{X} = (X_1, \ldots, X_n) \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$, en donde $\mathbf{\Sigma}$ es una matriz diagonal, entonces $X_1, \ldots, X_n$ son independientes.

Esto muestra una característica muy particular de la distribución normal multivariada, cual es que la no correlación equivale a la independencia. Vale la pena recordar que esto es, en general, falso, como se mostró anteriormente.

Otras propiedades de la distribución normal multivariada se resumen a continuación. Note que algunas de estas propiedades son válidas sólo para el caso en que $\mathbf{\Sigma}$ es invertible, pero algunas otras valen en general.

Proposición 4.7.3 Sea $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$, donde $\boldsymbol{\mu} \in \mathbb{R}^n$ y $\mathbf{\Sigma}$ es una matriz simétrica de $n \times n$.

(i) Sea $\boldsymbol{A}$ una matriz de $k \times n$. Si $\boldsymbol{Y} = \boldsymbol{AX}$ entonces $\boldsymbol{Y} \sim N_k(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A'})$.

(ii) Suponga $\boldsymbol{\Sigma}$ invertible, y considere su *descomposición de Cholesky* $\boldsymbol{\Sigma} = \boldsymbol{RR'}$, en donde $\boldsymbol{R}$ es una matriz triangular inferior. Entonces $\boldsymbol{Y} \stackrel{\text{def}}{=} \boldsymbol{R}^{-1}\boldsymbol{X} \sim N_n(\boldsymbol{R}^{-1}\boldsymbol{\mu}, \boldsymbol{I}_n)$.

(iii) Si $\boldsymbol{\Sigma}$ es invertible, entonces $(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2(n)$, la distribución chi-cuadrado con n grados de libertad.

Demostración: Haremos las demostraciones de estas propiedades, pues el procedimiento utilizado es de interés por sí mismo. Para mostrar (i), consideremos la función característica de $\boldsymbol{Y}$. Por (4.6.6) tenemos que $\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = \varphi_{\boldsymbol{X}}(\boldsymbol{A't})$, de modo que

$$\begin{aligned} \varphi_{\boldsymbol{Y}}(\boldsymbol{t}) &= e^{i(\boldsymbol{A't})'\boldsymbol{\mu} - (\boldsymbol{A't})'\boldsymbol{\Sigma}(\boldsymbol{A't})/2} \\ &= e^{i\boldsymbol{t'A\mu} - \boldsymbol{t'A\Sigma A't}/2}, \end{aligned}$$

y el resultado es inmediato. La propiedad (ii) es directa de (i), tomando $\boldsymbol{A} = \boldsymbol{R}^{-1}$. Para ver (iii), considere $\boldsymbol{Y} = \boldsymbol{R}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, donde $\boldsymbol{R}$ es la matriz triangular inferior mencionada en (ii). Por (ii), se tiene que $\boldsymbol{Y} \sim N_n(0, \boldsymbol{I}_n)$, y además

$$(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Y'Y} = \sum_{j=1}^{n} Y_j,$$

donde $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(0,1)$. El resultado es entonces una consecuencia de lo hecho en el Ejemplo 4.6.3. ■

Veamos a continuación algunas aplicaciones de estos resultados.

Ejemplo 4.7.2 Considere (X, Y) con densidad conjunta proporcional a $e^{-Q(x,y)/2}$, donde

$$Q(x, y) = x^2 + 2y^2 - 8x + 10y - 2xy + 17.$$

Por la forma que tiene la densidad, se deduce que (X, Y) tiene distribución normal bivariada, pero es necesario identificar sus parámetros. Consideremos la forma cuadrática (4.7.8) correspondiente con $n = 2$,

$$Q(x, y) = (x - \mu_1, y - \mu_2)\boldsymbol{\Sigma}^{-1} \begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix},$$

la cual se quiere igualar con la expresión dada inicialmente. Para ello, igualaremos sus derivadas, lo que da, escrito en forma vectorial:

$$\begin{pmatrix} 2x - 8 - 2y \\ 4y + 10 - 2x \end{pmatrix} = 2\boldsymbol{\Sigma}^{-1} \begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix}.$$

Note que igualando el lado derecho a $(0, 0)'$, se obtiene, después de multiplicar a la izquierda por $\boldsymbol{\Sigma}$ que $x - \mu_1 = 0$ e $y - \mu_2 = 0$, por lo que se concluye que μ_1 y μ_2 se obtienen de resolver el sistema $\nabla Q(x, y) = (0, 0)'$. En nuestro caso:

$$\begin{aligned} 2x - 2y &= 8 \\ 4y - 2x &= -10, \end{aligned}$$

cuya solución es $(\mu_1, \mu_2) = (3, -1)$. Para obtener $\mathbf{\Sigma}$, observe que al igualar las matrices Hessianas se tiene

$$\begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix} = 2\mathbf{\Sigma}^{-1},$$

de modo que $\mathbf{\Sigma} = 2(HQ(x,y))^{-1}$. En nuestro caso:

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix},$$

lo que termina de identificar los parámetros de la distribución normal bivariada buscada.

El método aquí empleado se puede extender fácilmente a más dimensiones.

Ejemplo 4.7.3 Sean $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ las coordenadas del vector $\boldsymbol{X}$, y defina $U = \frac{1}{n}\sum_{j=1}^{n} X_j$ y $V = \sum_{j=1}^{n}(X_j - U)^2$. Veamos que U y V son independientes. Considere el vector

$$\boldsymbol{Y} = \begin{pmatrix} U \\ X_1 - U \\ \vdots \\ X_n - U \end{pmatrix},$$

el cual puede interpretarse como una transformación lineal del vector $\boldsymbol{X}$ de la forma $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, donde $\boldsymbol{A}$ es una matriz de $(n+1) \times n$. Puesto que $\boldsymbol{X}$ tiene distribución normal multivariada, $\boldsymbol{Y}$ también. Además

$$\begin{aligned} Cov(U, X_j - U) &= Cov(U, X_j) - Cov(U, U) \\ &= \frac{1}{n}\sum_{k=1}^{n} Cov(X_k, X_j) - Var(U) \\ &= \frac{1}{n} Var(X_j) - \frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0, \end{aligned}$$

de modo que U es independiente de $X_1 - U, \ldots, X_n - U$, de donde se sigue que U y V son independientes. Note que este resultado no depende de los valores particulares que μ y σ^2 puedan tomar.

Ejemplo 4.7.4 En el Ejemplo 4.7.3, veamos ahora que $V/\sigma^2 \sim \chi^2(n-1)$. Para ello, considere primero el caso en que $\mu = 0$, $\sigma^2 = 1$, y defina las variables aleatorias

$$\begin{aligned} Y_1 &= \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \\ Y_j &= \frac{X_1 + X_2 + \cdots + X_{j-1} - (j-1)X_j}{\sqrt{j(j-1)}}, \qquad j = 2, \ldots, n. \end{aligned}$$

Sea

$$\boldsymbol{Q} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \cdots & \cdots & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \cdots & \cdots & \cdots & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix},$$

de modo que si $\boldsymbol{Y}$ es el vector cuyas coordenadas son $Y_1, \ldots, Y_n$, las transformaciones descritas (conocidas como transformaciones de Helmert) se escriben $\boldsymbol{Y} = \boldsymbol{QX}$, donde $\boldsymbol{Q}$ es una matriz de $n \times n$. Observe que $\boldsymbol{Q}$ es una matriz *unitaria*, es decir, $\boldsymbol{QQ}' = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_n$, de modo que $\boldsymbol{Y} \sim N_n(0, \boldsymbol{I}_n)$. Entonces:

$$\begin{aligned} \sum_{j=1}^{n} Y_j^2 &= \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{X} \\ &= \boldsymbol{X}'\boldsymbol{X} = \sum_{j=1}^{n} X_j^2 \\ &= n\overline{X}_n^2 + \sum_{j=1}^{n} (X_j - \overline{X}_n)^2 \\ &= Y_1^2 + V, \end{aligned}$$

de donde $V = Y_2^2 + Y_3^2 + \cdots + Y_n^2$, y vemos así que V se escribe como la suma de $n-1$ variables aleatorias i.i.d. con distribución $N(0,1)$, de donde se sigue que $V \sim \chi^2(n-1)$. Se sigue además que U y V son independientes. En el caso general, considere las mismas variables $Y_1, \ldots, Y_n$, definidas ahora en términos de $Z_j \stackrel{\text{def}}{=} (X_j - \mu)/\sigma \sim N(0,1)$, para $j = 1, \ldots, n$. Por último, observe que $E(V/\sigma^2) = n-1$, por lo que $E(V/(n-1)) = \sigma^2$, resultado independiente del valor de μ.

4.8. El Mejor Predictor Lineal

Para finalizar este capítulo resolveremos el siguiente problema. Suponga $\boldsymbol{X} \in \mathbb{R}^{k+l}$ es un vector aleatorio con $E(\boldsymbol{X}) = \boldsymbol{\mu}$ y $V(\boldsymbol{X}) = \boldsymbol{\Sigma}$, lo cual anotaremos $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suponga además que $\boldsymbol{X}$ se puede particionar de la siguiente forma:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Y} \end{pmatrix} \qquad \text{con} \qquad \boldsymbol{W} = \begin{pmatrix} W_1 \\ \vdots \\ W_k \end{pmatrix}, \qquad \text{e} \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_l \end{pmatrix}.$$

Si el valor de $\boldsymbol{W}$ es conocido, digamos $\boldsymbol{w}$, ¿cómo predecir el valor de $\boldsymbol{Y}$? Esta situación se suscita en casos donde las variables de interés se observan sólo en parte, de modo que se requiere "adivinar" el valor de las variables no observadas, pero asumiendo $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ conocidos.

El problema así planteado es un tanto vago. Para hacerlo más preciso, nos centraremos aquí en *predictores lineales*, esto es, predictores de la forma $\boldsymbol{a}+\boldsymbol{BW}$, donde $\boldsymbol{a}\in\mathbb{R}^l$, y $\boldsymbol{B}$ es una matriz de $l\times k$. Resta aún por definir un procedimiento para obtener $\boldsymbol{a}$ y $\boldsymbol{B}$. Para ello, recurrimos al criterio de minimizar el *error cuadrático medio*, es decir, resolveremos el problema de calcular $\boldsymbol{a}$ y $\boldsymbol{B}$ tales que

$$E\{(\boldsymbol{Y}-\boldsymbol{a}-\boldsymbol{BW})'(\boldsymbol{Y}-\boldsymbol{a}-\boldsymbol{BW})\} \tag{4.8.1}$$

sea mínimo.

Introducimos ahora la siguiente notación. Sean $\boldsymbol{\mu}_w=E(\boldsymbol{W})$ y $\boldsymbol{\mu}_y=E(\boldsymbol{Y})$ los vectores de valores esperados de $\boldsymbol{W}$ e $\boldsymbol{Y}$ respectivamente. Las matrices de varianza-covarianza correspondientes se denotarán por $\boldsymbol{\Sigma}_{ww}$ y $\boldsymbol{\Sigma}_{yy}$, y finalmente, la matriz de covarianzas entre $\boldsymbol{W}$ e $\boldsymbol{Y}$ se denotará por $\boldsymbol{\Sigma}_{wy}$, de modo que

$$\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_w\\ \boldsymbol{\mu}_y\end{pmatrix} \qquad \text{y} \qquad \boldsymbol{\Sigma}=\left(\begin{array}{c|c}\boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wy}\\ \hline \boldsymbol{\Sigma}_{yw} & \boldsymbol{\Sigma}_{yy}\end{array}\right),$$

donde $\boldsymbol{\Sigma}$ es una *matriz particionada*. Observe que puesto que $\boldsymbol{\Sigma}$ debe ser simétrica, se debe cumplir que $\boldsymbol{\Sigma}'_{yw}=\boldsymbol{\Sigma}_{wy}$. Con esta notación, se tiene el siguiente resultado.

Proposición 4.8.1 La solución al problema de minimización (4.8.1) está dada por

$$\boldsymbol{a}=\boldsymbol{\mu}_y-\boldsymbol{B}\boldsymbol{\mu}_w, \tag{4.8.2}$$

y

$$\boldsymbol{B}=\boldsymbol{\Sigma}_{yw}\boldsymbol{\Sigma}_{ww}^{-1}. \tag{4.8.3}$$

Demostración: Observe que la expresión en (4.8.1) se puede reescribir como

$$E(\boldsymbol{Y}'\boldsymbol{Y})-2\boldsymbol{a}'E(\boldsymbol{Y})+\boldsymbol{a}'\boldsymbol{a}-2E(\boldsymbol{Y}'\boldsymbol{BW})+2\boldsymbol{a}'\boldsymbol{B}E(\boldsymbol{W})+E(\boldsymbol{W}'\boldsymbol{B}'\boldsymbol{BW}),$$

la que a su vez es igual a

$$\begin{aligned}&\sum_{i=1}^{l}E(Y_i^2)-2\sum_{i=1}^{l}a_iE(Y_i)+\sum_{i=1}^{l}a_i^2-2\sum_{i=1}^{l}\sum_{j=1}^{k}B_{ij}E(Y_iW_j)\\&+2\sum_{i=1}^{l}\sum_{j=1}^{k}a_iB_{ij}E(W_j)+\sum_{i=1}^{l}E\{(\sum_{j=1}^{k}B_{ij}W_j)^2\}.\end{aligned}$$

Para minimizar, primero diferenciamos esta expresión con respecto a a_i e igualamos a 0, con lo que se obtiene

$$-2E(Y_i)+2a_i+2\sum_{j=1}^{l}B_{ij}E(W_j)=0,$$

o equivalentemente,

$$a_i=E(Y_i)-\sum_{j=1}^{l}B_{ij}E(W_j),$$

lo que escrito en forma vectorial resulta $\boldsymbol{a} = \boldsymbol{\mu}_y - \boldsymbol{B}\boldsymbol{\mu}_w$, lo que prueba (4.8.2). Para obtener $\boldsymbol{B}$, usamos un procedimiento análogo. Se deriva con respecto a B_{ij}, se iguala a 0, para obtener, después de acomodar términos y reemplazar el valor de a_i por $E(Y_i) - \sum_{m=1}^{k} B_{im}E(W_m)$ el conjunto de $k \times l$ ecuaciones

$$\sum_{m=1}^{k} B_{im}Cov(W_m, W_j) = Cov(Y_i, W_j) \qquad \forall\, i, j,$$

de donde se obtiene el resultado $\boldsymbol{B} = \boldsymbol{\Sigma}_{yw}\boldsymbol{\Sigma}_{ww}^{-1}$. ■

Así, el mejor predictor lineal (MPL), en el sentido explicado anteriormente, de $\boldsymbol{Y}$ dado un valor para $\boldsymbol{W}$ es

$$\boldsymbol{a} + \boldsymbol{B}\boldsymbol{W} = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yw}\boldsymbol{\Sigma}_{ww}^{-1}(\boldsymbol{W} - E(\boldsymbol{W})). \tag{4.8.4}$$

Se propone como ejercicio mostrar que la matriz de varianza-covarianza del MPL está dada por

$$\boldsymbol{\Sigma}_{yw}\boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{wy}. \tag{4.8.5}$$

Ejemplo 4.8.1 Sea $\boldsymbol{X} \in \mathbb{R}^3$ con

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad \text{y} \quad \boldsymbol{\Sigma} = \left(\begin{array}{cc|c} 5 & 1 & 2 \\ 1 & 3 & 3 \\ \hline 2 & 3 & 6 \end{array}\right),$$

y obtengamos el MPL de X_3 dados X_1 y X_2. Tenemos que, por (4.8.3):

$$\boldsymbol{B} = \begin{pmatrix} 3 & 5 \end{pmatrix} \begin{pmatrix} 5 & 1 \\ 1 & 3 \end{pmatrix}^{-1} = \frac{1}{14} \begin{pmatrix} 3 & 13 \end{pmatrix}.$$

Por otra parte, por (4.8.2):

$$\boldsymbol{a} = 0 - \frac{1}{14} \begin{pmatrix} 3 & 13 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = -\frac{16}{14}.$$

Así, el MPL buscado es, de acuerdo a (4.8.4):

$$\frac{-16 + 3X_1 + 13X_2}{14}.$$

La varianza del MPL se obtiene de (4.8.5), y está dada por

$$Var(MPL) = \frac{45}{14}.$$

Observe que en este caso el MPL es simplemente escalar.

Ejemplo 4.8.2 Sea $\boldsymbol{X} \in \mathbb{R}^5$ con

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 0 \\ -1 \\ 1 \\ -3 \end{pmatrix} \quad \text{y} \quad \boldsymbol{\Sigma} = \left(\begin{array}{cc|ccc} 12 & -1 & 3 & 6 & 0 \\ -1 & 36 & 5 & 5 & 0 \\ \hline 3 & 5 & 9 & -1 & 0 \\ 6 & 5 & -1 & 13 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{array}\right).$$

Calculemos el MPL de $\boldsymbol{Y}$ dado $\boldsymbol{W}$, donde

$$\boldsymbol{W} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad \text{e} \qquad \boldsymbol{Y} = \begin{pmatrix} X_3 \\ X_4 \\ X_5 \end{pmatrix}.$$

En este caso tenemos, aplicando (4.8.2) y (4.8.3) que

$$\boldsymbol{B} = \begin{pmatrix} 3 & 5 \\ 6 & 5 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 12 & -1 \\ -1 & 36 \end{pmatrix}^{-1} = \frac{1}{431} \begin{pmatrix} 113 & 63 \\ 221 & 66 \\ 0 & 0 \end{pmatrix},$$

y

$$\boldsymbol{a} = \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} - \frac{1}{431} \begin{pmatrix} 113 & 63 \\ 221 & 66 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \frac{1}{431} \begin{pmatrix} -657 \\ -11 \\ 1293 \end{pmatrix}.$$

Finalmente, el MPL buscado es

$$\frac{1}{431} \begin{pmatrix} -657 + 113X_1 + 63X_2 \\ -11 + 221X_1 + 66X_2 \\ 1293 \end{pmatrix}.$$

La matriz de varianza-covarianza del MPL es, de acuerdo a (4.8.5)

$$\frac{1}{431} \begin{pmatrix} 654 & 993 & 0 \\ 993 & 1656 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Observe la forma especial del MPL, particularmente su tercera coordenada, y la última fila y columna de $V(MPL)$. Esto no es coincidencia, y la razón es que X_5 es no correlacionada con las otras variables predictoras. Así, al no existir correlación, el MPL se transforma simplemente en $E(X_5) = 1293/431 = 3$, tal como se obtuvo.

4.9. Problemas

1. Sean $X_1, X_2, \ldots, X_n$ i.i.d. con función de probabilidad $p(x) = (1-p)p^x,\ x = 0, 1, \ldots$, es decir, con una distribución de tipo geométrico. Sea $Y_n = X_1 + X_2 + \cdots + X_n$.

   (a) Encuentre la función probabilidad de Y_2.

   (b) Demuestre que la función probabilidad p_n de Y_n está dada por

   $$p_n(y) = \frac{(y+n-1)!}{y!(n-1)!}\theta^n(1-\theta)^y,\ y = 0, 1, 2, \ldots$$

   Indicación: Demuestre que si Z tiene función probabilidad p_m, U tiene función probabilidad p_1 y Z y U son independientes, entonces $Z + U$ tiene función probabilidad p_{m+1}. Proceda luego por inducción.

   (c) Calcule la media de Y_n en base a la expresión obtenida para p_n.

   (d) Calcule el valor esperado de Y_n como la suma de los valores esperados de los X_i.

2. Un lote de tamaño N tiene D elementos defectuosos. Se extrae una muestra aleatoria de tamaño n y se cuenta el número X de elementos defectuosos en la muestra.

   (a) Calcule $E(X)$ a partir de la función probabilidad.

   (b) Exprese X como $X_1 + \cdots + X_n$ y use $E(X) = E(X_1) + \cdots + E(X_n)$. Use ésto para calcular $E(X)$.

3. Demostrar la desigualdad de Cauchy-Schwartz :

   $$(E(XY))^2 \leq E(X^2)E(Y^2).$$

   Hint: Considere $E((tX + Y)^2)$.

4. Sean X e Y variables aleatorias independientes con distribución uniforme en $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, $\theta \in \mathbb{R}$. Pruebe que la distribución de $X - Y$ no depende de θ, hallando su densidad.

5. Dado $f_{X_1,X_2}(x_1, x_2)$, encontrar $f_{U,V}(u, v)$ y $f_U(u)$, con:

   (a) $U = X_1 + X_2$, $V = X_2$

   (b) $U = X_1X_2$, $V = X_2$

   (c) $U = \frac{X_1}{X_2}$, $V = X_2$

   Explicitar el caso particular en el que X_1 y X_2 son independientes.

6. En el Problema 5, encuentre $f_U(u)$ e identifique, de ser posible, la distribución para X_1, X_2 iid$\sim N(0, 1)$.

7. Sean $X_1, X_2, \ldots, X_n$ iid con distribución de Rayleigh con parámetro $\theta > 0$:

   $$f(x) = \begin{cases} \frac{x}{\theta}\exp(\frac{-x^2}{2\theta^2}) & \text{si } x > 0 \\ 0 & \text{si } x \leq 0. \end{cases}$$

a.- Determine la densidad conjunta de $Y_1, Y_2, \ldots, Y_n$ donde $Y_i = X_i^2$.

b.- ¿Cuál es la distribución de $U = \min\{X_1, \ldots, X_n\}$?

c.- Calcule la distribución de $Z = \frac{X_1}{X_2}$.

8. Sean X e Y iid Exp(α). Muestre que $Z = \frac{X}{X+Y} \sim U(0,1)$.

9. Sean X_1, X_2 con densidad conjunta

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} \frac{1}{x_1^2} e^{-(x_2 - x_1^2)} & \text{si } x_1 \geq 1,\ x_2 \geq x_1^2 \\ 0 & \text{si no.} \end{cases}$$

Encontrar $f_{Y_1}(y_1)$, con $Y_1 = \log(X_1)$ y $f_{Y_2}(y_2)$, con $Y_2 = \frac{X_2}{X_1}$.

10. Sean X, Y y Z son variables aleatorias independientes que tienen igual función densidad $f(x) = e^{-x}$, $0 < x < \infty$. Encuentre la distribución conjunta de $U = X + Y$, $V = X + Z$, $W = Y + Z$.

11. Suponga que X_1, X_2 son variables aleatorias independientes con distribución uniforme sobre el intervalo $[0,1]$. Encuentre la distribución conjunta de $Y_1 = X_1 + X_2$ e $Y_2 = \frac{X_1}{X_2}$.

12. Cuando una corriente I (medida en ampéres) fluye a través de una resistencia R (medida en ohms), la potencia generada está dada por $W = I^2 R$ (medida en Watts). Si I y R son variables aleatorias independientes con densidades

$$f_I = 6x(1-x) \ \ 0 \leq x \leq 1$$

$$f_R(x) = 2x \ \ 0 \leq x \leq 1,$$

Determine f_W.

13. Sean $X_1, \ldots, X_n$ variables aleatorias i.i.d. con densidad

$$f_X(x) = x^{-2} \ \ si \ \ 1 < x < \infty.$$

Sea $Y = \min\{X_1, \ldots, X_n\}$. ¿Existe $E(X_1)$? Si es así encuéntrela. ¿Existe $E(Y)$? Si es así encuéntrela.

14. Sean X_1, X_2 variables aleatorias independientes cada una con distribución $N(0,1)$. Si $Y_1 = X_1^2 + X_2$, $Y_2 = X_2$, encuentre $f_{Y_1,Y2}$ y f_{Y_1}.

15. Suponga que los tiempos entre ocurrencias de un cierto fenómeno pueden ser representados por $T_1, \ldots, T_n$, variables aleatorias independientes cada una con distribución exponencial de parámetro λ. Si $T = T_1 + \cdots + T_n$, encuentre la distribución de T.

16. Si X e Y son las coordenadas de un punto seleccionado al azar del círculo unitario $\{(x,y) : x^2 + y^2 \leq 1\}$, ¿cuál es la distribución de la variable aleatoria $Z = X^2 + Y^2$?.

17. Si $T_1 \sim$ Exp(λ_1) y $T_2 \sim$ Exp(λ_2), encuentre la densidad de $T = T_1 + T_2$.

18. Dados $a < b$ y $c < d$, $X \sim U[a,b]$ e $Y \sim U[a,b]$, con X e Y independientes, calcule $f_X \star f_Y$.

19. Suponga $X_1, \ldots, X_n$ son variables aleatorias i.i.d. con distribución $U[0,1]$. Pruebe que

$$-2n\log(Y) \sim \text{Gama}(n, \frac{1}{2}),$$

donde Y es la media geométrica de las X_i, esto es,

$$Y = (\prod_{i=1}^{n} X_i)^{1/n}.$$

20. La densidad conjunta entre X e Y está dada por:

$$f_{X,Y}(x,y) = \frac{e^{-\frac{x}{y}}e^{-y}}{y} \qquad 0 \leq x \leq \infty \qquad 0 \leq y \leq \infty.$$

Encuentre $E(X)$.

21. Sean $X_1, X_2, \ldots, X_n$ iid U(0,1).

a.- Sean $Y_j = -\log(X_j) \;\; j = 1, \ldots, n$. Encontrar la función generadora de momentos de Y_j, y a partir de ella calcule $E(Y)$ y $Var(Y)$. ¿Qué distribución tiene Y?.

b.- Sea $Y = \lambda \sum_{j=1}^{n} Y_j$ con $\lambda > 0$. Encuentre la función generadora de momentos de Y. Calcule $E(Y)$, $Var(Y)$. ¿Qué distribución tiene Y?.

22. Si la variable aleatoria X tiene función generadora de momentos dada por $M_X(t) = \frac{3}{3-t}$, obtener la desviación estándar de X.

$Resp: \frac{1}{3}$

23. En un circuito se ponen n resistencias en serie. Supóngase que cada una de las resistencias está distribuida uniformemente en (0,1), y suponga además que todas las resistencias son independientes. Sea R la resistencia total.

(a) Encontrar la función generadora de momentos de R.

(b) Usando (a), encontrar $E(R)$ y $Var(R)$.

24. Suponga que la distribución conjunta de X_1 y X_2 es normal bivariada. Se definen las variables aleatorias $Y_1 = 3X_1 + 2X_2 + 1$ e $Y_2 = X_1 + 5X_2 - 4$. Demuestre que (Y_1, Y_2) tiene también distribución normal bivariada, e identifique sus parámetros.

25. Si $(X_1, X_2)^t \sim N_2(\mu, \Sigma)$ donde μ^t = (1,-2) y $\sigma_1^2 = 4, \sigma_{12} = -10, \sigma_2^2 = 25$, encuentre directamente las densidades marginales de X_1 y X_2.

26. Sean $Y_1, Y_2, \ldots, Y_n$ definidos por

$$Y_i = U + Y_{i-1} + Z_{i-1} \qquad i = 1, \ldots, n \qquad Z_0 = 0, \qquad Y_0 = 0,$$

en donde $U, Z_1, \ldots, Z_n$ son independientes de media cero, con Var$(U) = a$, Var$(Z_i) = b$.

a.- Encuentre la matriz de covarianzas de $(Y_1, \ldots, Y_n)^t$.

b.- Determine el MPL de Y_3 dado Y_2

c.- Determine el MPL de Y_4 dado $Y_2 + Y_3$

27. Sean Y_1, Y_2, Y_3 independientes de media cero y varianza uno, defina las variables aleatorias X_1, X_2, X_3 por:

$$X_1 = \frac{Y_1}{\sqrt{1-\alpha^2}} \quad , \quad X_2 = \alpha X_1 + Y_2 \quad , \quad X_3 = \alpha X_2 + Y_3$$

Encuentre Var$(X_1, X_2, X_3)^t$ y E$(X_1, X_2, X_3)^t$

28. Dados E$(X_1, X_2, X_3)^t = (1, 2, 3)^t$ y

$$\text{Var}(X) = \begin{pmatrix} a & a & a \\ a & a+b & a \\ a & a & a+c \end{pmatrix}$$

a.- Encuentre el MPL de X_1 dado $X_2 = x_2$.

b.- Encuentre el MPL de $X_3 - X_2$ dado $X_1 = 4$.

29. Sean X_1 y Y_2 variables aleatorias independientes con distribución $N(0, 1)$. Sean $Y_1 = \alpha + aX_1 + bX_2$, $Y_2 = \beta + cX_1 + dX_2$.

a.- Encuentre la distribución conjunta de Y_1 e Y_2.

b.- Calcule la varianza del error de predicción del MPL de Y_2 dado Y_1.

Capítulo 5

Distribución y Esperanza Condicional

5.1. Motivación

En el Capítulo 2 se discutió extensamente en qué sentido la información o conocimiento afecta las probabilidades de eventos. Surge entonces la noción de *probabilidad condicional*, que refleja cómo estas probabilidades cambian. Así, si A y F son eventos tales que $P(F) > 0$, en donde A representa el evento de interés, y F es la información disponible (esto es, se sabe que F ocurrió), entonces la probabilidad condicional de A dado F se define mediante

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

tal como lo expresa (2.2.1).

La inquietud natural que surge ahora se refiere a la posibilidad de implementar cálculos semejantes pero ahora referidos a variables o vectores aleatorios. En otras palabras, si X es una variable aleatoria de interés, y si se conoce el valor y que toma otra variable aleatoria Y, ¿de qué manera se afecta la distribución, y por ende, las probabilidades asociadas a X una vez conocida esta información adicional? Esto es, ¿cómo determinamos la *distribución condicional* de X dado que $Y = y$?

Hay algunos casos especiales en que esta pregunta se puede responder utilizando solamente los conceptos ya introducidos anteriormente. Comenzamos nuestra discusión abordando estos casos.

5.2. Distribución Condicional: Visión Preliminar

En el Capítulo 2 se tuvo ya un primer acercamiento al problema de determinar los cambios en la distribución de una variable aleatoria dada información relativa a una segunda variable aleatoria, en el caso que éstas son discretas. En efecto, si X e Y son discretas con función de probabilidad discreta conjunta $p_{X,Y}(x, y)$ para $(x, y) \in \mathcal{D}$, entonces la *función de probabilidad condicional* de X dado que $Y = y$ se define mediante

$$p_{X|Y=y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \qquad (x, y) \in \mathcal{D}. \tag{5.2.1}$$

En estricto rigor, esta definición se reduce simplemente a probabilidades condicionales para eventos. En efecto, si A es el evento $\{X = x\}$ y F es el evento $\{Y = y\}$, con $P(F) = P(Y = y) > 0$, entonces (5.2.1) no es otra cosa que (2.2.1). Es importante destacar que para que esta definición tenga sentido, debe cumplirse que $P(F) = P(Y = y) > 0$. En caso contrario, el cuociente (5.2.1) se indefine.

Otra característica interesante de la definición de función de probabilidad condicional, es que si X e Y son variables aleatorias independientes, entonces

$$p_{X|Y=y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x),$$

y en forma análoga, $p_{Y|X=x}(y|x) = p_Y(y)$. En otras palabras, cuando hay independencia entre las variables aleatorias en cuestión, información respecto de una de ellas no altera las probabilidades (distribución) de la otra. Esta característica es no sólo deseable, si no que, a nivel intuitivo, completamente natural.

La definición de función de probabilidad condicional se puede extender en forma natural a vectores aleatorios. Así, si $\boldsymbol{X}$ e $\boldsymbol{Y}$ son vectores aleatorios discretos, se define la *función de probabilidad discreta conjunta condicional* de $\boldsymbol{X}$ dado que $\boldsymbol{Y} = \boldsymbol{y}$ mediante

$$p_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})}.$$

Ejemplo 5.2.1 Sean $X \sim$ Poisson(λ) e $Y \sim$ Poisson(μ) independientes. Sea $Z = X + Y$, y calculemos la distribución condicional de X dado que $Z = z$. Primero, notemos que $\mathcal{Z} = \{0, 1, 2, \ldots\}$, y que para $z \in \mathcal{Z}$

$$\begin{aligned} p_Z(z) &= P(Z = z) = P(X + Y = z) = \sum_{x=0}^{z} P(X = x, Y = z - x) \\ &= \sum_{x=0}^{z} P(X = x)P(Y = z - x) = \sum_{x=0}^{z} \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{z-x} e^{-\mu}}{(z-x)!} \\ &= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^{z} \binom{z}{x} \lambda^x \mu^{(z-x)} = \frac{(\lambda+\mu)^z e^{-(\lambda+\mu)}}{z!}, \end{aligned}$$

por lo que $Z \sim$ Poisson$(\lambda + \mu)$. Luego,

$$\begin{aligned} p_{X|Z=z}(x|z) &= \frac{p_{X,Z}(x,z)}{p_Z(z)} = \frac{P(X = x, Z = z)}{P(Z = z)} = \frac{P(X = x, Y = z - x)}{P(Z = z)} \\ &= \frac{P(Y = y)P(Y = z - x)}{P(Z = z)} = \frac{\frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{z-x} e^{-\mu}}{(z-x)!}}{\frac{(\lambda+\mu)^z e^{-(\lambda+\mu)}}{z!}} \\ &= \binom{z}{x} \left(\frac{\lambda}{\lambda+\mu}\right)^x \left(1 - \frac{\lambda}{\lambda+\mu}\right)^{z-x}, x = 0, 1, \ldots, z, \end{aligned}$$

y hemos así probado que $X|Z = z \sim$ Bin$(z, \frac{\lambda}{\lambda+\mu})$.

Note que probar el hecho que $Z \sim$ Poisson$(\lambda + \mu)$ se puede hacer en forma alternativa, recurriendo a funciones generadoras. Se propone ésto como ejercicio.

Ejemplo 5.2.2 Se dispone de n monedas, cada una con probabilidad $0 < p < 1$ de dar cara. Considere el siguiente experimento. Se lanza cada moneda, independientemente de las demás. Posteriormente, aquellas monedas que dieron sello se lanzan una vez más, independientemente entre sí y de los lanzamientos en la etapa anterior. Obtengamos la distribución del número total de caras al final de este experimento.

Método I: Sean X e Y el número de caras registrados en la primera y segunda ronda de lanzamientos, respectivamente. Entonces, $X \sim \text{Bin}(n,p)$ e $Y|X = x \sim \text{Bin}(n-x,p)$, y la variable que nos interesa es $Z = X + Y$. Luego, para $z \in \mathcal{Z} = \{0,1,\ldots,n\}$,

$$\begin{aligned}
p_Z(z) &= P(Z=z) = P(X+Y=z) = \sum_{k=0}^{z} P(X=k, Y=z-k) \\
&= \sum_{k=0}^{z} P(X=k)P(Y=z-k|X=k) \\
&= \sum_{k=0}^{z} \binom{n}{k} p^k (1-p)^{n-k} \binom{n-k}{z-k} p^{z-k}(1-p)^{n-z} \\
&= \binom{n}{z} p^z (1-p)^{2n-z} \sum_{k=0}^{z} \binom{z}{k} (1-p)^{-k} \\
&= \binom{n}{z} p^z (1-p)^{2n-z} \left(1 + \frac{1}{1-p}\right)^z \\
&= \binom{n}{z} p^z (1-p)^{2n-z} \left(\frac{2-p}{1-p}\right)^z \\
&= \binom{n}{z} (p(2-p))^z \left((1-p)^2\right)^{n-z},
\end{aligned}$$

y notando que $p(2-p) + (1-p)^2 = 1$ para cualquier $p \in [0,1]$, se concluye que $Z \sim \text{Bin}(n, p(2-p))$.

Método II: Consideremos ahora variables aleatorias $X_1, \ldots, X_n$ tales que $X_i = 1$ si la i-ésima moneda dio cara *al final del experimento*, 0 en caso contrario. Es decir, $X_i = 1$ cuando la i-ésima moneda da cara después de ya sea el primer o segundo lanzamientos. Se tiene que la cantidad de interés se obtiene mediante la suma $X_1 + \cdots + X_n$, en donde las variables en esta suma son i.i.d. con distribución Bernoulli. Para calcular $P(X_i = 1)$, observe que $X_i = 0$ es equivalente a obtener dos sellos en igual número de lanzamientos independientes de una moneda con probabilidad $1-p$ de dar cara. Luego, para $i = 1, 2, \ldots, n$ se tiene

$$P(X_i = 1) = 1 - P(X_i = 0) = 1 - (1-p)^2 = p(2-p),$$

y se concluye que el número total de caras tiene distribución Binomial, correspondiente a n ensayos, cada uno con probabilidad de éxito dada por $p(2-p)$.

Ejemplo 5.2.3 Considere un par (X,Y) con distribución uniforme en el círculo unitario descrito por $\{(x,y) \ : \ x^2 + y^2 \leq 1\}$, y calcule $P(X > 0{,}5|Y < 0{,}25)$.

Note que la probabilidad buscada es igual a

$$\frac{P(X > 0{,}5, Y < 0{,}25)}{P(Y < 0{,}25)},$$

las cuales se calculan como cuocientes de áreas (ver Figura 5.2.1). Así,

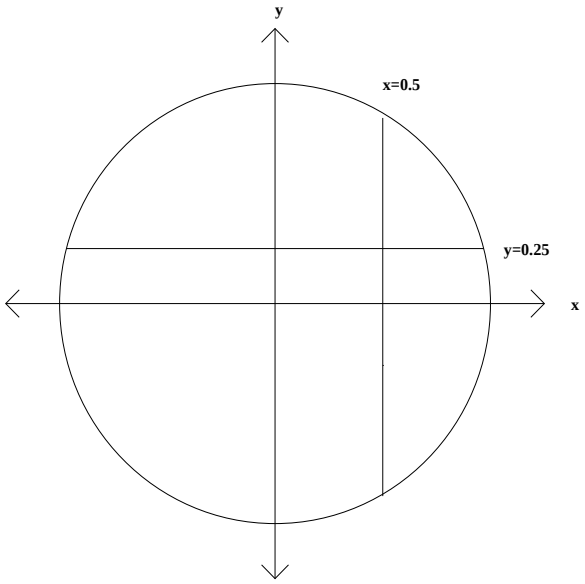


Figura 5.2.1: Diagrama para el Ejemplo 5.2.3.

$$\begin{aligned} P(Y < 0{,}25) &= \int_{-1}^{0{,}25} \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi} dx dy = \int_{-1}^{0{,}25} \frac{2\sqrt{1-y^2}}{\pi} dy \\ &= \frac{y\sqrt{1-y^2}}{\pi}\Big|_{-1}^{0{,}25} + \frac{\sin^{-1}(y)}{\pi}\Big|_{-1}^{0{,}25} = \frac{\sqrt{15}}{16\pi} + \frac{\sin^{-1}(0{,}25)}{\pi} + \frac{1}{2} \\ &\approx 0{,}6574811787. \end{aligned}$$

Por otra parte,

$$\begin{aligned} P(X < 0{,}5, Y > 0{,}25) &= \int_{-1}^{0{,}25} \int_{0{,}5}^{\sqrt{1-y^2}} \frac{1}{\pi} dx dy = \int_{-1}^{0{,}25} \frac{\sqrt{1-y^2} - 0{,}5}{\pi} dy \\ &= \left[\frac{y\sqrt{1-y^2}}{2\pi} + \frac{\sin^{-1}(y)}{2\pi} \right]\Big|_{-1}^{0{,}25} - \frac{0{,}25+1}{2\pi} \\ &= \frac{\sqrt{15}}{32\pi} + \frac{\sin^{-1}(0{,}25)}{2\pi} + \frac{1}{4} - \frac{5}{8\pi}, \\ &\approx 0{,}1297969105, \end{aligned}$$

y la probabilidad pedida es el cuociente entre dichas cantidades, lo que da aproximadamente 0.1974154009.

Note que aunque el Ejemplo 5.2.3 está originalmente planteado en términos de variables aleatorias continuas, la probabilidad condicional calculada corresponde básicamente a una discretización de dichas variables en términos de intervalos. Si la probabilidad pedida fuese $P(X > 0{,}5|Y = 0{,}25)$, nuestra actual definición de probabilidad condicional no se puede aplicar, pues por ser Y una variable aleatoria continua, se tiene que $P(Y = 0{,}25) = 0$.

Esto requiere entonces una definición más general de distribución condicional, lo que se discute a continuación.

5.3. Definición General de Distribución Condicional

Para motivar la definición, consideremos dos variables aleatorias X e Y con densidad conjunta $f_{X,Y}(x, y)$, definidas en un subconjunto apropiado de $\mathbb{R}^2$. Supongamos se quiere calcular la probabilidad del evento $X \in A$, sabiendo que Y tomó el valor y. Es necesario hacer la precisión que el hecho que $P(Y = y) = 0$ no significa que Y no pueda jamás tomar el valor y. Esta aparente contradicción es sólo producto del modelo matemático que hemos adoptado para tratar variables aleatorias. No obstante lo anterior, cuando se opera con variables aleatorias continuas, los eventos de interés son usualmente intervalos o uniones de ellos.

Para resolver el problema planteado, consideremos un peque no intervalo $(y - \epsilon, y + \epsilon]$ para $\epsilon > 0$. Para dar sentido a la expresión $P(X \in A|Y = y)$, usaremos un argumento basado en límites.

Definición 5.3.1 Sean X e Y variables aleatorias. Se define la probabilidad condicional que $X \in A$ dado que $Y = y$ mediante

$$P(X \in A|Y = y) = \lim_{\epsilon \to 0^+} P(X \in A|y - \epsilon < Y \leq y + \epsilon). \tag{5.3.1}$$

Más generalmente, si B es un evento definido en términos de una o más variables aleatorias $X_1, \ldots, X_n$, (por ejemplo, $\{X_1 + X_2 > X_3\}$), se define la probabilidad condicional de B dado que $Y = y$ mediante

$$P(B|Y = y) = \lim_{\epsilon \to 0^+} P(B|y - \epsilon < Y \leq y + \epsilon). \tag{5.3.2}$$

En particular, la *función de distribución acumulada condicional* de X dado que $Y = y$ se define mediante

$$F_{X|Y=y}(x|y) = P(X \leq x|Y = y) = \lim_{\epsilon \to 0^+} P(X \leq x|y - \epsilon < Y \leq y + \epsilon). \tag{5.3.3}$$

Veamos algunas consecuencias de la Definición 5.3.1. En primer lugar, si X e Y son independientes, entonces, para cualquier $\epsilon > 0$

$$P(X \in A|y - \epsilon < Y \leq y + \epsilon) = P(X \in A),$$

de modo que el límite en (5.3.1) se reduce a $P(X \in A)$, tal como se espera desde un punto de vista intuitivo. Note además que este resultado no depende del tipo de variable involucrada.

En segundo lugar, observe que si ambas variables son discretas, la definición (5.3.1) tendrá sentido sólo si y es un valor tal que $P(Y = y) > 0$. Observe además que puesto que hemos asumido que el soporte de Y contiene sólo puntos con probabilidad estrictamente positiva, se concluye que observar y tal que $P(Y = y) = 0$ es imposible. Así,

$$\begin{aligned}
\lim_{\epsilon\to 0^+} P(X \in A | y-\epsilon < Y \leq y+\epsilon) &= \lim_{\epsilon\to 0^+} \frac{P(X \in A, y-\epsilon < Y \leq y+\epsilon)}{P(y-\epsilon < Y \leq y+\epsilon)} \\
&= \frac{\lim_{\epsilon\to 0^+} P(X \in A, y-\epsilon < Y \leq y+\epsilon)}{\lim_{\epsilon\to 0^+} P(y-\epsilon < Y \leq y+\epsilon)} \\
&= \frac{P(X \in A, Y = y)}{P(Y = y)} = P(X \in A | Y = y) \\
&= \sum_{x \in A \cap \mathcal{X}} p_{X|Y=y}(x|y),
\end{aligned}$$

tal como se tenía hasta el momento.

En tercer lugar, y volviendo a la situación del comienzo de esta sección, suponga que X e Y tienen densidad conjunta $f_{X,Y}(x, y)$. Entonces

$$\begin{aligned}
P(X \leq x | Y = y) &= \lim_{\epsilon\to 0^+} P(X \leq x | y-\epsilon < Y \leq y+\epsilon) \\
&= \lim_{\epsilon\to 0^+} \frac{P(X \leq x, y-\epsilon < Y \leq y+\epsilon)}{P(y-\epsilon < Y \leq y+\epsilon)} \\
&= \lim_{\epsilon\to 0^+} \frac{\frac{1}{2\epsilon}\int_{-\infty}^{x}\int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(s,t)dtds}{\frac{1}{2\epsilon}\int_{y-\epsilon}^{y+\epsilon} f_Y(t)dt} \\
&= \frac{\int_{-\infty}^{x}\left(\lim_{\epsilon\to 0^+} \frac{1}{2\epsilon}\int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(s,t)dt\right)ds}{\lim_{\epsilon\to 0^+} \frac{1}{2\epsilon}\int_{y-\epsilon}^{y+\epsilon} f_Y(t)dt},
\end{aligned}$$

y usando el Teorema del Valor Medio para integrales se obtiene la siguiente expresión para la función de distribución acumulada condicional de X dado que $Y = y$:

$$F_{X|Y=y}(x|y) = \int_{-\infty}^{x} \frac{f_{X,Y}(s,y)}{f_Y(y)} ds. \tag{5.3.4}$$

Definición 5.3.2 Si X e Y poseen densidad conjunta $f_{X,Y}(x, y)$, se define la *densidad condicional* de X dado que $Y = y$ mediante

$$f_{X|Y=y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{5.3.5}$$

Observe que (5.3.5) se obtiene de (5.3.4) mediante diferenciación. Note que (5.3.5) es una función densidad. En efecto, ella es siempre no negativa, por ser un cuociente entre funciones no negativas, y además,

$$\int_{-\infty}^{\infty} f_{X|Y=y}(x|y)dx = \frac{1}{f_Y(y)}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \frac{f_Y(y)}{f_Y(y)} = 1.$$

Adicionalmente, si se asume que X e Y son independientes, entonces se tiene $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, de modo que

$$f_{X|Y=y}(x|y) = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x).$$

Así, en el caso de independencia, la densidad condicional de X dado que $Y = y$ se transforma simplemente en la densidad marginal de X, tal como debía esperarse intuitivamente.

Un resultado básico relativo a probabilidades condicionales para eventos es el Teorema de Probabilidades Totales. Enunciamos a continuación una generalización al caso continuo.

Teorema 5.3.1 *Sea B un evento, y X una variable aleatoria con densidad $f_X(x)$. Entonces*

$$P(B) = \int_{-\infty}^{\infty} P(B|X = x)f_X(x)dx. \tag{5.3.6}$$

Queda aún por discutir el caso mixto. Aquí lo usual es que la distribución conjunta de las variables involucradas se defina en términos de distribuciones condicionales de una variable aleatoria dada la otra, la que se combina con la distribución marginal de la variable que condiciona. Este enfoque es ligeramente distinto de lo expuesto hasta el momento, en el que las distribuciones condicionales se definieron a partir de la distribución conjunta. Así, por ejemplo, si $X|Y = y$ es discreta, con distribución dependiente de y, e Y es continua con densidad $f_Y(y)$, entonces la función de probabilidad discreta conjunta está dada por $p_{X,Y}(x,y) = p_{X|Y=y}(x|y)f_Y(y)$.

Otra situación que aparece con frecuencia, es una generalización del Teorema de Bayes visto en el Capítulo 2. Supongamos que se conoce la distribución condicional de X dado que $Y = y$, y la distribución marginal de Y. ¿Cómo se calcula la distribución de Y dado que $X = x$? La interpretación que se suele dar a este proceso es como sigue. Los estados de la naturaleza se describen, antes de hacer un experimento, mediante los valores de Y. La opinión que se tiene de esta naturaleza, se describe desde un punto de vista probabilístico mediante la distribución de Y, usualmente llamada distribución *a priori*. Suponiendo que el estado de la naturaleza es y, la variable aleatoria X, que representa el resultado de un cierto experimento a realizar, tiene distribución $X|Y = y$. Se realiza dicho experimento, y se observa el valor x de una variable aleatoria X. Como resultado de este experimento, actualizamos nuestra opinión de la naturaleza, mediante el cálculo de la distribución de Y dado que $X = x$, también llamada distribución *a posteriori*.

Veremos a continuación la forma de realizar estos cálculos.

1. X **e** Y **son discretas:** en este caso el cálculo es relativamente sencillo. Usando el hecho que $p_{X,Y}(x,y) = p_{X|Y=y}(x|y)p_Y(y)$, y $p_X(x) = \sum_y p_{X,Y}(x,y)$, se obtiene la fórmula

$$p_{Y|X=x}(y|x) = \frac{p_{X|Y=y}(x|y)p_Y(y)}{\sum_{s\in\mathcal{Y}} p_{X|Y=s}(x|s)p_Y(s)}. \tag{5.3.7}$$

2. X **e** Y **son continuas:** en este caso es posible probar que (X,Y) tiene densidad conjunta dada por $f_{X,Y}(x,y) = f_{X|Y=y}(x|y)f_Y(y)$. La densidad marginal de X se obtiene de $f_X(x) =$

$\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$, de modo que se tiene la expresión

$$f_{Y|X=x}(y|x) = \frac{f_{X|Y=y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y=s}(x|s)f_Y(s)ds}. \tag{5.3.8}$$

3. X **es discreta e** Y **es continua:** la distribución marginal de X se obtiene mediante la fórmula $p_X(x) = \int_{-\infty}^{\infty} p_{X|Y=y}(x|y)f_Y(y)dy$, expresión que se obtiene del Teorema 5.3.1, por lo que

$$f_{Y|X=x}(y|x) = \frac{p_{X|Y=y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} p_{X|Y=s}(x|s)f_Y(s)ds}. \tag{5.3.9}$$

4. X **es continua e** Y **es discreta:** mediante cálculos análogos a los mostrados, se obtiene que

$$p_{Y|X=x}(y|x) = \frac{f_{X|Y=y}(x|y)p_Y(y)}{\sum_{s\in\mathcal{S}} f_{X|Y=s}(x|s)p_Y(s)}. \tag{5.3.10}$$

Por último, la generalización de los conceptos vistos al caso de más variables es directa. Veamos a continuación algunos ejemplos.

Ejemplo 5.3.1 Sean X e Y variables aleatorias independientes con $X \sim \text{Exp}(\lambda)$ e $Y \sim \text{Exp}(\mu)$, donde $\lambda, \mu > 0$. Calcule $P(X > Y)$.

Una forma de resolver este problema consiste en calcular la densidad de $W = X - Y$, con lo que la probabilidad pedida es simplemente $\int_0^{\infty} f_W(w)dw$. Sin embargo, por el Teorema 5.3.1 se tiene

$$P(X > Y) = \int_0^{\infty} P(X > Y|Y = y)f_Y(y)dy.$$

El paso crucial del argumento consiste en calcular $P(X > Y|Y = y)$. Una vez que se condiciona en $Y = y$, se puede substituir dicho valor en el evento al lado izquierdo de la probabilidad condicional, lo que se conoce como *Principio de Substitución*. Así, $P(X > Y|Y = y) = P(X > y|Y = y)$. Pero una vez que se ha hecho esta substitución, el evento de interés $\{X > y\}$ en la probabilidad condicional, ya no depende de la variable aleatoria Y, esto es, depende sólo de X, y puesto que X e Y son independientes, se concluye que $P(X > y|Y = y) = P(X > y) = e^{-y/\lambda}$. Luego,

$$\begin{aligned} P(X > Y) &= \int_0^{\infty} e^{-y/\lambda}\frac{e^{-y/\mu}}{\mu} = \frac{1}{\mu\left(\frac{1}{\lambda}+\frac{1}{\mu}\right)} \\ &= \frac{\lambda}{\lambda+\mu}. \end{aligned}$$

Se propone como ejercicio obtener este resultado mediante el cálculo de la densidad de $W = X - Y$.

Ejemplo 5.3.2 Considere el par (X, Y) del Ejemplo 4.2.4, y calcule la densidad condicional de X dado que $Y = y$. Puesto que ambos $f_{X,Y}$ y f_Y se tienen de lo hecho en el Ejemplo 4.2.4, lo pedido se obtiene directamente de (5.3.5):

$$f_{X|Y=y}(x|y) = \frac{\frac{3}{4}(|x| + |y|)}{\frac{3}{4}(1 - y^2)} = \frac{|x| + |y|}{1 - y^2},$$

para $-(1 - |y|) \leq x \leq 1 - |y|$.

Ejemplo 5.3.3 Suponga que $X|Y = y \sim$ Poisson(y), e $Y \sim \Gamma(\alpha, \lambda)$, con $\alpha > 0$ y $\lambda > 0$. Calcule la densidad de Y dado que $X = x$.

Se tiene que para $x \in \{0, 1, 2, \ldots, \}$

$$\begin{aligned} p_X(x) &= \int_0^\infty \frac{y^x e^{-y}}{x!} \frac{y^{\alpha-1} e^{-y/\lambda}}{\Gamma(\alpha)\lambda^\alpha} dy = \frac{1}{x!\Gamma(\alpha)\lambda^\alpha} \int_0^\infty y^{\alpha+x-1} e^{-y(1+1/\lambda)} dy \\ &= \frac{\Gamma(\alpha + x)}{x!\Gamma(\alpha)\lambda^\alpha \left(1 + \frac{1}{\lambda}\right)^{\alpha+x}} \end{aligned}$$

Luego, por (5.3.9), y después de simplificar las expresiones se obtiene que

$$f_{Y|X=x}(y|x) = \frac{y^{\beta-1} e^{-y/\mu}}{\Gamma(\beta)\mu^\beta}, \qquad , y > 0,$$

en donde $\beta = \alpha + x$ y $\mu = \lambda/(1 + \lambda)$, y se concluye que $Y|X = x \sim \Gamma(\beta, \mu)$.

Ejemplo 5.3.4 Considere un vector aleatorio $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{y} \quad \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right),$$

en donde $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$ tienen dimensiones respectivas k y l, con $k + l = n$. Calculemos la distribución condicional de $\boldsymbol{X}_1$ dado que $\boldsymbol{X}_2 = \boldsymbol{x}_2$.

Para ello, considere el vector $\boldsymbol{W} = \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$. Se tiene que, por las propiedades de las matrices de covarianza, y recordando que puesto que $\boldsymbol{\Sigma}_{22}$ es simétrica, su inversa también lo es:

$$\begin{aligned} Cov(\boldsymbol{X}_2, \boldsymbol{W}) &= Cov(\boldsymbol{X}_2, \boldsymbol{X}_1) - Cov(\boldsymbol{X}_2, \boldsymbol{X}_2)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= 0. \end{aligned}$$

Así, usando la Proposición 4.7.3(i) con

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}_k & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & \boldsymbol{I}_l \end{pmatrix},$$

se tiene que el vector

$$\begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{X}_2 \end{pmatrix}$$

tiene distribución conjunta normal multivariada, por lo que $\boldsymbol{X}_2$ y $\boldsymbol{W}$ son vectores aleatorios independientes (recuerde que en el caso de la distribución normal multivariada, independencia es equivalente a la no correlación). Así, la distribución condicional de $\boldsymbol{W}$ dado que $\boldsymbol{X}_2 = \boldsymbol{x}_2$ es simplemente la distribución marginal (no condicional) de $\boldsymbol{W}$. Un cálculo directo, muestra que $\boldsymbol{W} \sim N_k(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$. Pero puesto que $\boldsymbol{W} = \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$, por el principio de substitución introducido en el Ejemplo 5.3.1, la distribución condicional de $\boldsymbol{W}$ dado que $\boldsymbol{X}_2 = \boldsymbol{x}_2$ coincide con aquella de $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{x}_2$ dado que $\boldsymbol{X}_2 = \boldsymbol{x}_2$. Puesto que después de condicionar en $\boldsymbol{X}_2 = \boldsymbol{x}_2$ la cantidad $-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{x}_2$ es simplemente una constante, el resultado final se obtiene de restar dicha constante a la distribución condicional de $\boldsymbol{W}$, para obtener

$$\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_k(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{x}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Para el caso particular en que $k = l = 1$ (esto es, $n = 2$), y con

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

se obtiene

$$X_1|X_2 = x_2 \sim N\left(\mu_1 - \frac{\rho\sigma_1}{\sigma_2}(\mu_2 - x_2), \sigma_1^2(1-\rho^2)\right).$$

Ejemplo 5.3.5 Sean $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, donde $\lambda > 0$. Considere las variables aleatorias definidas mediante

$$\begin{aligned} Y_1 &= X_{(1)}, \\ (Y_2, ..., Y_n) &= \begin{cases} (X_2 - X_1, \ldots, X_n - X_1) \text{ si } X_{(1)} = X_1 \\ (X_1 - X_2, X_3 - X_2, ..., X_n - X_2) \text{ si } X_{(1)} = X_2 \\ \qquad\qquad \vdots \\ (X_1 - X_n, \ldots, X_{n-1} - X_n) \text{ si } X_{(1)} = X_n \end{cases} \end{aligned}$$

Observe que la definición de $Y_2, \ldots, Y_n$ consiste en las variables $X_1 - X_{(1)}, \ldots, X_n - X_{(1)}$, después de eliminar aquella que es idénticamente 0. Obtengamos la distribución conjunta de $Y_1, \ldots, Y_n$. Para ello, defina los eventos

$$\begin{aligned} A &= \{Y_1 > y_1, \ldots, Y_n > y_n\} \\ B_i &= \{X_{(1)} = X_i\}, \;\; i = 1, \ldots, n, \end{aligned}$$

donde $y_1, \ldots, y_n > 0$. Se tiene entonces que

$$P(Y_1 > y_1, \ldots, Y_n \leq y_n) = \sum_{i=1}^{n} P(A \cap B_i).$$

Ahora, por el Teorema 5.3.1 se tiene

$$P(A \cap B_i) = \int_0^\infty P(A \cap B_i | X_i = x_i) f_{X_i}(x_i) dx_i.$$

Por otra parte,

$$\begin{aligned}\{A \cap B_i\} &= \{X_i > y_1, X_1 - X_i > y_2, \ldots, X_{i-1} - X_i > y_i, \\ &\quad X_{i+1} - X_i > y_{i+1}, \ldots, X_n - X_i > y_n, X_1 > X_i, \\ &\quad \ldots, X_{i-1} > X_i, X_{i+1} > X_i, \ldots X_n > X_i\} \\ &= \{X_i > y_1, X_1 > y_2 + X_i, \ldots, X_{i-1} > y_i + X_i, \\ &\quad \ldots, X_{i+1} > y_{i+1} + X_i, \ldots, X_n > y_n + X_i\},\end{aligned}$$

por lo que, usando el principio de substitución y el hecho que $X_1, \ldots, X_n$ son i.i.d. Exp(λ), se tiene que

$$\begin{aligned}P(A \cap B_i) &= \int_0^\infty P(X_i > y_1, X_1 > y_2 + X_i, ..., X_{i-1} > y_i + X_i, \\ &\quad ..., X_{i+1} > y_{i+1} + X_i, ..., X_n > y_n + X_i | X_i = x_i) \frac{e^{-x_i/\lambda}}{\lambda} dx_i \\ &= \int_0^\infty P(x_i > y_1, X_1 > y_2 + x_i, ..., X_{i-1} > y_i + x_i, \\ &\quad ..., X_{i+1} > y_{i+1} + x_i, ..., X_n > y_n + x_i | X_i = x_i) \frac{e^{-x_i/\lambda}}{\lambda} dx_i \\ &= \int_0^\infty P(x_i > y_1, X_1 > y_2 + x_i, ..., X_{i-1} > y_i + x_i, \\ &\quad ..., X_{i+1} > y_{i+1} + x_i, ..., X_n > y_n + x_i) \frac{e^{-x_i/\lambda}}{\lambda} dx_i \\ &= \int_{y_1}^\infty P(X_1 > y_2 + x_i, ..., X_{i-1} > y_i + x_i, ..., \\ &\quad X_{i+1} > y_{i+1} + x_i, ..., X_n > y_n + x_i) \frac{e^{-x_i/\lambda}}{\lambda} dx_i \\ &= \int_{y_1}^\infty \prod_{j=1}^{i-1} P(X_j > y_{j+1} + x_i) \prod_{j=i+1}^{n} P(X_j > y_j + x_i) \frac{e^{-x_i/\lambda}}{\lambda} dx_i \\ &= \int_{y_1}^\infty \prod_{j=1}^{i-1} e^{-(y_{j+1}+x_i)/\lambda} \prod_{j=i+1}^{n} e^{-(y_j+x_i)/\lambda} \frac{e^{-x_i/\lambda}}{\lambda} dx_i \\ &= e^{-\sum_{j=2}^{n} y_j/\lambda} \int_{y_1}^\infty e^{-nx_i/\lambda} dx_i \\ &= \frac{1}{n} e^{-\sum_{j=2}^{n} y_j/\lambda} e^{-ny_1/\lambda}\end{aligned}$$

Luego,

$$P(A) = \sum_{i=1}^{n} \frac{1}{n} e^{-\sum_{j=2}^{n} y_j/\lambda} e^{-ny_1/\lambda} = e^{-\sum_{j=2}^{n} y_j/\lambda} e^{-ny_1/\lambda},$$

y se concluye que $Y_1, \ldots, Y_n$ son independientes, con $Y_1 \sim \text{Exp}(\lambda/n)$, e $Y_j \sim \text{Exp}(\lambda)$ para $j = 2, \ldots, n$. En particular, se deduce el siguiente resultado, que es útil en Infe-

rencia Estadística: $X_{(1)} \sim \text{Exp}(\lambda/n)$ y $\sum_{i=1}^{n}(X_i - X_{(1)}) = \sum_{j=2}^{n} Y_j \sim \Gamma(n-1,\lambda)$ son independientes.

5.4. Esperanza Condicional

Pasamos a definir ahora el concepto de *esperanza condicional*, y a estudiar algunas de sus propiedades básicas. En forma intuitiva, la esperanza condicional es simplemente la esperanza de una distribución condicional.

Definición 5.4.1 Sean X e Y variables aleatorias. Se define la *esperanza condicional* de X dado que $Y = y$ mediante

$$E(X|Y=y) = \begin{cases} \sum_x x p_{X|Y=y}(x|y) & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} x f_{X|Y=y}(x|y)dx & \text{si } X \text{ es continua,} \end{cases} \tag{5.4.1}$$

si la suma o integral correspondiente converge absolutamente, lo que se tiene si $E(|X|) < \infty$.

La esperanza condicional así definida, tiene todas las propiedades que posee la esperanza $E(\cdot)$ definida en el Capítulo 3. Por ejemplo, si las expresiones involucradas existen, entonces dadas constantes a y b se tiene $E(aX + bZ|Y = y) = aE(X|Y = y) + bE(Z|Y = y)$. La razón de esto es que la esperanza condicional de X dado que $Y = y$ se puede ver simplemente como el valor esperado correspondiente a una cierta variable aleatoria W cuya distribución coincide con la de $X|Y = y$. De este modo, todas las propiedades para $E(\cdot)$ se cumplen para $E(\cdot|Y = y)$, incluyendo la correspondiente versión del Teorema 3.8.1:

$$E(g(X)|Y=y) = \int_{-\infty}^{\infty} g(x) f_{X|Y=y}(x|y)dx,$$

si X es continua, o bien reemplazando la integral por una suma si X es discreta. Este resultado permite definir momentos de la distribución condicional, y en particular la varianza condicional, en forma análoga a las versiones no condicionales correspondientes.

Definición 5.4.2 Sean X e Y variables aleatorias. En la medida que las expresiones involucradas existan, se define:

(a) El *momento condicional de orden* k de X dado que $Y = y$ mediante

$$\mu_k(X|Y=y) = E(X^k|Y=y). \tag{5.4.2}$$

(b) La *varianza condicional* de X dado que $Y = y$ mediante

$$Var(X|Y=y) = E(X^2|Y=y) - (E(X|Y=y))^2. \tag{5.4.3}$$

Algunas propiedades de la esperanza condicional son consecuencia del Principio de Substitución. Así, para una función de dos variables g, se tiene que

$$E(g(X,Y)|Y=y) = E(g(X,y)|Y=y),$$

y en particular,

$$E(g(X)h(Y)|Y=y) = E(g(X)h(y)|Y=y) = h(y)E(g(X)|Y=y),$$

siempre y cuando las expresiones involucradas existan.

Ejemplo 5.4.1 En el Ejemplo 5.2.1 se tiene que $X|Z=z \sim \text{Bin}(z, \frac{\lambda}{\lambda+\mu})$, de modo que $E(X|Z=z) = \frac{z\lambda}{\lambda+\mu}$.

Ejemplo 5.4.2 Si $X|Y=y \sim \text{Bin}(n,y)$ e $Y \sim \text{Beta}(a,b)$ con $a,b>0$, calcule $E(Y|X=x)$.

De acuerdo a la Definición 5.4.1, necesitamos previamente obtener la distribución de Y dado que $X=x$. Por el Teorema 5.3.1, y para $x \in \{0,1,2,\ldots,n\}$:

$$\begin{aligned} p_X(x) &= \int_0^1 \binom{n}{x} y^x (1-y)^{n-x} \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)} dy \\ &= \frac{\binom{n}{x}}{B(a,b)} \int_0^1 y^{a+x-1}(1-y)^{b+n-x-1} dy \\ &= \frac{\binom{n}{x} B(a+x, b+n-x)}{B(a,b)} \end{aligned}$$

Luego, por (5.3.9), y después de simplificar términos, se obtiene que

$$f_{Y|X=x}(y|x) = \frac{y^{a+x-1}(1-y)^{b+n-x-1}}{B(a+x, b+n-x)},$$

de modo que $Y|X=x \sim \text{Beta}(a+x, b+n-x)$. Por consiguiente, $E(Y|X=x) = \frac{a+x}{a+b+n}$ (ver Ejemplo 3.8.3).

Ejemplo 5.4.3 En el Ejemplo 5.3.4 tenemos $E(\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{x}_2)$.

En todos estos ejemplos, se tiene que la esperanza condicional de una variable o vector aleatorio dado el valor de otra variable o vector aleatorio, se expresa como una función del valor de la variable que condiciona. Así, en el Ejemplo 5.4.1, $E(X|Z=z) = \frac{z\lambda}{\lambda+\mu}$, que es una función de z. Esto motiva la siguiente definición.

Definición 5.4.3 Sean X e Y variables aleatorias con $E(|X|) < \infty$. La *esperanza condicional* de X dado Y es una variable aleatoria que se denota por $E(X|Y)$, y definida como $\varphi(Y)$, donde

$$\varphi(y) = E(X|Y=y). \tag{5.4.4}$$

Así, en el Ejemplo 5.4.1 se tiene que $\varphi(z) = E(X|Z=z) = \frac{\lambda z}{\lambda+\mu}$, de modo que

$$E(X|Z) = \varphi(Z) = \left(\frac{\lambda}{\lambda+\mu}\right) Z.$$

A modo de receta, para calcular $E(X|Y)$, basta reemplazar "y" por "Y", una vez calculado el valor de $E(X|Y=y)$.

En el mismo Ejemplo 5.4.1, note que puesto que $Z \sim \text{Poisson}(\lambda+\mu)$, entonces $E(Z) = \lambda+\mu$, y se tiene que

$$E(E(X|Z)) = E\left(\left\{\frac{\lambda}{\lambda+\mu}\right\} Z\right) = \frac{\lambda}{\lambda+\mu} E(Z) = \lambda = E(X).$$

Lejos de ser una coincidencia, esto es resultado de una de las propiedades básicas de esperanzas condicionales.

Teorema 5.4.1 *Sean X e Y variables aleatorias con $E(|X|) < \infty$. Entonces*

$$E(X) = E(E(X|Y)). \tag{5.4.5}$$

Aunque no daremos una demostración del Teorema 5.4.1 en el caso general, es ilustrativo considerar lo que sucede en el caso que X e Y poseen densidad conjunta $f_{X,Y}(x,y)$. Puesto que

$$\varphi(y) = E(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x|y)dx = \int_{-\infty}^{\infty} \frac{x f_{X,Y}(x,y)}{f_Y(y)} dx,$$

entonces

$$\begin{aligned} E(E(X|Y)) &= E(\varphi(Y)) = \int_{-\infty}^{\infty} \varphi(y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \frac{x f_{X,Y}(x,y)}{f_Y(y)} dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy \\ &= E(X). \end{aligned}$$

El tratamiento del caso general requiere conceptos de *Teoría de la Medida*, que van más allá de los objetivos de este texto.

Veamos ahora otras dos propiedades útiles de la esperanza condicional, que son consecuencias del Teorema 5.4.1.

Proposición 5.4.1 Sean X, Y y Z variables aleatorias.

(a) Si $E(X^2) < \infty$, entonces

$$Var(X) = E(Var(X|Y)) + Var(E(X|Y)) \tag{5.4.6}$$

(b) Si $E(X^2)$ y $E(Y^2)$ son ambas finitas, entonces

$$Cov(X,Y) = E(Cov((X,Y)|Z)) + Cov(E(X|Z), E(Y|Z)), \tag{5.4.7}$$

en donde la *covarianza condicional* entre X e Y dado Z se define como $Cov((X,Y)|Z) = E(XY|Z) - E(X|Z)E(Y|Z)$.

Demostración:

(a) Se tiene que

$$\begin{aligned} E(Var(X|Y)) &= E\{E(X^2|Y) - (E(X|Y))^2\} = E(E(X^2|Y)) - E\{(E(X|Y))^2\} \\ &= E(X^2) - E\{(E(X|Y))^2\}. \end{aligned}$$

y por otra parte,

$$\begin{aligned} Var(E(X|Y)) &= E\{(E(X|Y))^2\} - \{E(E(X|Y))\}^2 \\ &= E\{(E(X|Y))^2\} - (E(X))^2, \end{aligned}$$

y el resultado se obtiene de sumar estas expresiones.

(b) El procedimiento para este caso es similar al de (a). En efecto,

$$\begin{aligned} E(Cov((X,Y)|Z)) &= E(E(XY|Z) - E(X|Z)E(Y|Z)) \\ &= E(XY) - E(E(X|Z)E(Y|Z)), \end{aligned}$$

y además

$$\begin{aligned} Cov(E(X|Z), E(Y|Z)) &= E(E(X|Z)E(Y|Z)) - E(E(X|Z))E(E(Y|Z)) \\ &= E(E(X|Z)E(Y|Z)) - E(X)E(Y), \end{aligned}$$

y el resultado se obtiene de sumar las expresiones obtenidas. ■

Ejemplo 5.4.4 En el Ejemplo 5.2.2, suponga que sólo nos interesa calcular el valor esperado y varianza del número total de monedas que dan cara al final de las dos rondas del experimento. Con la notación usada en su momento, dicho número es $Z = X + Y$, donde $X \sim \text{Bin}(n,p)$, e $Y|X = x \sim \text{Bin}(n-x,p)$. Se tiene que

$$\begin{aligned} E(Z|X=x) &= E(X+Y|X=x) = E(x+Y|X=x) \\ &= x + E(Y|X=x) = x + (n-x)p \\ &= np + (1-p)x, \end{aligned}$$

de modo que $E(Z|X) = np + (1-p)X$, y así

$$\begin{aligned} E(Z) &= E(E(Z|X)) = E(np + (1-p)X) = np + (1-p)E(X) \\ &= np + (1-p)np = np(2-p). \end{aligned}$$

Por otra parte,

$$\begin{aligned} Var(Z|X=x) &= Var(X+Y|X=x) = Var(x+Y|X=x) \\ &= Var(Y|X=x) = (n-x)p(1-p), \end{aligned}$$

de donde, $Var(Z|X) = p(1-p)(n-X)$, y

$$\begin{aligned} E(Var(Z|X)) &= E(p(1-p)(n-X)) = p(1-p)E(n-X) \\ &= p(1-p)(n-E(X)) = p(1-p)(n-np) \\ &= np(1-p)^2. \end{aligned}$$

Además,

$$\begin{aligned} Var(E(Z|X)) &= Var(np+(1-p)X) = Var((1-p)X) \\ &= (1-p)^2 Var(X) = (1-p)^2 np(1-p) \\ &= np(1-p)^3. \end{aligned}$$

Así, por (5.4.6),

$$\begin{aligned} Var(Z) &= np(1-p)^2 + np(1-p)^3 = np(1-p)^2(1+1-p) \\ &= np(1-p)^2(2-p). \end{aligned}$$

Ejemplo 5.4.5 Sean W, X e Y variables aleatorias con densidad conjunta

$$f_{W,X,Y}(w,x,y) = \begin{cases} c(1+wxy) & \text{si } 0 \le w,x,y \le 1 \\ 0 & \text{si no.} \end{cases}$$

Obtengamos primero el valor de c. Se debe tener

$$\begin{aligned} 1 &= c \iiint_{[0,1]^3} (1+wxy)\,dwdxdy \\ &= c \iiint_{[0,1]^3} 1\,dwdxdy + c \iiint_{[0,1]^3} wxy\,dwdxdy \\ &= c\left(1+\frac{1}{8}\right) = \frac{9c}{8}, \end{aligned}$$

de modo que $c = \frac{8}{9}$. Calculemos ahora la distribución condicional de (W,X) dado que $Y=y$. Para ello, se necesita la densidad $f_Y(y)$, la que se calcula mediante

$$\begin{aligned} f_Y(y) &= \frac{8}{9}\int_0^1\int_0^1 (1+wxy)\,dwdx = \frac{8}{9}\int_0^1 \left(1+\frac{xy}{2}\right)dx = \frac{8}{9}\left(1+\frac{y}{4}\right) \\ &= \frac{8+2y}{9}, \qquad 0 \le y \le 1. \end{aligned}$$

Así, se obtiene

$$f_{W,X|Y=y}(w,x|y) = \frac{f_{W,X,Y}(w,x,y)}{f_Y(y)} = \frac{4(1+wxy)}{4+y}.$$

Verifiquemos ahora que (5.4.7) se cumple:

$$E(WX|Y=y)=\frac{4}{4+y}\int_0^1\int_0^1 wx(1+wxy)dxdw=\frac{9+4y}{36+9y},$$

y luego,

$$E(WX|Y)=\frac{9+4Y}{36+9Y}.$$

Además

$$E(W|Y=y)=\frac{4}{4+y}\int_0^1\int_0^1 w(1+wxy)dxdw=\frac{6+2y}{12+3y},$$

y por la simetría del problema se obtienen las esperanzas condicionales

$$E(W|Y)=\frac{6+2Y}{12+3Y}\qquad\text{y}\qquad E(X|Y)=\frac{6+2Y}{12+3Y}.$$

Luego,

$$Cov((W,X)|Y)=E(WX|Y)-E(W|Y)E(X|Y)=\frac{Y}{9(4+Y)^2},$$

de donde se obtiene

$$\begin{aligned}E(Cov((W,X)|Y)) &= \int_0^1\left(\frac{y}{9(4+y)^2}\times\frac{8+2y}{9}\right)dy\\ &= \frac{2-8\log(5)+16\log(2)}{81}.\end{aligned}$$

Por otra parte,

$$\begin{aligned}Cov(E(W|Y),E(X|Y)) &= Cov\left(\frac{6+2Y}{12+3Y},\frac{6+2Y}{12+3Y}\right)\\ &= Var\left(\frac{6+2Y}{12+3Y}\right).\end{aligned}$$

Ahora,

$$E\left(\frac{6+2Y}{12+3Y}\right)=\int_0^1\left(\frac{6+2y}{12+3y}\times\frac{8+2y}{9}\right)dy=\frac{14}{27},$$

y además

$$\begin{aligned}E\left\{\left(\frac{6+2Y}{12+3Y}\right)^2\right\} &= \int_0^1\left(\frac{(6+2y)^2}{(12+3y)^2}\frac{8+2y}{9}\right)dy\\ &= \frac{20+8\log(5)-16\log(2)}{81}\end{aligned}$$

por lo que

$$\begin{aligned}Var\left(\frac{6+2Y}{12+3Y}\right) &= \frac{20+8\log(5)-16\log(2)}{81}-\frac{14^2}{27^2}\\ &= \frac{8\log(5)-16\log(2)}{81}-\frac{16}{729}.\end{aligned}$$

Finalmente, sumando $E(Cov((W,X)|Y))$ y $Cov(E(X|Y),E(W|Y))$, se obtiene

$$\frac{2-8\log(5)+16\log(2)}{81}+\frac{8\log(5)-16\log(2)}{81}-\frac{16}{729}=\frac{2}{729}.$$

Calculemos ahora $Cov(W,X)$ directamente de la distribución conjunta inicial. Se tiene

$$E(WX)=\frac{8}{9}\int_0^1\int_0^1 wx(1+wxy)dxdwdy=\frac{22}{81},$$

y

$$E(W)=\frac{8}{9}\int_0^1\int_0^1 w(1+wxy)dxdwdy=\frac{14}{27},$$

donde, por simetría, $E(W)=E(X)$. Se obtiene así

$$\begin{aligned} Cov(W,X) &= E(WX)-E(W)E(X)=\frac{22}{81}-\frac{14^2}{27^2} \\ &= \frac{2}{729}, \end{aligned}$$

lo que coincide con lo que se obtuvo anteriormente. Cálculos semejantes permiten concluir que $Var(W)=Var(X)=\frac{121}{1458}$, por lo que $\rho(W,X)=\frac{4}{121}\approx 0{,}033$.

Ejemplo 5.4.6 Consideremos la situación del Ejemplo 4.5.4, la cual generalizamos suponiendo que $X_1,X_2,\ldots$ son i.i.d. con media μ y varianza σ^2, y consideramos N una variable aleatoria con soporte incluido en $\{1,2,\ldots\}$, con media ν y varianza τ^2. Así, definimos $S_N=\sum_{i=1}^{N}X_i$, esto es, $X_1+\cdots+X_n$ si $N=n$, con $n\geq 1$. Se asume además que N es independiente de $X_1,X_2,\ldots$. Calculemos ahora $E(S_N)$ y $Var(S_N)$. Se tiene que

$$E(S_N|N=n)=E(\sum_{i=1}^{n}X_i|N=n)=\sum_{i=1}^{n}E(X_i|N=n).$$

Pero como N es independiente de cada X_i, se tiene que $E(X_i|N=n)=E(X_i)=\mu$. Luego,

$$E(S_N|N=n)=\sum_{i=1}^{n}\mu=n\mu,$$

de donde $E(S_N|N)=N\mu$, y

$$E(S_N)=E(E(S_N|N))=E(N\mu)=\mu E(N)=\mu\nu.$$

Por otra parte,

$$Var(E(S_N|N))=Var(N\mu)=\mu^2Var(N)=\mu^2\tau^2,$$

y usando la independencia de los X's,

$$\begin{aligned} Var(S_N|N=n) &= Var(\sum_{i=1}^{n} X_i|N=n) = \sum_{i=1}^{n} Var(X_i|N=n) \\ &= \sum_{i=1}^{n} Var(X_i) = n\sigma^2. \end{aligned}$$

Luego, $Var(S_N|N) = N\sigma^2$, por lo que $E(Var(S_N|N)) = \nu\sigma^2$, de modo que por (5.4.6)

$$Var(S_N) = \mu^2\tau^2 + \nu\sigma^2.$$

Ejemplo 5.4.7 Considere X, Y y Z variables aleatorias independientes con $X, Y \sim N(0,1)$, y Z tiene densidad f_Z. Calcule la distribución de

$$W = \frac{X + ZY}{\sqrt{1+Z^2}},$$

y obtenga una expresión para $\rho(X,W)$ y $\rho(Y,W)$ en términos de la distribución de Z. Evalúe estas correlaciones para el caso en $Z \sim U(0,1)$.

En este ejemplo queda de manifiesto la utilidad de los argumentos basados en condicionamiento. Puesto que Z tiene la distribución más "complicada", condicionemos en un valor de Z. Entonces, dado que $Z = z$, W se transforma en, por el principio de substitución,

$$\frac{X + zY}{\sqrt{1+z^2}}.$$

Pero ahora z es simplemente una constante, de modo que la distribución condicional de W dado que $Z = z$ corresponde a una combinación lineal de las variables X e Y (condicionadas en z). Pero tanto X como Y son independientes de Z, de modo que $X|Z = z \sim N(0,1)$ e $Y|Z = z \sim N(0,1)$. Más aún, dado que $Z = z$, X e Y siguen siendo independientes (¿por qué?) por lo que se concluye que la distribución condicional mencionada es también normal (ver Ejemplo 4.6.2). Se tiene que

$$E(W|Z=z) = \frac{1}{\sqrt{1+z^2}}E(X|Z=z) + \frac{z}{\sqrt{1+z^2}}E(Y|Z=z) = 0,$$

y

$$Var(W|Z=z) = \frac{Var(X|Z=z)}{1+z^2} + \frac{z^2 Var(Y|z=z)}{1+z^2} = \frac{1+z^2}{1+z^2} = 1,$$

y entonces

$$W|Z=z \sim N(0,1).$$

Pero puesto que esta distribución condicional no depende de z, ella es también no condicional, y así, $W \sim N(0,1)$. Ahora bien,

$$\begin{aligned} Cov((X,W)|Z=z) &= \frac{Cov((X, X+zY)|Z=z)}{\sqrt{1+z^2}} \\ &= \frac{Cov((X,X)|Z=z)}{\sqrt{1+z^2}} + \frac{zCov((X,Y)|Z=z)}{\sqrt{1+z^2}} \\ &= \frac{Var(X|Z=z)}{\sqrt{1+z^2}} = \frac{1}{\sqrt{1+z^2}}, \end{aligned}$$

de modo que $Cov((X,W)|Z) = \frac{1}{\sqrt{1+z^2}}$. Por otra parte, $E(X|Z) = E(X) = 0$, y de (5.4.7) se tiene

$$Cov(X,W) = E\left(\frac{1}{\sqrt{1+Z^2}}\right),$$

lo cual coincide con $\rho(X,W)$, pues $Var(X) = Var(W) = 1$. Análogamente se obtiene

$$\rho(Y,W) = E\left(\frac{Z}{\sqrt{1+Z^2}}\right).$$

En el caso que $Z \sim U(0,1)$, se obtiene $\rho(X,W) = \operatorname{arcsinh}(1) \approx 0{,}8813$ y $\rho(Y,W) = \sqrt{2}-1$. Los detalles de estos últimos cálculos se proponen como ejercicio.

5.5. El Mejor Predictor

En esta sección retomamos el tema de predecir el valor de una variable o vector aleatorio, dado el valor de otra variable o vector aleatoria. En la Sección 4.8 abordamos este problema restringiéndonos a predictores lineales. Predicción lineal es atractiva por su simplicidad, pero muchas veces es posible encontrar mejores predictores, si uno no se limita solamente a aquellos que tienen forma lineal.

Consideremos el caso de dos vectores aleatorios $\boldsymbol{X} \in \mathbb{R}^k$ e $\boldsymbol{Y} \in \mathbb{R}^l$, y encontremos el *mejor predictor* (MP) de $\boldsymbol{X}$ dado $\boldsymbol{Y}$, es decir, hallar alguna función $g(\boldsymbol{Y})$ que minimice el *error cuadrático medio* de predicción

$$E\{(\boldsymbol{X} - g(\boldsymbol{Y}))'(\boldsymbol{X} - g(\boldsymbol{Y}))\}. \tag{5.5.1}$$

Para ello, usaremos el siguiente resultado preliminar.

Proposición 5.5.1 Sea $\boldsymbol{X} \in \mathbb{R}^k$ un vector aleatorio tal que $V(\boldsymbol{X})$ existe. Entonces, la solución del problema

$$\min_{\boldsymbol{c}\in\mathbb{R}^k} E\{(\boldsymbol{X} - \boldsymbol{c})'(\boldsymbol{X} - \boldsymbol{c})\} \tag{5.5.2}$$

es $\boldsymbol{c} = E(\boldsymbol{X})$.

Demostración: Sea $h(\boldsymbol{c}) = E\{(\boldsymbol{X} - \boldsymbol{c})'(\boldsymbol{X} - \boldsymbol{c})\}$. Entonces

$$h(\boldsymbol{c}) = E(\boldsymbol{X}'\boldsymbol{X} - 2\boldsymbol{c}'\boldsymbol{X} + \boldsymbol{c}'\boldsymbol{c}) = E(\boldsymbol{X}'\boldsymbol{X}) - 2\boldsymbol{c}'E(\boldsymbol{X}) + \boldsymbol{c}'\boldsymbol{c}.$$

Suponiendo $\boldsymbol{c} = (c_1, \ldots, c_k)'$, y diferenciando $h(\boldsymbol{c})$ con respecto a c_j e igualando a 0 se obtiene $-2E(X_j) + 2c_j = 0$, de donde $c_j = E(X_j)$ para $j = 1, \ldots, k$. Puesto que la matriz Hessiana de $h(\boldsymbol{c})$ es $2\boldsymbol{I}_k$, que es definida positiva, se concluye que $\boldsymbol{c} = E(\boldsymbol{X})$ es efectivamente el mínimo buscado. ■

Observe que el resultado de la Proposición 5.5.1 resuelve una versión restringida del problema que motiva esta sección, cual es la de hallar el mejor vector de constantes, predictor de $\boldsymbol{X}$, en el

sentido de resolver el problema 5.5.2. El error de predicción es, con $c = E(\boldsymbol{X})$,

$$\begin{aligned} E(\boldsymbol{X}'\boldsymbol{X}) - 2E(\boldsymbol{X})'E(\boldsymbol{X}) + E(\boldsymbol{X}'E(\boldsymbol{X})) &= E(\boldsymbol{X}'\boldsymbol{X}) - E(\boldsymbol{X})'E(\boldsymbol{X}) \\ &= E(\sum_{j=1}^{k} X_i^2) - \sum_{j=1}^{k} E(X_j)^2 \\ &= \sum_{j=1}^{k} Var(X_j). \end{aligned}$$

Volviendo al problema original, consideremos la cantidad a minimizar, dada por (5.5.2), entre todas las posibles funciones $g : \mathbb{R}^k \to \mathbb{R}^l$. Observe que por el Teorema 5.4.1

$$E\{(\boldsymbol{X} - g(\boldsymbol{Y}))'(\boldsymbol{X} - g(\boldsymbol{Y}))\} = E(E\{(\boldsymbol{X} - g(\boldsymbol{Y}))'(\boldsymbol{X} - g(\boldsymbol{Y}))|\boldsymbol{Y}\}).$$

Ahora, para minimizar

$$\begin{aligned} h(\boldsymbol{y}) &\stackrel{\text{def}}{=} E\{(\boldsymbol{X} - g(\boldsymbol{Y}))'(\boldsymbol{X} - g(\boldsymbol{Y}))|\boldsymbol{Y} = \boldsymbol{y}\} \\ &= E\{(\boldsymbol{X} - g(\boldsymbol{y}))'(\boldsymbol{X} - g(\boldsymbol{y}))|\boldsymbol{Y} = \boldsymbol{y}\}, \end{aligned}$$

la Proposición 5.5.1 establece que la función g elegida debe estar definida por $g^\star(\boldsymbol{y}) = E(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y})$, y por lo tanto, el MP es $g^\star(\boldsymbol{Y}) = E(\boldsymbol{X}|\boldsymbol{Y})$. En efecto, puesto que para cualquier función g, y para cualquier $\boldsymbol{y}$ se tiene

$$E\{(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}))'(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}))|\boldsymbol{Y} = \boldsymbol{y}\} \leq h(\boldsymbol{y}),$$

entonces

$$E\{(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y}))'(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y}))|\boldsymbol{Y}\} \leq h(\boldsymbol{Y}),$$

y tomando valor esperado a cada lado de esta última desigualdad se obtiene

$$E\{(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y}))'(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y}))\} \leq E\{(\boldsymbol{X} - g(\boldsymbol{Y}))'(\boldsymbol{X} - g(\boldsymbol{Y}))\},$$

cualquiera que se g.

Así, hemos deducido que el mejor predictor de $\boldsymbol{X}$ dado $\boldsymbol{Y}$, es simplemente

$$MP = E(\boldsymbol{X}|\boldsymbol{Y}). \tag{5.5.3}$$

Por otra parte, por (5.4.6) aplicado a cada elemento de las matrices en cuestión, se tiene que

$$V(MP) = V(\boldsymbol{X}) - E(V(\boldsymbol{X}|\boldsymbol{Y})), \tag{5.5.4}$$

que es una matriz al menos semi-definida positiva. Más aún, para cualquier vector de constantes $\boldsymbol{d} = (d_1, \ldots, d_k)$ se cumple que

$$Var(\boldsymbol{d}'E(\boldsymbol{X}|\boldsymbol{Y})) \leq Var(\boldsymbol{d}'\boldsymbol{X}).$$

lo que en particular muestra que cada coordenada $E(X_j|\boldsymbol{Y})$ del MP tiene siempre varianza inferior o igual a $Var(X_j)$, que corresponde al error de predecir X_j mediante la constante $E(X_j)$. Además, el error de predicción (5.5.1) está dado por

$$E\{(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y}))'(\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{Y}))\} = \sum_{j=1}^{k}\{Var(X_i) - Var(E(X_i|\boldsymbol{Y}))\}, \tag{5.5.5}$$

resultado cuya demostración se propone como ejercicio.

En algunos casos, como en los Ejemplos 5.2.1 y 5.4.3 el MP tiene forma lineal en la variable predictora. No es difícil convencerse que en este caso el MP y el MPL deben necesariamente coincidir. Sin embargo, esto no es la regla, puesto que el MPL usa sólo $E((\boldsymbol{X}, \boldsymbol{Y})')$ y $V((\boldsymbol{X}, \boldsymbol{Y})')$, mientras que el MP hace uso de la distribución condicional de $\boldsymbol{X}$ dado $\boldsymbol{Y}$, la cual, salvo excepciones, no queda siempre determinada por dichas cantidades.

Por último, en el caso que $\boldsymbol{X}$ e $\boldsymbol{Y}$ son independientes, se verifica la igualdad $E(\boldsymbol{X}|\boldsymbol{Y}) = E(\boldsymbol{X})$, y el MP se reduce simplemente a $E(\boldsymbol{X})$.

Ejemplo 5.5.1 Suponga que $X|Y = y \sim N(y, \tau^2)$, y que $Y \sim N(\mu, \sigma^2)$, donde μ, τ^2 y σ^2 son conocidos. Calcule el MP de Y dado X, y obtenga la varianza y error de predicción correspondientes.

Por las condiciones del problema, se tiene que

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{\tau^2\sigma^2}} \exp\left(-\frac{(x-y)^2}{2\tau^2} - \frac{(y-\mu)^2}{2\sigma^2}\right).$$

Note que el argumento de la función exponencial en esta densidad conjunta es una forma cuadrática, de modo que la distribución conjunta de (X, Y) es normal bivariada. Los parámetros de esta distribución se pueden obtener en forma similar a la del Ejemplo 4.7.2. Otra alternativa consiste simplemente en calcularlos directamente, como haremos a continuación. En primer lugar, se tiene que $E(Y) = \mu$, y $Var(Y) = \sigma^2$. Por otra parte,

$$E(X) = E(E(X|Y)) = E(Y) = \mu,$$
$$Var(X) = Var(E(X|Y)) + E(Var(X|Y)) = Var(Y) + E(\tau^2) = \sigma^2 + \tau^2,$$

y

$$\begin{aligned} E(XY) &= E(E(XY|Y)) = E(YE(X|Y)) = E(Y^2) \\ &= Var(Y) + E(Y)^2 = \sigma^2 + \mu^2, \end{aligned}$$

de modo que $Cov(X, Y) = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$, y por lo tanto $\rho(X, Y) = \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}$. Así, por lo hecho en los Ejemplos 5.3.4 y 5.4.3 el MP es

$$E(Y|X) = \mu - \left(\frac{\sigma^2}{\sigma^2 + \tau^2}\right)(\mu - X) = \frac{\sigma^2 X + \tau^2\mu}{\sigma^2 + \tau^2}.$$

La varianza del MP es $\sigma^4/(\sigma^2 + \tau^2)$, y el error de predicción es

$$Var(Y) - Var(E(Y|X)) = \sigma^2 - \frac{\sigma^4}{\sigma^2 + \tau^2} = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

Los detalles se proponen como ejercicio.

Ejemplo 5.5.2 Suponga que la vida útil T de una ampolleta es una variable aleatoria con distribución exponencial de media $\lambda > 0$. En el instante $t = 0$ la ampolleta se enciende. En un instante posterior $s > 0$ se constata que la ampolleta se había quemado. Calcule el MP del instante en que la ampolleta se quemó.

Necesitamos calcular $E(T|T < s)$, para lo cual se requiere la distribución condicional de T dado que $T < s$. Se tiene

$$\begin{aligned} P(T \leq t|T < s) &= \frac{P(T \leq t, T < s)}{P(T < s)} = \frac{P(T \leq \min\{t, s\})}{P(T < s)} \\ &= \frac{1 - e^{-\min\{t,s\}/\lambda}}{1 - e^{-s/\lambda}}, \end{aligned}$$

de donde, mediante diferenciación se obtiene

$$f_{T|T<s}(t) = \begin{cases} \frac{e^{-t/\lambda}}{\lambda(1-e^{-s/\lambda})} & \text{si } 0 < t < s \\ 0 & \text{si no.} \end{cases}$$

Así, el MP buscado es

$$E(T|T < s) = \int_0^s \frac{te^{-t/\lambda}}{\lambda(1 - e^{-s/\lambda})} dt = \lambda - \frac{e^{-s/\lambda}}{1 - e^{-s/\lambda}}.$$

Ejemplo 5.5.3 Suponga que dos ampolletas, cuyos tiempos de vida son independientes, con distribución exponencial de medias $\lambda > 0$ y $\mu > 0$ respectivamente, se ponen en funcionamiento simultáneamente. Se observa que la primera de ellas se quema en un instante $t > 0$. Calcule el MP de la vida útil de la otra ampolleta.

Si X e Y representan los tiempos de vida de estas ampolletas, se sabe que $X \sim \text{Exp}(\lambda)$ e $Y \sim \text{Exp}(\mu)$, y que X e Y son independientes. Lo que se observa es $U = \min\{X, Y\}$, y se quiere predecir $V = \max\{X, Y\}$, de modo que se necesita $E(V|U = t)$. Usando una modificación del argumento que lleva a concluir (4.4.9), se tiene para $u < v$:

$$\begin{aligned} P(U > u, V \leq v) &= P(u < X \leq v, u < Y \leq v) \\ &= P(u < X \leq v)P(u < Y \leq v) \\ &= (e^{-u/\lambda} - e^{-v/\lambda})(e^{-u/\mu} - e^{-v/\mu}). \end{aligned}$$

Por otra parte,

$$F_V(v) = P(X \leq v)P(Y \leq v) = (1 - e^{-v/\lambda})(1 - e^{-v/\mu}),$$

de modo que

$$\begin{aligned} F_{U,V}(u, v) &= F_V(v) - P(U > u, V \leq v) \\ &= (1 - e^{-v/\lambda})(1 - e^{-v/\mu}) - (e^{-u/\lambda} - e^{-v/\lambda})(e^{-u/\mu} - e^{-v/\mu}). \end{aligned}$$

Derivando parcialmente esta expresión con respecto a u y v, se obtiene la densidad conjunta

$$f_{U,V}(u,v) = \frac{e^{-v/\mu}e^{-u/\lambda} + e^{-u/\mu}e^{-v/\lambda}}{\lambda\mu},$$

definida en la región $0 < u < v < \infty$. Por otra parte,

$$\begin{aligned} f_U(u) &= \int_u^\infty \frac{(e^{-v/\mu}e^{-u/\lambda} + e^{-u/\mu}e^{-v/\lambda}}{\lambda\mu} du \\ &= \left(\frac{1}{\lambda} + \frac{1}{\mu}\right) e^{-u(1/\lambda+1/\mu)}, \end{aligned}$$

para $u > 0$. Note que $U \sim \text{Exp}((\lambda^{-1} + \mu^{-1})^{-1})$. Luego, la densidad condicional de V dado que $U = u$ es, después de simplificar,

$$f_{V|U=u}(v|u) = \frac{f_{U,V}(u,v)}{f_U(u)} = \frac{e^{-(v-u)/\lambda} + e^{-(v-u)/\lambda}}{\lambda + \mu},$$

para $v > u$, de donde

$$E(V|U=u) = \int_u^\infty v f_{V|U=u}(v|u)dv = u + \frac{\lambda^2 + \mu^2}{\lambda + \mu},$$

de modo que el MP buscado es $s + (\lambda^2 + \mu^2)/(\lambda + \mu)$.

5.6. Problemas

1. Sea X una variable aleatoria con distribución de Bernoulli con parámetro p. Si $E(Y|X = 0) = 1$ y $E(Y|X = 1) = 2$, encuentre $E(Y)$.

2. Sea N una variable aleatoria discreta positiva de media μ, y suponga que $X_1, X_2, ...$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con $E(X_1) = m$. Si N es independiente de las variables aleatorias X_i, pruebe que:

$$E(X_1 + X_2 + ... + X_N) = \mu.$$

3. Suponga que el número de personas que entran a un supermercado el día Lunes es una variable aleatoria de media 50. Suponga además que los montos de dinero gastado por los clientes en el supermercado son variables aleatorias independientes de media común 8. Si dichos montos son independientes del número total de clientes que entran al supermercado, ¿cuál es el monto esperado de dinero gastado en la tienda ese día?.

4. Un dado insesgado es sucesivamente arrojado. Sean X e Y variables aleatorias que denotan el número de lanzamientos necesarios para obtener un 6 y un 5 respectivamente. Encontrar

   a.- $E(X)$.
   b.- $E(X|Y = 1)$.
   c.- $E(X|Y = 5)$.

5. Una población de individuos da lugar a una nueva población. Suponga que la probabilidad que un individuo de lugar a k individuos (descendientes) es p_k, $k = 0, 1, \ldots$, y el número de individuos que se obtienen a partir de individuos diferentes son variables aleatorias independientes. La población nueva forma la nueva generación, que a su vez, da lugar a la segunda generación, y así sucesivamente. Para $n = 0, 1, \ldots$ sea X_n el tama no de la n-ésima generación. Nótese que:

$$X_{n+1} = Z_1(n) + \cdots + Z_{X_n}(n),$$

   donde $Z_j(n)$ es el número de individuos de la generación $(n + 1)$-ésima que proceden del individuo j-ésimo de la generación n-ésima. Suponga que el número de descendientes de un individuo tiene media finita μ. Pruebe que:

$$M_n = E(X_n|X_0 = 1) = \mu^n.$$

6. Una urna contiene 4 bolas blancas y 6 bolas negras. Se sacan, en forma consecutiva y sin reemplazo, dos muestras aleatorias, de tama nos 3 y 5 respectivamente. Sean X e Y variables aleatorias que denotan el número de bolas blancas en las dos muestras. Calcule $E(X|Y = i)$ para $i = 1, 2, 3, 4$.

7. Sean X_1, X_2 variables aleatorias independientes e idénticamente distribuidas $N(0, 1)$. Sea U independiente de X_1 y X_2, y suponga que U distribuye uniforme en $[0, 1]$. Definamos $Z = UX_1 + (1 - U)X_2$.

   a.- Encuentre la distribución condicional de Z dado que $U = u$.

b.- Encuentre $E(Z)$ y $Var(Z)$.

c.- Encontrar la distribución de Z.

8. La siguiente tabla nos da la distribución conjunta de X e Y:

x/ y	1	2	3
1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
2	0	$\frac{1}{4}$	$\frac{1}{12}$
3	0	0	$\frac{1}{6}$

a.- Encontrar F_x, F_y, $F_{y|x}$, $F_{x|y}$.

b.- Encontrar $E(Y|X)$, $E(X|Y)$, $Cov(X,Y)$.

9. La densidad conjunta entre X e Y esta dada por :

$$f_{X,Y}(x,y) = \frac{e^{-\frac{x}{y}}e^{-y}}{y}, \qquad 0 \leq x \leq \infty, \;\; 0 \leq y \leq \infty$$

Encuentre $E(X^2|Y=y)$.

10. Sea (X,Y) con distribución uniforme entre las rectas $x+y=1$, $y=0$, y la curva $y=x^2$. Determine $f_X(x)$, $f_Y(y)$, $f_{X|Y}(x|y)$, $f_{Y|X}(y|x)$, y verifique que son densidades.

11. Suponga que $X|Y=y \sim N(y,1)$ e $Y \sim N(0,1)$.

a.- Calcule $E(X)$ y $Var(X)$.

b.- Calcule $\rho(X,Y)$

12. Sean X e Y independientes con $X \sim$ Geom(p), $Y \sim$ Poisson(λ), y $Z = X+Y$. Calcule $E(X|Y)$ y $E(Y|Z)$.

13. Sean $V|T=t \sim U(0,t)$ y T con densidad

$$f_T(t) = (r-1)t^{-r} \qquad 0<t<1; \qquad r \leq 1.$$

a.- Determine $f_V(v)$.

b.- Determine $f_{T|V=v}(t|v)$.

c.- Determine $E(T|V=v)$.

14. Sea $X \sim N(0,1)$ e $Y|X=x \sim N(\alpha x, 1-\alpha^2)$, para $0<\alpha<1$. Encontrar $E(Y)$.

15. Sean X e Y i.i.d. con distribución $N(0,\sigma^2)$, y sea $Z=\sqrt{X^2+Y^2}$. Obtenga las distribuciones condicionales $X|Z=z$ e $Y|Z=z$, y pruebe que $E(X|Z)=E(Y|Z)=0$.

16. Si (X, Y) tiene función densidad dada por:

$$f_{X,Y}(x, y) = \frac{e^{-y}}{y} \qquad 0 \leq x \leq y, \;\; 0 \leq y \leq \infty,$$

determine $E(X^3|Y = y)$.

17. Una cierta lámpara tiene una vida útil en *horas* cuya distribución es exponencial de media 1. Una persona enciende dicha lámpara y comienza a lanzar un dado equilibrado cada 15 segundos, continuando de esta manera mientras la lámpara esté encendida. Obtenga el valor esperado y la varianza del número de ases que se obtiene antes que la lámpara se apague.

18. Se tiene dos lámparas cuyas vidas útiles son variables aleatorias i.i.d. con distribución exponencial de media $\lambda > 0$. Suponiendo que ambas lámparas se encienden simultáneamente, denote por X el tiempo que transcurre hasta que la primera lámpara se apague, e Y el tiempo transcurrido hasta que la segunda lámpara se apague (note que $X \leq Y$).

    (a) Obtenga las distribuciones condicionales de Y dado que $X = x$, y de X dado que $Y = y$.

    (b) Calcule la esperanza y varianza condicional de cada una de las distribuciones en (a).

19. Suponga que el número esperado de accidentes por semana en una planta industrial es 5. Suponga también que el número de trabajadores heridos en cada accidente son variables aleatorias independientes con media común de 2.5. Si el número de trabajadores heridos en cada accidente es independiente del número de accidentes que ocurren, calcule el número esperado de trabajadores heridos.

    $Resp$: $12,5$

20. Se dispone de dos urnas A y B, la primera contiene tres bolas rojas y dos bolas negras, la segunda contiene tres bolas negras y dos bolas rojas. Se realiza el siguiente experimento :

    (a) Se escoge al azar un número entre 1 y 5, se saca igual cantidad de bolas de A y se introducen en B.

    (b) Si en (a) se transfirieron x bolas, entonces se escoge al azar un número entre x y $5 + x$, y se saca igual cantidad de bolas de B, las que se introducen en A. Sean X e Y los números respectivos de bolas transferidas en (a) y (b).

    Calcule $E(X)$, $E(Y)$, $\rho(X, Y)$. ¿Cuál es la probabilidad que la configuración de bolas al terminar el experimento coincida con la inicial?

21. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d. con función de distribución continua F. Sea $X = \text{máx}\{X_1, \ldots, X_n\}$.

    (a) Demuestre que para todo $k = 1, 2, \ldots, n$ se tiene que

$$P(X_k \leq x|X = t) = \begin{cases} \frac{(n-1)F(x)}{nF(t)} & \text{si } x < t \\ 1 & \text{si } x \geq t \end{cases}$$

(b) Suponga que F es diferenciable. ¿Existe densidad condicional en (a)? ¿Por qué?

(c) En el caso que F es la distribución $U(0,1)$, calcule $E(X_k|X)$ para $k = 1, 2, \ldots, n$.

22. Sean X_1, X_2, X_3 tres puntos escogidos en forma independiente y al azar en el intervalo $[0,1]$. Obtenga $E(X_{(1)}|X_{(2)}, X_{(3)})$, $E(X_{(2)}|X_{(1)}, X_{(3)})$ y $E(X_{(3)}|X_{(1)}, X_{(2)})$, en donde $X_{(1)}, X_{(2)}, X_{(3)}$ son los estadísticos de orden correspondientes.

23. Suponga que X_1 y X_2 tienen distribución conjunta normal bivariada tal que $E(X_1|X_2) = 3{,}7 - 0{,}15X_2$, $E(X_2|X_1) = 0{,}4 - 0{,}6X_1$ y $Var(X_2|X_1) = 3{,}64$. Determine la media y la varianza de X_1, la media y la varianza de X_2, y la correlación entre X_1 y X_2.

24. Sean X, Y con distribución conjunta normal bivariada, y con

$$Q(x,y) \quad = \quad x^2 + 2y^2 - xy - 3x - 2y + 4$$

a.- Escriba la densidad conjunta de X e Y.

b.- Obtenga $E(\boldsymbol{Z})$ y $V(\boldsymbol{Z})$ si $\boldsymbol{Z} = (X,Y)'$.

c.- Determine las densidades marginales de X e Y.

d.- Obtenga $P(X < 3|Y = 2)$, y determine además E$(X|Y = y)$ y $Var(X|Y = y)$.

25. Si el mejor predictor de X dado Y coincide con $E(X)$, ¿es necesariamente cierto que X e Y son independientes?

$Hint$: Considere (X,Y) con distribución uniforme en el círculo $\{(x,y) : x^2 + y^2 \leq 1\}$.

26. Pruebe que si X e Y son independientes y X posee densidad $f_X(x)$ entonces

$$P(X < Y) = \int_{-\infty}^{\infty} (1 - F_Y(x)) f_X(x) dx.$$

Aplique lo anterior al caso en que X tiene distribución exponencial con parámetro λ e $Y \sim U(0,\lambda)$, donde $\lambda > 0$.

27. Sean X e Y i.i.d. $U(0,1)$, y defina $U = \text{mín}\{X,Y\}$ y $V = \text{máx}\{X,Y\}$.

(a) Obtenga la densidad condicional de U dado que $V = v$, y la densidad condicional de V dado que $U = u$.

(b) Calcule $E(U|V)$ y $E(V|U)$.

28. Si X e Y son no correlacionadas, ¿es necesariamente cierto que el MP coincide con $E(X)$?

$Hint$: Considere $Y \sim U(-1,1)$ y $X = Y^2$.

29. Suponga que $X|Z = z \sim$ Poisson(z) y que $Z \sim \Gamma(\alpha, 1)$, con densidad

$$f_Z(z) = \begin{cases} \frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z} & \text{si } z > 0 \\ 0 & \text{si no,} \end{cases}$$

y en donde $\alpha > 0$.

(a) Demuestre que para $k = 0, 1, 2, \ldots$ se tiene

$$P(X = k) = \frac{\Gamma(k+\alpha)}{\Gamma(a)\Gamma(k+1)} \left(\frac{1}{2}\right)^{k+\alpha}.$$

(b) Usando métodos probabilísticos muestre que

$$\sum_{k=1}^{\infty} \binom{k+n-1}{n} \frac{1}{2^k} = 2^n \qquad \text{para } n = 1, 2, \ldots$$

(*Indicación:* Calcule $E(X)$ de dos maneras distintas.)

30. Se escoge al azar un número en el intervalo $[0, 1]$. Si el resultado es x, se procede a lanzar n veces y en forma independiente una moneda cuya probabilidad de dar cara es x. Sea Y la variable aleatoria que representa el número de caras que se obtuvo al cabo de los n lanzamientos.

(a) Calcule $E(Y)$ y $Var(Y)$ sin calcular previamente la distribución de Y.

(b) Repita (a) usando ahora la distribución de Y.

Capítulo 6

Nociones de Convergencia y sus Aplicaciones

6.1. Motivación

Supongamos una moneda honesta se lanza repetidamente y en forma independiente. De acuerdo a la *interpretación frecuentista* de la probabilidad introducida en el Capítulo 1, la frecuencia relativa del número de caras (esto es, la proporción de veces que se obtuvo cara), debe oscilar en torno a $1/2$, y de hecho, converge a este valor. Es decir, si

$$f_n = \frac{\text{Número de caras en los primeros } n \text{ ensayos}}{n},$$

entonces $\lim_{n\to\infty} f_n = 1/2$. Sin embargo, no hemos precisado en qué sentido dicha convergencia ha de entenderse. De partida, note que f_n se puede reescribir de la siguiente manera:

$$f_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

donde $X_i = 1$ si se obtuvo cara en el i-ésimo lanzamiento, y 0 si no. Así, f_n es una variable aleatoria, de modo que se debe definir alguna noción de convergencia para variables aleatorias. Surgen varias alternativas que iremos revisando en este Capítulo. Por ejemplo, y recordando que las variables aleatorias son funciones a valores reales definidas en un cierto espacio muestral, es posible considerar nociones de convergencia para una sucesión de funciones. Por otra parte, es también posible apelar al aspecto probabilístico de dichas variables, y así definir nociones de convergencia que utilicen su distribución.

El aspecto formal del tratamiento de las nociones que definiremos involucra usualmente un alto nivel de sofisticación teórica que no será cubierta en este texto, de modo que nos centraremos más en las aplicaciones.

6.2. Definición de Nociones de Convergencia

Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias, definidas en un espacio muestral Ω común. La sucesión se denotará usualmente por $\{X_n\}$. Se definen a continuación 4 tipos distintos de convergencia.

Definición 6.2.1

(a) Se dice que $\{X_n\}$ converge *en distribución* a una variable aleatoria X, lo que se denota $X_n \xrightarrow{D} X$, si

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \tag{6.2.1}$$

para todo $x \in \mathbb{R}$ tal que x es un punto de continuidad de F_X.

(b) Se dice que $\{X_n\}$ converge *en probabilidad* a una variable aleatoria X, lo que se denota $X_n \xrightarrow{P} X$, si

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0. \tag{6.2.2}$$

(c) Se dice que $\{X_n\}$ converge *en media cuadrática* a una variable aleatoria X, lo que se denota $X_n \xrightarrow{\text{m.c.}} X$, si

$$\lim_{n\to\infty} E\{(X_n - X)^2\} = 0. \tag{6.2.3}$$

(d) Se dice que $\{X_n\}$ converge *casi seguramente* a una variable aleatoria X, lo que se denota $X_n \xrightarrow{\text{c.s.}} X$, si

$$P(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1. \tag{6.2.4}$$

Estos tipos de convergencia se refieren a comportamiento asintótico de la sucesión $\{X_n\}$, pero en aspectos esencialmente diferentes. Así, la convergencia en distribución (6.2.1) usa solamente la función de distribución de las variables aleatorias. Puesto que no existe una identificación entre una variable aleatoria y su distribución (por ejemplo, si $X \sim N(0,1)$ entonces $-X \sim N(0,1)$, pero $X \neq -X$), este tipo de convergencia no usa los valores de las variables en cuestión, si no que las probabilidades asociadas. El hecho que la convergencia de F_{X_n} a F_X se requiera sólo para aquellos puntos en que F_X es continua, obedece a razones técnicas.

En el otro extremo, la convergencia casi segura (6.2.4), también llamada *convergencia con probabilidad 1*, trata las variables aleatorias como funciones, y requiere que exista convergencia puntual en un conjunto de puntos del espacio muestral cuya probabilidad es 1. En otras palabras, la convergencia puntual no se cumple en un conjunto que, desde el punto de vista probabilístico, se puede despreciar.

La convergencia en probabilidad (6.2.2) y en media cuadrática (6.2.3) representan situaciones intermedias, en que ambas, las variables aleatorias y su distribución se combinan. La convergencia en probabilidad requiere que la probabilidad que un elemento genérico de la sucesión difiera del límite en una cantidad arbitrariamente peque na converja a 0. Por otra parte, la convergencia en media cuadrática requiere que el error cuadrático medio de predecir la variable límite por un elemento de la sucesión, sea asintóticamente 0.

Veremos ahora un resultado fundamental concerniente a las relaciones que existen entre estos modos de convergencia.

Teorema 6.2.1 (Relación entre los Modos de Convergencia) Sean $X, X_1, X_2, \ldots$ variables aleatorias.

(a) Si $X_n \xrightarrow{P} X$, entonces $X_n \xrightarrow{D} X$.

(b) Si $X_n \xrightarrow{\text{m.c.}} X$, entonces $X_n \xrightarrow{P} X$.

(c) Si $X_n \xrightarrow{\text{c.s.}} X$, entonces $X_n \xrightarrow{P} X$.

La demostración de este resultado será omitida, por ser de caracter esencialmente técnico. Sin embargo, y como veremos en los ejemplos que siguen, las recíprocas de estos resultados son, en general, falsas. Por otra parte, la convergencia en distribución suele recibir el nombre alternativo de *convergencia débil*, pues es implicada por todos los otros tipos de convergencia. En forma análoga, la convergencia casi segura, suele también recibir el nombre de *convergencia fuerte*.

Ejemplo 6.2.1 Sea $X_n \sim \text{Exp}(\lambda_n)$, donde $\{\lambda_n\}$ es una sucesión de números positivos tales que $\lim_{n\to\infty} \lambda_n = 0$. Observe que $E(X_n) = \lambda_n$, de modo que se intuye que el límite, en caso de existir, debe ser 0. Veamos en qué sentido se produce esta potencial convergencia a 0.

Note que $F_{X_n}(x) = 1 - e^{-x/\lambda_n}$ si $x > 0$, y 0 si no, de modo que para $x > 0$ se tiene

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} (1 - e^{-x/\lambda_n}) = 1,$$

y $\lim_{n\to\infty} F_{X_n}(x) = 0$, si $x < 0$. Si $X = 0$, se tiene entonces que $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$, para $x \neq 0$. El caso $x = 0$ es irrelevante, pues es precisamente el único punto de discontinuidad de F_X. Luego, $X_n \xrightarrow{D} X$. Por otra parte, observe que para $\epsilon > 0$ se tiene

$$P(|X_n - X| > \epsilon) = P(X_n > \epsilon) = e^{-\epsilon/\lambda_n} \to 0,$$

si $n \to \infty$, de modo que se concluye también que $X_n \xrightarrow{P} X$. En estricto rigor, este último resultado implica la convergencia en distribución, pero es ilustrativo, ocasionalmente, mostrar algunas propiedades en forma directa. Pero eso no es todo. Note que

$$E\{(X_n - X)^2\} = E(X_n^2) = 2\lambda_n^2 \to 0,$$

si $n \to \infty$, así que además se cumple que $X_n \xrightarrow{\text{m.c.}} 0$.

Ejemplo 6.2.2 Sean $X, X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} N(0, 1/2)$. Entonces, dado que $F_{X_n}(x) = F_X(x)$ para cualquier x, se cumple en forma trivial que $X_n \xrightarrow{D} X$. Sin embargo, note que $X_n - X \sim N(0, 1)$, de modo que

$$P(|X_n - X| > \epsilon) = 2(1 - \Phi(\epsilon)), \quad n \geq 1,$$

por lo que no hay convergencia en probabilidad, y en virtud del Teorema 6.2.1, tampoco puede haber convergencia en media cuadrática o casi segura.

Ejemplo 6.2.3 Sea $Y \sim U(0,1)$, y defina para $m = 0, 1, 2, \ldots$ e $i = 0, 1, \ldots, 2^m - 1$ los intervalos $I_{2^m+i} = [i/2^m, (i+1)/2^m]$, y las variables aleatorias $X_1, X_2, \ldots$

$$X_n = \begin{cases} 1 & \text{si } Y \in I_n \\ 0 & \text{si no} \end{cases}$$

Así, los intervalos I_n van en forma cíclica cubriendo el intervalo $[0,1]$. Es claro que para cualquier $\omega \in \Omega$, hay una infinidad de valores de n tales que $X_n(\omega) = 1$, de modo que puntualmente, $X_n(\omega)$ no converge a valor alguno. Sin embargo, si $X = 0$, y $0 < \epsilon \leq 1$, entonces

$$P(|X_n - X| > \epsilon) = P(X_n > \epsilon) = P(Y \in I_n) = \text{largo de } I_n,$$

que converge a 0 cuando $n \to \infty$, de modo que $X_n \xrightarrow{P} X$. En forma similar, se prueba que $E(X_n^2) \to 0$ cuando $n \to \infty$, de modo que hay convergencia en media cuadrática (y en probabilidad), pero no casi segura.

Ejemplo 6.2.4 Sea X con densidad

$$f_X(x) = \begin{cases} 2x^{-3} & \text{si } x > 1 \\ 0 & \text{si no,} \end{cases}$$

y sea $I_n = [1, n+1]$, para $n \geq 1$. Defina $X_n(\omega) = X(\omega)$ si $X(\omega) \in I_n$, y $X_n(\omega) = 0$ si no. Es claro que para cualquier $\omega \in \Omega$ se cumple $\lim_{n\to\infty} X_n(\omega) = X(\omega)$, de modo que $X_n \xrightarrow{\text{c.s.}} X$, y en particular, $X_n \xrightarrow{P} X$. Por otra parte, note que

$$X_n - X = \begin{cases} 0 & \text{si } 1 < X \leq n+1 \\ X & \text{si no,} \end{cases}$$

de modo que

$$E\{(X_n - X)^2\} = \int_{n+1}^{\infty} \frac{2x^2}{x^3} dx = \int_{n+1}^{\infty} \frac{2}{x} dx = \infty,$$

y entonces no existe convergencia en media cuadrática.

A pesar de lo evidenciado en estos ejemplos, hay un caso particular en que convergencia en distribución implica convergencia en probabilidad, como lo muestra el siguiente resultado.

Proposición 6.2.1 Si $\{X_n\}$ es una sucesión de variables aleatorias tales que $X_n \xrightarrow{D} c$, una variable aleatoria constante, entonces $X_n \xrightarrow{P} c$.

Demostración: Puesto que $X_n \xrightarrow{D} c$, se tiene entonces que para $x \neq c$ se cumple

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 1 & \text{si } x > c \\ 0 & \text{si } x < c. \end{cases}$$

Por otra parte, dado $\epsilon > 0$ se tiene

$$\begin{aligned} P(|X_n - c| \leq \epsilon) &= P(c - \epsilon \leq X_n \leq c + \epsilon) \geq P(c - \epsilon < X_n \leq c + \epsilon) \\ &= F_{X_n}(c + \epsilon) - F_{X_n}(c - \epsilon) \to 1 - 0 = 1, \end{aligned}$$

cuando $n \to \infty$, de donde se concluye que

$$P(|X_n - c| > \epsilon) = 1 - P(|X_n - c| \leq \epsilon) \to 0,$$

cuando $n \to \infty$, y entonces $X_n \xrightarrow{P} X$. ■

En el caso particular de variables aleatorias discretas, se tiene la siguiente caracterización de la convergencia en distribución.

Proposición 6.2.2 Sean $X, X_1, X_2, \ldots$ variables aleatorias discretas con valores en $0, 1, 2, \ldots$. Entonces $X_n \xrightarrow{D} X$ si y sólo si $\lim_{n\to\infty} p_{X_n}(k) = p_X(k)$ para todo $k = 0, 1, 2, \ldots$.

Demostración: Se propone como ejercicio. ■

> **Ejemplo 6.2.5** Si $X_n \sim \text{Bin}(n, p_n)$, donde $\lim_{n\to\infty} np_n = \lambda > 0$, y $\lim_{n\to\infty} p_n = 0$, entonces el desarrollo que conduce a (3.10.12) muestra que $X_n \xrightarrow{D} X$, donde $X \sim \text{Poisson}(\lambda)$.

Como ya hemos visto en capítulos anteriores, se puede construir nuevas variables aleatorias mediante transformaciones de aquellas disponibles. Esto es, si X es una variable aleatoria, y si $g : \mathbb{R} \longrightarrow \mathbb{R}$ es una función continua, entonces $g(X)$ es una variable aleatoria. Surge entonces la siguiente pregunta. Si $\{X_n\}$ converge a X en algún sentido, ¿es cierto que $\{g(X_n)\}$ converge a $g(X)$ en ese (u otro) sentido? La respuesta está dada por el siguiente resultado.

Proposición 6.2.3 Sean $X, X_1, X_2, \ldots$ variables aleatorias, y sea $g : \mathbb{R} \longrightarrow \mathbb{R}$ una función continua.

(a) Si $X_n \xrightarrow{D} X$ entonces $g(X_n) \xrightarrow{D} g(X)$.

(b) Si $X_n \xrightarrow{P} X$ entonces $g(X_n) \xrightarrow{P} g(X)$.

(c) Si $X_n \xrightarrow{\text{c.s.}} X$ entonces $g(X_n) \xrightarrow{\text{c.s.}} g(X)$.

Observe que de acuerdo al resultado de la Proposición 6.2.3, la convergencia de sucesiones de variables aleatorias no se altera debido a transformaciones continuas, *excepto en el caso de la convergencia en media cuadrática*. La razón que esto no funcione en dicho caso es fácil de ver mediante el siguiente contraejemplo, que es una ligera variación del Ejemplo 6.2.4.

Ejemplo 6.2.6 Considere

$$f_X(x) = \begin{cases} 3x^{-4} & \text{si } x > 1 \\ 0 & \text{si no,} \end{cases}$$

y sea $I_n = [1, n+1]$, para $n \geq 1$. Defina $X_n(\omega) = X(\omega)$ si $X(\omega) \in I_n$, y $X_n(\omega) = 0$ si no. El mismo tipo de razonamiento del Ejemplo 6.2.4 permite concluir que $X_n \xrightarrow{\text{c.s.}} X$, de modo que si $g(x) = x^2$, que es una función continua, entonces se obtiene que $X_n^2 \xrightarrow{\text{c.s.}} X^2$. Por otra parte, observe que

$$E\{(X_n - X)^2\} = \int_{n+1}^{\infty} \frac{3x^2}{x^4} dx = \frac{3}{n+1} \to 0$$

si $n \to \infty$, de donde $X_n \xrightarrow{\text{m.c.}} X$. Sin embargo se puede comprobar que $E(\{(X_n^2 - X^2)^2\}) = \infty$ para cualquier n, por lo que no existe convergencia en media cuadrática para $g(X_n)$.

Ejemplo 6.2.7 Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con distribución común $U(0,1)$, y sea $Y_n = \text{mín}\{X_1, \ldots, X_n\}$. Por (4.4.7), se tiene que

$$f_{Y_n}(y) = n(1-y)^{n-1}, \qquad 0 < y < 1.$$

Así, $Y_n \sim \text{Beta}(1, n)$, por lo que $E(Y_n) = 1/(n+1)$, de modo que se sospecha que en caso de existir el límite de Y_n, éste debiera ser 0. Dado $0 < \epsilon < 1$, se tiene que

$$P(Y_n > \epsilon) = \int_{\epsilon}^{1} n(1-y)^{n-1} dy = (1-\epsilon)^n,$$

y tomando límite cuando $n \to \infty$ se concluye que $\lim_{n\to\infty} P(Y_n > \epsilon) = 0$. Puesto que si $\epsilon > 1$ se tiene $P(Y_n > \epsilon) = 0$, hemos mostrado que $Y_n \xrightarrow{P} 0$. Consideremos ahora $Z_n = nY_n$. Ya no es cierto que exista la misma convergencia anterior, pues ahora $E(Z_n) = n/(n+1) \to 1$ si $n \to \infty$. Veremos que Z_n converge en distribución a una variable aleatoria $Z \sim \text{Exp}(1)$. Para ello, consideremos $F_{Z_n}(z)$. Se tiene, para $z > 0$:

$$\begin{aligned} F_{Z_n}(z) &= P(Z_n \leq z) = P(nY_n \leq z) = P(Y_n \leq z/n) \\ &= 1 - (1 - z/n)^n \longrightarrow 1 - e^{-z} \end{aligned}$$

cuando $n \to \infty$, de modo que $Z_n \xrightarrow{D} Z \sim \text{Exp}(1)$.

Veamos a continuación cómo se extienden estas nociones de convergencia al caso de vectores aleatorios. Para ello, recordamos la definición de la *norma euclidiana* de un vector $\boldsymbol{v} = (v_1, \ldots, v_k) \in \mathbb{R}^k$:

$$||\boldsymbol{v}|| = \sqrt{v_1^2 + v_2^2 + \cdots + v_k^2}.$$

Definición 6.2.2 Sean $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ vectores aleatorios en $\mathbb{R}^k$, donde $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nk})$ y $\boldsymbol{X} = (X_1, \ldots, X_k)$.

(a) Se dice que $\boldsymbol{X}_n$ converge en probabilidad a $\boldsymbol{X}$ si $||\boldsymbol{X}_n - \boldsymbol{X}|| \xrightarrow{P} 0$, es decir, si para cualquier $\epsilon > 0$ se cumple

$$\lim_{n\to\infty} P(||\boldsymbol{X}_n - \boldsymbol{X}|| > \epsilon) = 0. \tag{6.2.5}$$

(b) Se dice que $\boldsymbol{X}_n$ converge en media cuadrática a $\boldsymbol{X}$ si $||\boldsymbol{X}_n - \boldsymbol{X}|| \xrightarrow{\text{m.c.}} 0$, es decir, si

$$\lim_{n\to\infty} E\{||\boldsymbol{X}_n - \boldsymbol{X}||^2\} = 0. \tag{6.2.6}$$

(c) Se dice que $\boldsymbol{X}_n$ converge casi seguramente a $\boldsymbol{X}$ si

$$P(\{\omega \in \Omega : \lim_{n\to\infty} X_{ni}(\omega) = X_i(\omega),\ i = 1, \ldots, k\}) = 1. \tag{6.2.7}$$

Observación: Hemos diferido la discusión de la noción de convergencia en distribución de vectores aleatorios para una sección posterior, debido a varias complicaciones técnicas que van más allá del ámbito de este libro. La Sección 6.4 discute este tema, dando una caracterización muy útil y que permite evitar dichos problemas.

El siguiente resultado es útil para chequear convergencia en probabilidad, media cuadrática y casi segura de vectores aleatorios.

Proposición 6.2.4 Sean $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ vectores aleatorios en $\mathbb{R}^k$, donde $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nk})$ y con $\boldsymbol{X} = (X_1, \ldots, X_k)$. Entonces

(a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ si y sólo si $X_{ni} \xrightarrow{P} X_i$ para $i = 1, \ldots, k$.

(b) $\boldsymbol{X}_n \xrightarrow{\text{m.c.}} \boldsymbol{X}$ si y sólo si $X_{ni} \xrightarrow{\text{m.c.}} X_i$ para $i = 1, \ldots, k$.

(c) $\boldsymbol{X}_n \xrightarrow{\text{c.s.}} \boldsymbol{X}$ si y sólo si $X_{ni} \xrightarrow{\text{c.s.}} X_i$ para $i = 1, \ldots, k$.

La Proposición 6.2.4 establece que para verificar los tipos de convergencia de vectores aleatorios, basta con mostrar que cada coordenada (que es una variable aleatoria), converge a la correspondiente coordenada del vector límite, y de acuerdo al tipo de convergencia adecuado.

La próxima sección retoma la idea planteada al comienzo de este capítulo, esta vez dándole un sentido formal.

6.3. Leyes de Grandes Números

En términos intuitivos, las leyes de grandes números (LGN) establecen que si $X_1, X_2, \ldots$ constituyen una *muestra aleatoria* de una cierta distribución F (esto es, $X_1, X_2, \ldots$ son i.i.d. con distribución común F), y si dicha distribución posee valor esperado μ, entonces

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu,$$

en un sentido a especificar. Así, la LGN establece que la secuencia de medias aritméticas de las primeras n variables converge a la esperanza de la distribución, que ciertamente coincide con la esperanza de cualquiera de las variables en cuestión.

Hay dos tipos de LGN que estudiaremos aquí: la ley débil (LDGN), y la ley fuerte (LFGN), que establecen resultados de convergencia en probabilidad y casi segura, respectivamente. Comenzamos esta discusión con el primer caso. Para ello, necesitamos un resultado previo.

Proposición 6.3.1 (Desigualdad de Tchebyshev) Considere un real $\alpha > 0$, y una variable aleatoria X.

(a) Si X es no negativa (esto es, $P(X \geq 0) = 1$), y si $E(X)$ es finita, entonces

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}. \tag{6.3.1}$$

(b) Si $\text{Var}(X)$ es finita, entonces

$$P(|X - E(X)| \geq \alpha) \leq \frac{\text{Var}(X)}{\alpha^2}. \tag{6.3.2}$$

Demostración:

(a) Observe que

$$\alpha I\{X \geq \alpha\} \leq X I\{X \geq \alpha\} \leq X,$$

y tomando valor esperado a cada lado de la desigualdad se obtiene

$$\alpha P(X \geq \alpha) \leq E(X I\{X \geq \alpha\}) \leq E(X),$$

de donde $\alpha P(X \geq \alpha) \leq E(X)$, lo que prueba el resultado.

(b) Note que

$$P(|X - E(X)| \geq \alpha) = P((X - E(X))^2 \geq \alpha^2),$$

y el resultado se obtiene de aplicar (a) al lado derecho de esta última igualdad. ■

La aplicación fundamental de la Proposición 6.3.1 es el siguiente resultado.

Proposición 6.3.2 (Ley débil de Tchebyshev) Sean $X_1, X_2, \ldots$ variables aleatorias no correlacionadas (lo que significa $\text{Cov}(X_i, X_j) = 0$ si $i \neq j$), con varianzas finitas y tales que existe un número $M > 0$ tal que $\text{Var}(X_n) \leq M$ para todo $n \geq 1$. Si $S_n = \sum_{i=1}^{n} X_i$, entonces se cumple que

$$\frac{S_n - E(S_n)}{n} \xrightarrow{P} 0.$$

Demostración: Puesto que las variables son no correlacionadas, se tiene que

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \leq nM.$$

Luego, por la desigualdad de Tchebyshev (6.3.2) se tiene que para cualquier $\epsilon > 0$:

$$P(|S_n - E(S_n)| \geq n\epsilon) \leq \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} \leq \frac{M}{n\epsilon^2} \longrightarrow 0,$$

lo que prueba el resultado. ■

Ejemplo 6.3.1 (Ley de Grandes Números de Bernoulli)
Considere un proceso de Bernoulli $X_1, X_2, \ldots$, con probabilidad de éxito p. Se tiene que $S_n \sim \mathrm{Bin}(n, p)$, de modo que $E(S_n) = np$. Además, $\mathrm{Var}(X_n) = p(1-p)$, de modo que tomando $M = p(1-p)$, y considerando que las variables son independientes (en particular, no correlacionadas), las hipótesis de la LDGN se cumplen, y se concluye que

$$\frac{S_n - np}{n} \xrightarrow{P} 0,$$

o equivalentemente,

$$\frac{S_n}{n} \xrightarrow{P} p.$$

Aun cuando este resultado es una aplicación directa de la Proposición 6.3.2, lo interesante es que fue probado por Bernoulli en 1713, muchos a nos antes que Tchebyshev publicara su resultado.

Veremos a continuación la LFGN, resultado que se enuncia sin demostración.

Proposición 6.3.3 (Ley Fuerte de Kolmogorov)
Sean $X_1, X_2, \ldots$ variables aleatorias independientes e idénticamente distribuidas con $E(|X_n|) < \infty$, y $E(X_n) = \mu$. Entonces

$$\frac{S_n}{n} \xrightarrow{\text{c.s.}} \mu.$$

Observe que la LFGN, en contraste con la LDGN, no requiere existencia de la varianza de las variables aleatorias, aun cuando el supuesto que éstas sean i.i.d. es fundamental

Ejemplo 6.3.2 En el Ejemplo 6.3.1, la convergencia en probabilidad de S_n/n, es en realidad casi segura. Esto es una consecuencia directa de la LFGN de Kolmogorov.

Ejemplo 6.3.3 (Función de distribución empírica)
Considere una muestra aleatoria $X_1, X_2, \ldots$ de una cierta función de distribución F. La *función de distribución empírica* de esta muestra se define mediante

$$\hat{F}_n(x) = \frac{\text{Número de } X_i \text{ que son} \leq x}{n}.$$

Esta función se puede interpretar como una aproximación a la verdadera función de distribución $F(x) = P(X \leq x)$. Veamos que $\hat{F}_n(x) \xrightarrow{\text{c.s.}} F(x)$. Para ello, defina las variables $Y_1, Y_2, \ldots$:

$$Y_i = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{si no.} \end{cases}$$

Se tiene que $Y_1, Y_2, \ldots$ es un proceso de Bernoulli, con probabilidad de éxito

$$p = P(Y_1 = 1) = P(X_1 \leq x) = F(x).$$

Además, note que si $S_n = \sum_{i=1}^{n} Y_i$, entonces

$$\hat{F}_n(x) = \frac{S_n}{n} \xrightarrow{\text{c.s.}} p = F(x),$$

de donde se tiene el resultado. En otras palabras, la función de distribución empírica converge a la función de distribución F. Este resultado es útil para identificar la distribución F cuando se dispone de una muestra de F, y F no se conoce. Esta situación es común en problemas de Estadística.

Ejemplo 6.3.4 (Aproximación de una integral)
Considere una función $f(x)$ a valores reales, continua, definida en un intervalo $[a, b]$, y suponga que interesa calcular $I = \int_a^b f(x)dx$. Para ello utilizaremos el siguiente procedimiento. Supongamos en primer instancia que $f(x) \geq 0$ para $a \leq x \leq b$. Sea $M > 0$ un número real tal que $f(x) \leq M$ para todo $x \in [a, b]$. Tal número existe, pues cualquier función continua es acotada sobre intervalos cerrados. Así, el gráfico de la función queda comprendido en el rectángulo $[a, b] \times [0, M]$ (ver Figura 6.3.1).

Sean $(U_{11}, U_{12}), (U_{21}, U_{22}), \ldots$ vectores aleatorios i.i.d. con distribución uniforme en el rectángulo $[a, b] \times [0, M]$ (note que ello implica que U_{i1} es independiente de U_{i2} para todo $i \geq 1$), y defina las variables aleatorias $X_1, X_2, \ldots$ mediante

$$X_i = \begin{cases} 1 & \text{si } f(U_{i1}) > U_{i2} \\ 0 & \text{si no.} \end{cases}$$

Así, la variable X_i toma el valor 1 si el punto $\boldsymbol{U}_i = (U_{i1}, U_{i2})$ está por debajo del gráfico de la curva $y = f(x)$, y toma el valor 0 si no. Puesto que los vectores $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots$ son i.i.d., $X_1, X_2, \ldots$ es un proceso de Bernoulli con probabilidad de éxito p dada por $p = P(X_1 = 1)$. Para calcular dicha probabilidad, notemos que la densidad conjunta de $\boldsymbol{U}_1$ es

$$f_{U_{11},U_{12}}(u_{11}, u_{12}) = \frac{1}{M(b-a)}, \qquad \text{si } (u_{11}, u_{12}) \in [a, b] \times [0, M].$$

Luego,

$$p = \int_a^b \int_0^{f(u_{11})} \frac{1}{M(b-a)}\, du_{12} du_{11} = \frac{1}{M(b-a)} \int_a^b f(u_{11}) du_{11} = \frac{I}{M(b-a)}.$$

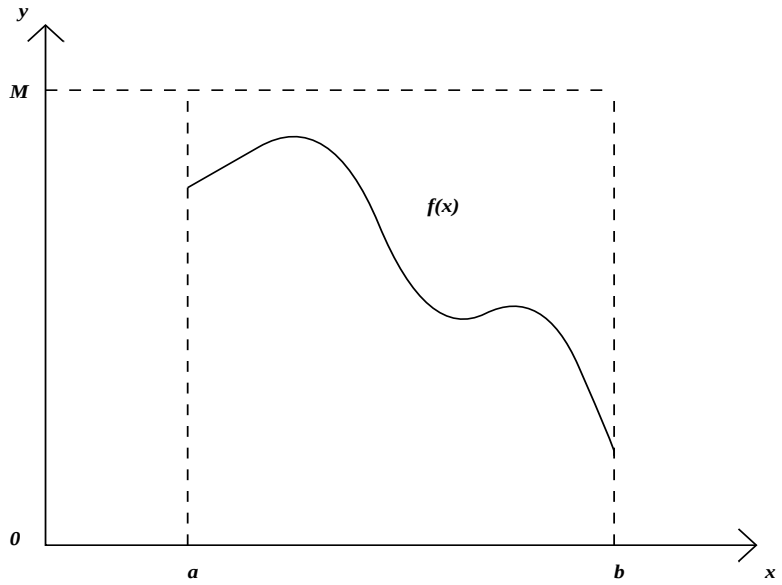


Figura 6.3.1: Aproximación de una integral, correspondiente al área bajo la curva $y = f(x)$, entre a y b.

Luego, de la LFGN se deduce que

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{\text{c.s.}} \frac{I}{M(b-a)},$$

o equivalentemente,

$$\frac{M(b-a)(X_1 + X_2 + \cdots + X_n)}{n} \xrightarrow{\text{c.s.}} I.$$

Este resultado sugiere el siguiente método para aproximar una integral del tipo de I:

- Generar una gran cantidad de puntos al azar en el rectángulo $[a, b] \times [0, M]$.
- Calcular la fracción de puntos que cae bajo el gráfico de la curva $y = f(x)$.
- Dicha fracción coincide con $(X_1 + \cdots + X_n)/n$, y multiplicada por $M(b-a)$, es una aproximación a I

La calidad de dicha aproximación ciertamente dependerá de la cantidad de puntos que se utilice, y además de cuan cerca esté M de la cantidad $\max\limits_{a \le x \le b} f(x)$. Volveremos a este punto en la Sección 6.5.

Puesto que virtualmente todos los paquetes estadísticos, y muchos lenguajes de programación poseen rutinas para generar números aleatorios, el método se puede implementar fácilmente.

Por último, si la función $f(x)$ no es positiva, entonces consideramos la función $g(x) = f(x) - m$, donde $m = \min\{f(x) \ : \ a \le x \le b\}$. Así, $I = m(b-a) + \int_a^b g(x)dx$, y el método se aplica a $g(x)$.

Ejemplo 6.3.5 Sean $X_1, X_2, \ldots$ i.i.d. con distribución común uniforme en el intervalo $[0,1]$, y considere la sucesión $Y_1, Y_2, \ldots$, con

$$Y_n = \left(\prod_{i=1}^{n} X_i \right)^{\frac{1}{n}}.$$

Y_n recibe el nombre de *media geométrica* de $X_1, \ldots, X_n$. Calculemos el límite casi seguro de $\{Y_n\}$. Note que Y_n no tiene la forma de un promedio de variables aleatorias, de modo que la LFGN no se puede aplicar directamente. Sin embargo, si $Z_n = \log(Y_n)$, entonces

$$Z_n = \frac{\log(X_1) + \log(X_2) + \cdots + \log(X_n)}{n},$$

la que tiene la forma apropiada. Además, por lo hecho en el Ejemplo 3.9.3, se tiene que $-\log(X_1) \sim \text{Exp}(1)$, de donde se sigue que $E(\log(X_1)) = -1$. Por otra parte, las variables aleatorias $\log(X_1), \log(X_2), \ldots$ son i.i.d., y la LFGN permite concluir que

$$Z_n \xrightarrow{\text{c.s.}} E(\log(X_1)) = -1.$$

Pero $Y_n = \exp(Z_n)$, y por la Proposición 6.2.3(c) se obtiene $Y_n \xrightarrow{\text{c.s.}} e^{-1}$.

6.4. Función Característica y Convergencia en Distribución

Retomamos aquí el estudio de la noción de convergencia en distribución. En la Sección 3.8.3 vimos que existe una correspondencia uno a uno entre la distribución de una variable aleatoria X y su función característica $\varphi_X(t)$. Tomando este hecho en consideración, es intuitivo pensar que debe existir alguna relación entre la convergencia en distribución de la sucesión $\{X_n\}$, y la sucesión de funciones características $\{\varphi_{X_n}(t)\}$. Similares argumentos se pueden aplicar al caso de vectores aleatorios.

La respuesta a esta inquietud está dada por el siguiente resultado.

Teorema 6.4.1 Sean X, X_1, X_2 variables aleatorias. Entonces, $X_n \xrightarrow{D} X$ si y sólo si se cumple que $\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t)$ para todo $t \in \mathbb{R}$.

Este resultado es en realidad una caracterización de la convergencia en distribución. De hecho, lo utilizaremos como una definición de convergencia para el caso de vectores aleatorios.

Definición 6.4.1 Sean $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ vectores aleatorios en $\mathbb{R}^k$. Diremos que $\{\boldsymbol{X}_n\}$ *converge en distribución* a $\boldsymbol{X}$ si para cualquier $\boldsymbol{t} \in \mathbb{R}^k$ se tiene que $\lim_{n\to\infty} \varphi_{\boldsymbol{X}_n}(t) = \varphi_{\boldsymbol{X}}(\boldsymbol{t})$.

Existe una caracterización alternativa de convergencia en distribución de vectores aleatorios, que damos a continuación.

Teorema 6.4.2 (Cramér-Wold)
Sean $\boldsymbol{X}_n = (X_{n1}, \ldots, X_{nk})$ y $\boldsymbol{X} = (X_1, \ldots, X_k)$ vectores aleatorios en $\mathbb{R}^k$. Entonces $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ si y sólo si para cualquier $\boldsymbol{t} = (t_1, \ldots, t_k) \in \mathbb{R}^k$ se cumple

$$t_1 X_{n1} + \cdots + t_k X_{nk} \xrightarrow{D} t_1 X_1 + \cdots + t_k X_k,$$

cuando $n \to \infty$.

Este resultado hace uso del hecho que la distribución de un vector aleatorio queda determinada por la distribución de todas las combinaciones lineales posibles de sus coordenadas.

La siguiente variación del Teorema 6.4.1 resulta ser muy útil para establecer convergencia en distribución de una sucesión de vectores aleatorios.

Teorema 6.4.3 (Paul Lévy)
Sean $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ vectores aleatorios definidos en $\mathbb{R}^k$, con funciones características respectivas $\varphi_{\boldsymbol{X}_1}(\boldsymbol{t}), \varphi_{\boldsymbol{X}_2}(\boldsymbol{t}), \ldots$. Si $\varphi_{\boldsymbol{X}_n}(\boldsymbol{t})$ converge puntualmente a un límite $\varphi(\boldsymbol{t})$, y si $\varphi(\boldsymbol{t})$ es continua en $\boldsymbol{t} = \boldsymbol{0}$, entonces

(a) Existe un vector aleatorio $\boldsymbol{X}$ tal que $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, y

(b) $\varphi(\boldsymbol{t})$ es la función característica de $\boldsymbol{X}$.

El Teorema 6.4.3 también vale para el caso particular $k = 1$, es decir, para variables aleatorias. Veamos a continuación algunas aplicaciones de estos resultados.

Ejemplo 6.4.1 Si $X_n \sim N(\mu_n, \sigma_n^2)$, donde $\{\mu_n\}$ y $\{\sigma_n^2\}$ son sucesiones convergentes a μ y $\sigma^2 > 0$ respectivamente, entonces $X_n \xrightarrow{D} X \sim N(\mu, \sigma^2)$. En efecto, tenemos que $\varphi_{X_n}(t) = \exp(i\mu_n t - t^2\sigma_n^2/2)$, y tomando límite, se encuentra que $\varphi_{X_n}(t)$ converge a $\varphi(t) = \exp(i\mu t - t^2\sigma^2/2)$. Puesto que este límite es claramente una función continua en $t = 0$ (más aun, es continua en todo $t \in \mathbb{R}$), el Teorema 6.4.3 asegura la existencia de una variable aleatoria X tal que $X_n \xrightarrow{D} X$. Pero puesto que el mismo Teorema garantiza que $\varphi_X(t) = \varphi(t)$, y $\varphi(t)$ es la función característica de una variable aleatoria con distribución $N(\mu, \sigma^2)$, el resultado se tiene por la correspondencia uno a uno entre la distribución de una variable aleatoria y su función característica.

Ejemplo 6.4.2 El resultado del Ejemplo 6.4.1 se puede generalizar a vectores aleatorios. Si $\boldsymbol{X}_n \sim N_k(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, con $\lim_{n\to\infty} \boldsymbol{\mu}_n = \boldsymbol{\mu}$ y $\lim_{n\to\infty} \boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$, donde $\boldsymbol{\Sigma}$ es semidefinida positiva, entonces $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Los detalles son muy parecidos a los del Ejemplo 6.4.1, y se proponen como ejercicio.

Ejemplo 6.4.3 Sean $X_1, X_2, \ldots$ i.i.d. con $P(X_k = 1) = P(X_k = -1) = 1/2$. Vamos a probar que el límite en distribución de $Y_n = \sum_{k=1}^{n} X_k/2^k$ es una variable aleatoria

$Y \sim U(-1,1)$. Para ello, note que

$$\begin{aligned}\varphi_{X_k}(t) &= E(e^{itX_k}) = \frac{e^{it}+e^{-it}}{2} \\ &= \frac{\cos(t)+i\sin(t)+\cos(-t)+i\sin(-t)}{2} = \frac{2\cos(t)}{2} \\ &= \cos(t).\end{aligned}$$

Así, se tiene que

$$\varphi_{Y_n}(t) = \prod_{k=1}^{n} \varphi_{X_k/2^k}(t) = \prod_{k=1}^{n} \cos(t/2^k).$$

Por otra parte, notemos que de la identidad $\sin(2t) = 2\sin(t)\cos(t)$ se concluye que

$$\cos(t/2^k) = \frac{\sin(t/2^{k-1})}{2\sin(t/2^k)},$$

para $k = 1, 2, \ldots$. Luego,

$$\varphi_{Y_n}(t) = \prod_{k=1}^{n} \frac{\sin(t/2^{k-1})}{2\sin(t/2^k)} = \frac{\sin(t)}{2^n \sin(t/2^n)},$$

y recordando que $\lim\limits_{x\to 0} \sin(x)/x = 1$, vemos que

$$\lim_{n\to\infty} \varphi_{Y_n}(t) = \frac{\sin(t)}{t}.$$

Notemos que este límite se puede definir como 1 para $t = 0$, caso en el que la función resultante es continua en 0 (recuerde que cualquier función característica evaluada en $t = 0$ vale 1). Sea ahora $Y \sim U(-1,1)$, y calculemos su función característica. Se tiene

$$\begin{aligned}\varphi_Y(t) &= \int_{-1}^{1} \cos(tx)\frac{1}{2}dx + i\int_{-1}^{1} \sin(tx)\frac{1}{2}dx = \frac{1}{2}\int_{-1}^{1} \cos(tx)dx \\ &= \frac{\sin(t)-\sin(-t)}{2t} = \frac{\sin(t)}{t},\end{aligned}$$

que coincide con el límite de $\varphi_{Y_n}(t)$. En virtud del Teorema 6.4.3, hemos mostrado que $Y_n \xrightarrow{D} Y \sim U(-1,1)$.

Para terminar esta sección, veremos dos resultados adicionales de convergencia en distribución, los que resultan ser muy útiles en una variedad de aplicaciones.

Teorema 6.4.4 (Scheffé)
Sean $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ vectores aleatorios en $\mathbb{R}^k$, con densidades respectivas $f(\boldsymbol{x}), f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots$. Si para todo $\boldsymbol{x} \in \mathbb{R}^k$ se cumple $\lim\limits_{n\to\infty} f_n(\boldsymbol{x}) = f(\boldsymbol{x})$, entonces $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

En otras palabras, el Teorema 6.4.4 establece que si la sucesión de densidades converge puntualmente a una cierta densidad, entonces existe convergencia en distribución. Por otra parte, no es necesario que la convergencia ocurra para absolutamente todos los puntos $\boldsymbol{x} \in \mathbb{R}^k$, pudiendo ésta no verificarse en un conjunto numerable de puntos en $\mathbb{R}^k$.

Ejemplo 6.4.4 En el Ejemplo 6.2.7 se mostró que $Z_n = n\,\mín\{X_1, \ldots, X_n\}$ converge en distribución a una cierta variable aleatoria Z con distribución exponencial de media 1, usando la Definición 6.2.1(a). Veamos ahora lo mismo usando el Teorema 6.4.4. Puesto que $F_{Z_n}(z) = 1 - (1 - z/n)^n$, entonces

$$f_{Z_n}(z) = (1 - z/n)^{n-1}, \qquad 0 < z/n < 1,$$

de donde se tiene que

$$\lim_{n\to\infty} f_{Z_n}(z) = e^{-z}, \qquad z > 0,$$

que corresponde a la densidad de Z.

Definamos ahora $W_n = n(1 - \máx\{X_1, \ldots, X_n\})$. Por (4.4.8), y usando la transformación $g(x) = n(1 - x)$, se tiene que

$$f_{W_n}(w) = (1 - w/n)^{n-1}, \qquad 0 < w/n < 1,$$

y por el argumento anterior, se concluye que $W_n \xrightarrow{D} W \sim \text{Exp}(1)$. Veamos ahora qué sucede con la distribución *conjunta* de (Z_n, W_n). Por lo hecho en el Ejemplo 4.4.8, y usando el cambio de variables $g(x, y) = (nx, n(1 - y))$, se tiene que

$$f_{Z_n,W_n}(z, w) = \begin{cases} \frac{n(n-1)}{n^2}(1 - \frac{z}{n} - \frac{w}{n})^{n-2} & \text{si } 0 \le \frac{z}{n} < 1 - \frac{w}{n} < 1 \\ 0 & \text{si no.} \end{cases}$$

Tomando límite, se tiene que esta densidad conjunta converge a

$$f_{Z,W}(z, w) = e^{-z-w}, \qquad z, w > 0,$$

y se concluye que $(Z_n, W_n) \xrightarrow{D} (Z, W)$, donde Z y W son i.i.d. con distribución exponencial de media 1.

Teorema 6.4.5 (Slutsky)
Sean $X, X_1, X_2, \ldots$ e $Y_1, Y_2, \ldots$ variables aleatorias tales que $X_n \xrightarrow{D} X$ e $Y_n \xrightarrow{P} c$, donde c es una constante. Entonces:

(a) $X_n + Y_n \xrightarrow{D} X + c$.

(b) $X_n - Y_n \xrightarrow{D} X - c$.

(c) $Y_n X_n \xrightarrow{D} cX$.

(d) Si $c \neq 0$ y $P(Y_n \neq 0) = 1$,

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}.$$

El Teorema 6.4.5 se usa fundamentalmente para construir nuevas sucesiones de variables aleatorias que convergen en distribución a partir de casos en que se conozca dicha convergencia previamente. Este resultado se usa habitualmente en combinación con el Teorema Central del Límite, tema de nuestra próxima sección.

6.5. El Teorema Central del Límite

Hemos dejado para esta última sección uno de los resultados fundamentales de la Teoría de Probabilidades. Hasta ahora hemos visto en la Sección 6.3 que promedios de variables aleatorias i.i.d. con valor esperado finito μ, convergen a μ. Este resultado permite justificar una interpretación de probabilidad desde un punto de vista *frecuentista*, es decir, las probabilidades se pueden concebir como límites de frecuencias relativas de eventos, si el experimento en cuestión se repite indefinidamente en forma independiente y siempre bajo las mismas condiciones. Sin embargo, las Leyes de Grandes Números no establecen cuan cerca está – en términos de probabilidades – este promedio de variables aleatorias del valor μ al que converge. En otras palabras, sería deseable saber cuál es la probabilidad que este promedio difiera de μ en menos que una cantidad prefijada $\delta > 0$.

Establecemos a continuación el resultado básico que nos permite calcular (al menos aproximadamente) probabilidades como las descritas en el párrafo anterior, del que veremos primero la versión univariada.

Teorema 6.5.1 (Teorema Central del Límite (TCL))
Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d., con $E(X_1) = \mu$ y $\mathrm{Var}(X_1) = \sigma^2 > 0$. Entonces se tiene que

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1). \tag{6.5.1}$$

Demostración: Supondremos en primer lugar que $\mu = 0$ y $\sigma = 1$. En este caso, $Z_n = \sqrt{n}\overline{X}_n$, y debemos probar que Z_n converge en distribución a una variable aleatoria Z con distribución $N(0,1)$. Para ello, usaremos el Teorema 6.4.3, en virtud del que basta probar que la sucesión de funciones características $\{\varphi_{Z_n}(t)\}$ converge para todo t a $e^{-t^2/2}$. Si $S_n = \sum_{j=1}^{n} X_j$, entonces $\overline{X}_n = S_n/n$. Por la independencia de $X_1, X_2, \ldots$ se tiene

$$\varphi_{S_n}(t) = \prod_{j=1}^{n} \varphi_{X_j}(t) = \varphi(t)^n,$$

donde $\varphi(t) = \varphi_{X_1}(t)$. Luego,

$$\varphi_{Z_n}(t) = \varphi_{S_n/\sqrt{n}}(t) = \left(\varphi(t/\sqrt{n})\right)^n.$$

Puesto que $E(X_1^2) < \infty$, es posible probar (no lo haremos) que su función característica (que hemos denotado $\varphi(t)$), posee dos derivadas continuas. Luego, podemos hacer un desarrollo en serie de Taylor de orden 2, para obtener

$$\varphi(t) = \varphi(0) + \varphi'(0) \cdot t + \varphi''(\theta(t)) \cdot \frac{t^2}{2},$$

donde $|\theta(t)| \leq |t|$. Luego,

$$\varphi(t) = \varphi(0) + \varphi'(0) \cdot t + \varphi''(0) \cdot \frac{t^2}{2} + \frac{t^2}{2} e(t),$$

donde $e(t) = \varphi''(\theta(t)) - \varphi''(0)$, y $e(t) \to 0$ cuando $t \to 0$. Por otra parte, y usando propiedades de funciones características, se tiene que $\varphi(0) = 1$, $\varphi'(0) = i\mu = 0$ y $\varphi''(0) = i^2 E(X_1^2) = -E(X_1^2) = -1$. Por lo tanto,

$$\varphi(t) = 1 - \frac{t^2}{2} + \frac{t^2 e(t)}{2},$$

de donde se deduce que

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + \frac{t^2 e(t/\sqrt{n})}{2n}\right)^n = \left(1 - \frac{t^2}{2n}\left\{1 - e(t/\sqrt{n})\right\}\right)^n.$$

El resultado se obtiene directamente, una vez que se prueba que si $\{c_n\}$ es una sucesión de números complejos tales que $c_n \to c$ cuando $n \to \infty$ entonces

$$\left(1 + \frac{c_n}{n}\right)^n \to e^c,$$

lo cual se propone como ejercicio. Finalmente, para el caso general $\mu \in \mathbb{R}$ y $\sigma^2 > 0$, defina

$$Y_n = \frac{X_n - \mu}{\sigma},$$

de modo que $Z_n = \sqrt{n}\overline{Y}_n$, y lo hecho recientemente se aplica a las variables aleatorias (i.i.d.) $Y_1, Y_2, \ldots$. ■

Uno de los aspectos más interesantes del Teorema 6.5.1 es que la convergencia vale *cualquiera que sea* la distribución original de las variables aleatorias involucradas. Por ejemplo, no hace falta que las variables sean continuas, la convergencia también vale para variables aleatorias discretas, aun cuando es necesario tener cierto cuidado en aproximar distribuciones discretas por una normal.

Veamos algunas aplicaciones del TCL.

Ejemplo 6.5.1 Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con distribución $N(\mu, \sigma^2)$. El TCL establece entonces que

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1).$$

Sin embargo, en este caso particular, dicha convergencia es trivial, pues se tiene que la distribución *exacta* de Z_n es $N(0,1)$.

Supongamos ahora que $\mu = 0$. La LFGN establece que

$$\frac{X_1^2 + \cdots + X_n^2}{n} \xrightarrow{\text{c.s.}} E(X_1^2) = \sigma^2.$$

Además, $\text{Var}(X_1) = 2\sigma^4$ (ver Ejemplo 3.8.9), de modo que si $\overline{Y}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, el TCL implica que

$$\frac{\sqrt{n}(\overline{Y}_n - \sigma^2)}{\sqrt{2}\sigma^2} \xrightarrow{D} Z \sim N(0,1).$$

Además, usando el Teorema 6.4.5, es fácil ver que

$$\frac{\sqrt{n}(\overline{Y}_n - \sigma^2)}{\overline{Y}_n\sqrt{2}} \xrightarrow{D} Z \sim N(0,1).$$

Luego, si $\delta > 0$ se tiene que

$$\begin{aligned} P(|\overline{Y}_n - \sigma^2| < \delta) &= P\left(\left|\frac{\sqrt{n}(\overline{Y}_n - \sigma^2)}{\overline{Y}_n\sqrt{2}}\right| < \frac{\sqrt{n}\delta}{\overline{Y}_n\sqrt{2}}\right) \\ &\approx P\left(|Z| < \frac{\sqrt{n}\delta}{\overline{Y}_n\sqrt{2}}\right) \\ &= 2\Phi\left(\frac{\sqrt{n}\delta}{\overline{Y}_n\sqrt{2}}\right) - 1, \end{aligned}$$

de modo que dado el valor de $\delta > 0$, y conocido el valor de $\overline{Y}_n$ (a partir de una muestra de tama no n) el valor de $P(|\overline{Y}_n - \sigma^2| < \delta)$ se puede aproximar. Por ejemplo, si $n = 100$, $\delta = 1$ e $\overline{Y}_n = 2{,}7$, la probabilidad se aproxima por 0.991179. Note que para realizar este cálculo, no se requiere conocer el valor de σ^2.

Ejemplo 6.5.2 Supongamos que $X_1, X_2, \ldots, X_{100}$ son i.i.d. con distribución exponencial de media 5, y calculemos aproximadamente $P(S_{100} > 600)$, donde $S_{100} = \sum_{k=1}^{100} X_k$. Tenemos que $E(X_1) = 5$, $\text{Var}(X_1) = 25$, de modo que la variable Z_{100} en (6.5.1) se transforma en

$$Z_{100} = \frac{10(\overline{X}_{100} - 5)}{5} = 2(\overline{X}_{100} - 5),$$

la que tiene distribución aproximadamente $N(0,1)$. Ahora,

$$\begin{aligned} P(S_{100} > 600) &= P(\overline{X}_{100} > 6) = P(2(\overline{X}_{100} - 5) > 2 * (6 - 5)) \\ &\approx P(Z > 2), \end{aligned}$$

donde $Z \sim N(0,1)$, y usando las tablas adecuadas, se puede obtener que

$$P(S_{100} > 600) \approx P(Z > 2) = 0{,}0228.$$

Por otra parte, y recordando que $S_{100} \sim \Gamma(100, 5)$, la probabilidad *exacta* se expresa mediante

$$\int_{600}^{\infty} \frac{x^{99} e^{-x/5}}{99!\, 5^{100}} dx,$$

y mediante integración numérica se obtiene que este valor es 0.0279, de modo que la aproximación es razonablemente buena.

Es claro que la calidad de aproximaciones basadas en el Teorema 6.5.1 dependen del valor de n. Para tener una mejor idea al respecto, el siguiente resultado es útil.

Teorema 6.5.2 (Berry-Esséen)
Bajo las hipótesis del Teorema 6.5.1, y si $G_n(t) = P(Z_n \leq t)$, donde Z_n fue definida en (6.5.1), entonces se tiene la siguiente cota:

$$\sup_{t \in \mathbb{R}} |G_n(t) - \Phi(t)| \leq \frac{33}{4} \frac{E(|X_1 - \mu|^3)}{\sigma^3 \sqrt{n}}, \qquad \forall n. \tag{6.5.2}$$

En la práctica, esta cota resulta ser casi siempre muy difícil de calcular. No obstante, lo interesante del resultado es que el máximo posible error cometido en las aproximaciones es del orden de $n^{-1/2}$. Para visualizar un poco mejor esta aproximación, note que el Teorema 6.5.1 implica que para n grande, la distribución de $\overline{X}_n$ es aproximadamente $N(\mu, \sigma^2/n)$. A este efecto, se generaron en un computador 1000 muestras de tama no $n = 100$ cada una, de la distribución exponencial con media 5, tal como en el Ejemplo 6.5.2. Por cada muestra se obtuvo el promedio de los valores generados en dicha muestra, los que designamos por $\overline{X}^1, \ldots, \overline{X}^{1000}$, y cuya distribución aproximada es $N(5, 1/4)$. Estos valores se usaron para construir un histograma, y la función de distribución empírica (ver Ejemplo 6.3.3), los que se muestran en la Figura 6.5.2. El histograma se construyó de modo que la suma de las área de las distintas barras sea igual a 1, de modo que la figura que se obtiene es una aproximación a la densidad $N(5, 1/4)$, que aparece representada en línea continua. Por otra parte, la función de distribución empírica (línea punteada) es una aproximación a $F_Y(y)$, donde $Y \sim N(5, 1/4)$ (línea continua). Se aprecia que la aproximación es, en términos generales, bastante buena.

Ejemplo 6.5.3 (Aproximando la distribución Binomial)
En el Ejemplo 6.2.5 se mostró que si $X \sim \text{Bin}(n, p_n)$ con $np_n \to \lambda > 0$, entonces la distribución de X se puede aproximar por la distribución de Poisson con parámetro λ, provisto que n es grande. Consideremos ahora el siguiente enfoque alternativo. Sean $Y_1, Y_2, \ldots, Y_n$ i.i.d. con distribución Bernoulli de parámetro p_n. Entonces la distribución de $Y = \sum_{k=1}^{n} Y_k$ coincide con la de X. Puesto que $E(Y_1) = p_n$ y $\text{Var}(Y_1) = p_n(1 - p_n)$, el Teorema Central del Límite implica que la distribución de Y es también aproximadamente $N(np_n, np_n(1 - p_n))$. Puesto que $np_n \approx \lambda$, se deduce que Y tiene distribución $N(\lambda, \lambda)$, aproximadamente, donde hemos despreciado el término np_n^2.

Ahora bien, en el momento de aproximar probabilidades binomiales (más generalmente, aquellas provenientes de distribuciones discretas) mediante la distribución normal,

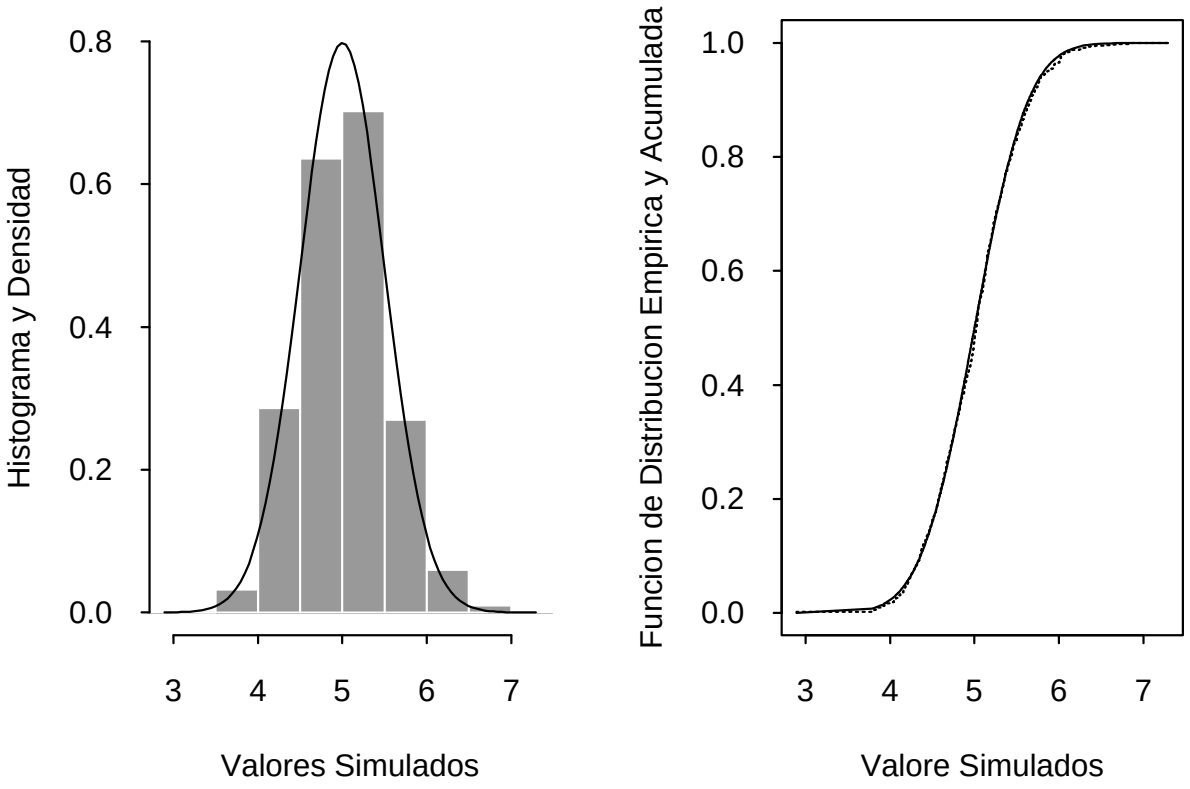


Figura 6.5.2: Distribución del promedio de 100 variables aleatorias i.i.d. con distribución exponencial de media 5, y aproximación normal mediante Teorema Central del Límite.

es necesario tener el siguiente cuidado. Si $X \sim \text{Bin}(n, p_n)$, entonces $P(X = 2)$ es una cantidad positiva (aun cuando su valor puede ser despreciable en ciertos casos). Puesto que X tiene distribución aproximadamente igual a la de $Y \sim N(\lambda, \lambda)$, al usar esta aproximación nos encontramos con que $P(Y = 2) = 0$, pues Y es continua, y esto sucede para cualquier otro valor particular de interés. Para corregir este problema se usa la llamada *corrección de continuidad*, que consiste en aproximar $P(X = k)$ mediante $P(k - \frac{1}{2} < Y < k + \frac{1}{2})$, es decir,

$$P(X = k) \approx \Phi\left(\frac{k + \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{k - \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right). \tag{6.5.3}$$

En otras palabras, se asume que el punto k representa el intervalo $[k - \frac{1}{2}, k + \frac{1}{2}]$ para la distribución normal, al momento de usar la aproximación.

A modo de ejemplo, consideremos el caso $n = 100$, $p_n = 0{,}05$, con lo que $\lambda = np_n =$

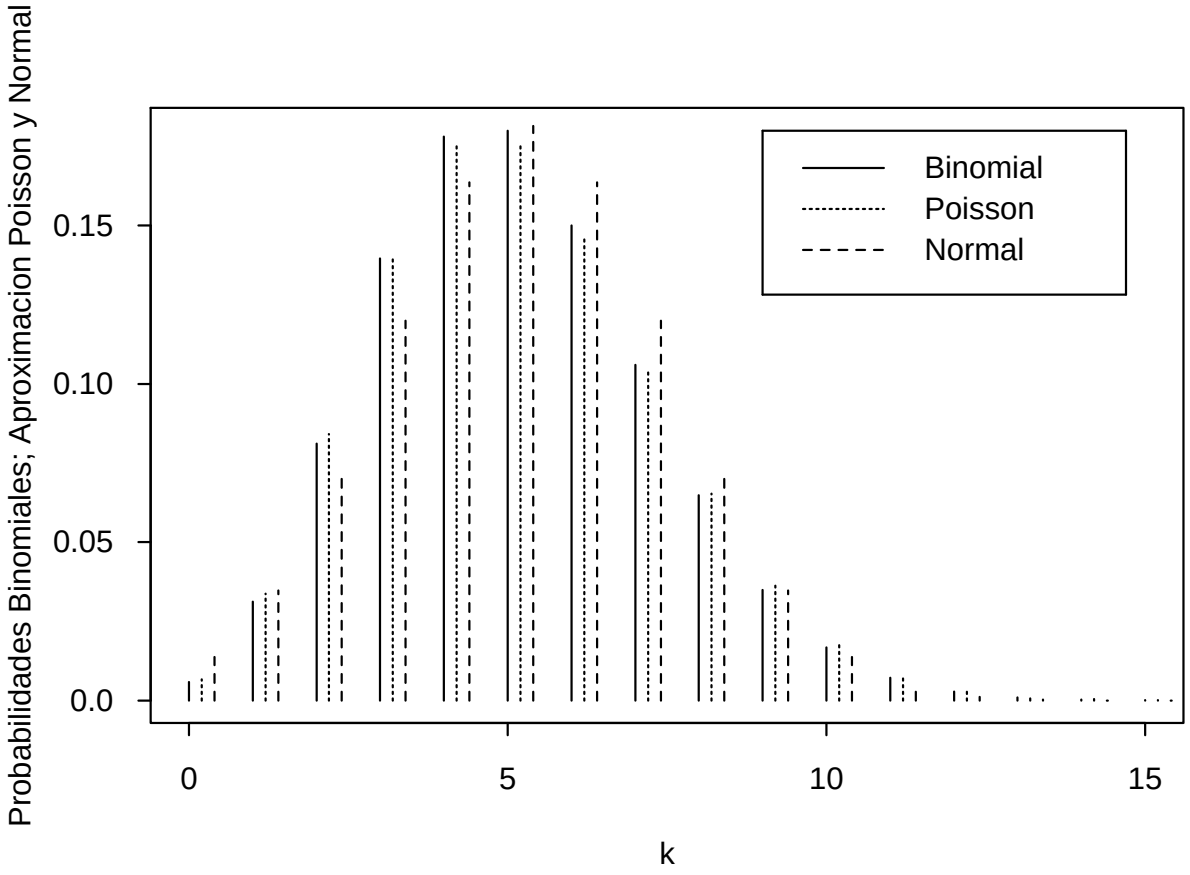


Figura 6.5.3: Aproximaciones $\mathrm{Poisson}(5)$ y $N(5,5)$ a la distribución $\mathrm{Bin}(100, 0{,}05)$.

5. La Figura 6.5.3 muestra las probabilidades exactas correspondientes a dicha distribución, así como las aproximaciones derivadas de la distribución de Poisson y Normal, como se detalló anteriormente. Es claro que la aproximación Poisson es superior a la Normal para este caso. Sin embargo, la aproximación Normal es usualmente más simple de calcular, y su precisión aumenta a medida que n crece.

Veamos ahora la extensión multivariada del Teorema 6.5.1.

Teorema 6.5.3 (Teorema Central del Límite Multivariado)
Sean $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ vectores aleatorios i.i.d. en $\mathbb{R}^k$, con $E(\boldsymbol{X}_1) = \boldsymbol{\mu}$ y $V(\boldsymbol{X}_1) = \boldsymbol{\Sigma}$, donde $\boldsymbol{\mu} \in \mathbb{R}^k$ y $\boldsymbol{\Sigma}$ es una matriz definida positiva. Entonces

$$\boldsymbol{Z}_n = \sqrt{n}(\overline{\boldsymbol{X}_n} - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Z} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{6.5.4}$$

donde $\overline{\boldsymbol{X}_n} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{X}_k$, y la suma se interpreta coordenada a coordenada.

Ejemplo 6.5.4 Sean $Y_1, Y_2, \ldots$ variables aleatorias i.i.d. con distribución $N(\mu, \sigma^2)$, donde $\mu \in \mathbb{R}$ y $\sigma^2 > 0$. Defina los vectores $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ en $\mathbb{R}^2$ mediante

$$\boldsymbol{X}_k = \begin{pmatrix} Y_k \\ Y_k^2 \end{pmatrix}.$$

Se tiene que $E(Y_k) = \mu$, $E(Y_k^2) = \mathrm{Var}(Y_k) + (E(Y_k))^2$, de modo que

$$E(\boldsymbol{Y}_k) = \begin{pmatrix} \mu \\ \sigma^2 + \mu^2 \end{pmatrix}.$$

Por otra parte, se tiene que $E(Y_k^3) = \mu^3 + 3\mu\sigma^2$ y además $E(Y_k^4) = 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4$ (verificar esto como ejercicio). Luego,

$$\mathrm{Cov}(Y_k, Y_k^2) = E(Y_k^3) - E(Y_k)E(Y_k^2) = \mu^3 + 3\mu\sigma^2 - \mu(\mu^2 + \sigma^2) = 2\mu\sigma^2,$$

y además

$$\mathrm{Var}(Y_k^2) = E(Y_k^4) - (E(Y_k^2))^2 = 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4 - (\mu^2 + \sigma^2)^2 = 2\sigma^4 + 4\mu^2\sigma^2,$$

y por lo tanto se tiene que

$$\boldsymbol{\Sigma} = V(\boldsymbol{X}_k) = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 2\sigma^4 + 4\mu^2\sigma^2 \end{pmatrix}.$$

Luego, el Teorema 6.5.4 asegura que

$$\sqrt{n}\left\{ \begin{pmatrix} \frac{1}{n}\sum_{k=1}^{n} Y_k \\ \frac{1}{n}\sum_{k=1}^{n} Y_k^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix} \right\} \xrightarrow{D} N_2(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Para finalizar esta sección, veamos otro resultado muy útil para verificar convergencia en distribución de funciones de promedios de variables o vectores aleatorios.

Teorema 6.5.4 (Método Delta)
Sea $\{\boldsymbol{X}_n\}$ una sucesión de vectores aleatorios en $\mathbb{R}^k$ tales que

$$\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

y sea $g : \mathbb{R}^k \to \mathbb{R}$ una función continuamente diferenciable en $\boldsymbol{x} = \boldsymbol{\mu}$. Entonces

$$\sqrt{n}(g(\boldsymbol{X}_n) - g(\boldsymbol{\mu})) \xrightarrow{D} N(0, \nabla g(\boldsymbol{\mu})' \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu})),$$

donde $\nabla g(\boldsymbol{x})$ es el vector (columna) de derivadas parciales de primer orden (o gradiente) de g evaluado en $\boldsymbol{x} = \boldsymbol{\mu}$.

Nota: Si $k = 1$, esto es, en el caso univariado, entonces la varianza de la distribución límite normal es $\sigma^2(g'(\mu))^2$.

Ejemplo 6.5.5 Sean $X_1, X_2, \ldots$ i.i.d. con distribución de Poisson con parámetro $\lambda > 0$. Por el Teorema 6.5.1 se tiene que

$$\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{D} N(0, \lambda).$$

Sea $g(x) = \sqrt{x}$, la que es continuamente diferenciable en $x = \lambda$. Puesto que $g'(\lambda) = \lambda^{-1/2}$, se concluye en virtud del Teorema 6.5.4, con $k = 1$,

$$\sqrt{n}(\sqrt{\overline{X}_n} - \sqrt{\lambda}) \xrightarrow{D} N(0, 1/4),$$

y observe que la distribución límite *no depende* de λ.

Ejemplo 6.5.6 En el Ejemplo 6.5.4, defina

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(Y_k - \overline{Y}_n)^2.$$

Es sencillo verificar que

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n} Y_k^2 - (\overline{Y}_n)^2,$$

de modo que por la LFGN se tiene que

$$\hat{\sigma}_n^2 \xrightarrow{\text{c.s.}} \mu^2 + \sigma^2 - (\mu)^2 = \sigma^2.$$

Para obtener la distribución límite (asintótica) de $\hat{\sigma}_n^2$ considere la función $g(x, y) = y - x^2$. Note que $g(\mu, \mu^2 + \sigma^2) = \sigma^2$. Además, se tiene que $g(\frac{1}{n}\sum_{k=1}^{n} Y_k, \frac{1}{n}\sum_{k=1}^{n} Y_k^2) = \hat{\sigma}_n^2$, y

$$\nabla g(x, y) = \begin{pmatrix} -2x \\ 1 \end{pmatrix},$$

por lo que es fácil verificar que

$$\nabla g(\mu, \mu^2 + \sigma^2)' \mathbf{\Sigma} \nabla g(\mu, \mu^2 + \sigma^2) = 2\sigma^4,$$

y por lo tanto hemos mostrado que

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{D} N(0, 2\sigma^4).$$

6.6. Problemas

1. En este problema se le pide verificar directamente el Teorema Central del Límite en algunos casos particulares, utilizando la convergencia de la f.g.m. a la de la distribución $N(0,1)$.

   Sean $X_1, \ldots, X_n$ iid con función generadora de momentos $M(t)$ y $X_i \sim F$ con media μ y varianza σ^2. Para cada uno de los siguientes casos: (i) $F = N(a, b^2)$ (ii) $F = \text{Exp}(\lambda)$ (iii) $F \sim \text{Poisson}(\lambda)$ (iv) $F \sim \text{Bin}(n, p)$:

   a.- Encuentre la f.g.m. de $S_n = \sum_{i=1}^{n} X_i$ y de $\overline{X}_n$.

   b.- Encuentre la f.g.m. $M_n(t)$ de $Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \sqrt{n}(\frac{\overline{X}_n - \mu}{\sigma})$.

   c.- Verifique que $M_n(t)$ tiende a $e^{\frac{t^2}{2}}$, cuando $n \to \infty$.

2. Una máquina empaquetadora de detergentes ha sido observada durante un largo tiempo, a través del cual se determinó que la varianza del peso de llenado es $\sigma^2 = 10$ gramos. Por otra parte el peso medio de llenado μ, depende del ajuste hecho a la máquina por cada operador.

   a.- Si mientras labora un mismo operador se realizan 25 observaciones, calcule aproximadamente la probabilidad que el peso medio observado se aleje en menos de 1 gramo de la media real de la máquina.

   b.- ¿Cuántas observaciones deben realizarse para asegurar que la probabilidad que lo mismo ocurra sea al menos 0.95?.

   $Resp: a) 0{,}88\ b) 39$

3. Suponga que dos dados se lanzan 600 veces. Sea X el número de veces en que se obtiene una suma de 7. Use el teorema central del límite para aproximar $P(90 < X < 110)$.

   $Resp.: 0{,}726.$

4. Si $X_1, \ldots, X_{20}$ son variables aleatorias iid Poisson con media 1, use el teorema central del límite para aproximar $P(\sum_{i=1}^{20} X_i > 15)$.

5. Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con $E(X_1) = 0$ y $Var(X_1) = \sigma^2$, donde $0 < \sigma^2 < \infty$. Sean $Y_1, Y_2, \ldots$ variables aleatorias i.i.d. tales que $E(Y_1) = \mu$, donde μ es un número real. Si $U_n = \frac{X_1 + \cdots + X_n}{n}$ y $V_n = \frac{Y_1 + \cdots + Y_n}{n}$, pruebe que

   $$U_n + \sqrt{n} V_n \longrightarrow N(\mu, \sigma^2)$$

   en distribución cuando $n \to \infty$.

6. Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con distribución Poisson(λ). Encuentre el límite en probabilidad de

   $$Y_n = \frac{X_1^2 + \cdots X_n^2}{n}.$$

   ¿Existe convergencia casi segura?

7. Sean $X_1, X_2, \ldots$ variables aleatorias tales que $E(X_n) \to \alpha$ y $Var(X_n) \to 0$. Pruebe que $X_n \xrightarrow{P} \alpha$.

8. Sean $X_1, X_2, \ldots$ variables aleatorias independientes con $X_1 = 0$, y tales que para $j \geq 2$ se tiene
$$P(X_j = k) = \begin{cases} j^{-3} & \text{si } k = \pm 1, \pm 2, \ldots, \pm j \\ 1 - 2j^{-2} & \text{si } k = 0. \end{cases}$$
Demuestre que si $\alpha > 1/2$
$$\frac{1}{n^\alpha} \sum_{j=1}^{n} X_j \xrightarrow{P} 0$$
cuando $n \longrightarrow \infty$. (*Indicación:* Use el hecho que $\sum_{k=1}^{j} k^2 = \frac{1}{6} j(j+1)(2j+1)$.)

9. Sean $X_1, X_2, \ldots$ independientes con distribución común $N(0,1)$. Calcule el límite casi seguro de
$$\frac{X_1^2 + \cdots X_n^2}{(X_1 - 1)^2 + \cdots + (X_n - 1)^2}.$$

10. Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con $X_1 \sim U(0, \theta)$ donde $\theta > 0$. Demuestre que $Y_n = \sqrt{3n}\{\log(2n^{-1} \sum_{i=1}^{n} X_i) - \log(\theta)\} \xrightarrow{D} Y$, con $Y \sim N(0,1)$.

11. Sean $X_1, X_2, \ldots$ variables aleatorias i.i.d. con $E(X_1) = 0$ y $E(X_1^2) = 2$. Encuentre el límite en distribución de las siguientes secuencias:

    (a) $Y_1, Y_2, \ldots$ donde
    $$Y_n = \frac{\sqrt{n}(X_1 + \cdots X_n)}{X_1^2 + \cdots X_n^2}.$$

    (b) $Z_1, Z_2, \ldots$ donde
    $$Z_n = \frac{X_1 + \cdots X_n}{\sqrt{X_1^2 + \cdots X_n^2}}.$$

12. (a) Suponga que $X_n \xrightarrow{D} N(0,1)$, $Y_n \xrightarrow{D} N(0,1)$ y que, para todo n, X_n sea independiente de Y_n. Muestre que $X_n + Y_n \xrightarrow{D} N(0,2)$.

    (b) Generalice el resultado de (a), probando que si $X_n \xrightarrow{D} F$, $Y_n \xrightarrow{D} G$, con X_n independiente de Y_n para todo n, entonces $X_n + Y_n \xrightarrow{D} Z$, donde la distribución de Z coincide con la de $X + Y$ tales que X e Y sean independientes y verificando $X \sim F$ e $Y \sim G$.

    Indicación: Use funciones características.

13. Sean $X_1, X_2, \ldots$ e $Y_1, Y_2, \ldots$ dos secuencias de variables aleatorias i.i.d. y tales que los X_j son independientes de los Y_k. Suponga que $E(X_1) = \mu_X$, $Var(X_1) = \sigma_X^2$, $E(Y_1) = \mu_Y$ y $Var(Y_1) = \sigma_Y^2$. Sea
$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n}{\overline{X}_n} - \frac{\mu_Y}{\mu_X}\right),$$

donde

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \qquad \text{e} \qquad \overline{Y}_n = \frac{Y_1 + \cdots + Y_n}{n}.$$

(a) Encuentre el límite en distribución de Z_n, usando el Teorema Central del Límite bivariado aplicado a $(X_1, Y_1), (X_2, Y_2), \ldots$, y el método delta.

(b) Repita (a) usando ahora el hecho que

$$Z_n = \sqrt{n}\left(\frac{\mu_X \overline{Y}_n - \mu_Y \overline{X}_n}{\mu_X \overline{X}_n}\right),$$

y el resultado del ejercicio anterior.

14. Sean $X, X_1, X_2, \ldots$ e $Y_1, Y_2, \ldots$ variables aleatorias verificando $P(X_n = 0) = 0 = P(X = 0)$, $X_n \xrightarrow{D} X$ e $Y_n \xrightarrow{P} c$, donde c es una constante. Demuestre que

$$\frac{Y_n}{X_n} \xrightarrow{D} \frac{c}{X}.$$

15. En el Ejemplo 6.5.6, determine la distribución asintótica de

$$\sqrt{n}\left(\frac{\hat{\sigma}_n}{\overline{Y}_n} - \frac{\sigma}{\mu}\right),$$

suponiendo que $\mu \neq 0$.

Apéndice A

Cálculo Combinatorial

A.1. Introducción y Principios Básicos

El cálculo combinatorial responde preguntas básicas tales como ¿de cuántas maneras se puede hacer algo?, o ¿cuántas configuraciones de cierto tipo hay?. En principio, bastaría con confeccionar un listado exhaustivo y contar sus elementos, pero esto es difícil cuando la naturaleza de los elementos complica su ordenamiento lógico o cuando el listado es demasiado largo. Desde el punto de vista abstracto, se trata de calcular la cardinalidad $\operatorname{card} A$ de un conjunto A, lo que sugiere el siguiente truco útil:

Identificar cada *manera* o *configuración* con los elementos de un conjunto A y calcular $\operatorname{card} A$. Si no podemos calcular directamente esta cardinalidad, ponemos a los elementos de A en correspondencia biunívoca con los de un nuevo conjunto B y encontramos la cardinalidad de B, la que siempre coincide con la de A. En vez de escribir explícitamente el conjunto A, podemos identificar las configuración o maneras con las que aparecen en ciertos problemas combinatoriales clásicos, para los que conocemos su solución.

La acción de contar es básica para definir los números naturales. De ella se deduce la suma y la multiplicación (como suma repetida). No es de sorprender que existan reglas *aditivas* y *multiplicativas* para la resolución de problemas combinatoriales.

$$\operatorname{card} \bigcup_{i=1}^{r} A_i = \sum_{i=1}^{r} \operatorname{card} A_i, \text{ si los } A_i \text{ son disjuntos.} \qquad \text{(A.1.1)}$$

(Regla Aditiva)

Antes de formalizar la regla multiplicativa, es conveniente discutir previamente el concepto de árbol.

Sea $\Omega \subseteq S_1 \times S_2 \times \cdots \times S_k$. Los *arreglos* $\boldsymbol{x} \in \Omega$ se pueden identificar con los caminos que parten de cierto *origen* (*nodo* de orden 0) y pasan consecutivamente por un nodo de orden 1, un nodo de orden 2, y así sucesivamente hasta los nodos de orden k, a los que denominamos nodos finales. Un nodo de orden i se conecta con otro de orden $i+1$ mediante un *arco*. El camino que conecta el origen con un nodo de orden i consiste de i arcos y lo denominamos rama cuando $i = k$, o sea,

cuando el camino conecta el origen con un nodo terminal. Un *árbol* es la colección de todas estas ramas. En el caso general, el número de ramas se obtiene por enumeración directa. Sin embargo, hay un caso importante, donde existe un atajo, cual es el de un árbol *regular*, para el cual el número de arcos que emana de un nodo depende sólo de su orden.

El número de ramas de un árbol regular es

$$n_1 \times n_2 \times \cdots \times n_k, \quad \textit{(Regla Multiplicativa)} \tag{A.1.2}$$

donde n_i es el número de arcos que emana de un nodo de orden i.

Los nodos de orden i reciben una etiqueta $x_i \in S_i$, $i = 1, 2, \ldots, k$. Aunque hay muchos caminos para los cuales el i-ésimo nodo lleva la *etiqueta* x_i, para cada nodo existe un único camino que lo une al origen. De esta forma, el nodo lleva la etiqueta x_i, pero realmente representa, además, a todos los valores previos. En otras palabras, un nodo de orden i contiene a $(x_1, \ldots, x_i)$ como información, lo que resulta muy eficiente.

Los árboles están muy bien adaptados a definiciones recursivas. Para definir recursivamente Ω a partir de $S_1 \times S_2 \times \cdots \times S_k$, se deben indicar los valores que puede tener x_i dados los valores de x_j para $j < i$. En términos del árbol, esto equivale a conocer las posibles etiquetas para un nodo de orden i dada la información contenida en el nodo precedente, es decir, dadas las etiquetas de todos los previos. Denotamos al conjunto de tales valores por $S_i(x_1, \ldots, x_{i-1})$, para $1 < i \leq k$, y usamos la convención que para $i = 1$ este conjunto coincide con S_1.

La regla multiplicativa se puede escribir más formalmente como sigue:

$$\text{card } S_i(x_1, \ldots, x_{i-1}) = n_i,\ i = 1, \ldots, k \Rightarrow \text{ card } \Omega = \prod_{i=1}^{r} n_i. \quad \textit{(Regla multiplicativa para función recursiva)} \tag{A.1.3}$$

Tomando $S_1 = A_1$ y $S_i(x_1, \ldots, x_{i-1}) = A_i$, $i = 2, \ldots, k$ se obtiene

$$\text{card} \prod_{i=1}^{r} A_i = \prod_{i=1}^{r} \text{ card } A_i. \quad \textit{(Regla Multiplicativa Básica)} \tag{A.1.4}$$

donde $\prod_{i=1}^{r} A_i = A_1 \times \cdots \times A_r$, o sea es el producto cartesiano de $A_1, A_2, \ldots, A_r$.

Relaciones de equivalencia:

Dado un conjunto de N elementos y una relación de equivalencia, interesa contar el número t de clases de equivalencia, es decir la cardinalidad de la partición inducida por esta clase. Si la partición es regular, en el sentido que todos los conjuntos tienen igual cardinalidad p, hay obviamente $t = \frac{N}{p}$ clases de equivalencia. Esto puede verse como consecuencia de la regla multiplicativa si construimos un árbol regular en que el nodo primario es la clase de equivalencia y los nodos secundarios corresponden a los elementos de esta clase.

Número de clases de equivalencia

Si las clases de equivalencia tienen la misma cardinalidad, su número se obtiene dividiendo $\text{card}\ \Omega$ por la cardinalidad de una de las clases.

(A.1.5)

A.2. Formulación funcional

Todos los problemas combinatorios clásicos se pueden formular en términos de funciones. Este grado de abstracción es útil para una aplicabilidad amplia de los resultados, pero debe ser complementado con modelos más concretos. Consideramos dos conjuntos $\mathcal{X}$ e $\mathcal{Y}$ de cardinalidades I y J respectivamente, así como el conjunto de funciones $\mathcal{F}$ de $\mathcal{X}$ en $\mathcal{Y}$. Para ciertos propósitos conviene enumerar los elementos, o sea, escribir $\mathcal{X} = \{x_i,\, i = 1, \ldots, I\}$ e $\mathcal{Y} = \{y_j,\, j = 1, \ldots, J\}$.

La función f está en correspondencia uno a uno con la función f^{-1}, que a $y \in \mathcal{Y}$ le asocia el conjunto $\{x/\ f(x) = y\}$. Cuando los elementos de $\mathcal{X}$ son indistinguibles (en un sentido a precisar) se considera la función $g = \text{card}\ f^{-1}$, definida por $g(y) = \text{card}\ f^{-1}(y)$. Interesa calcular cuantas funciones $\text{card}\ f^{-1}$ existen, así como cuantas funciones f satisfacen la condición $\text{card}\ f^{-1} = g$, para una función g especificada previamente. Cuando, además, los elementos de y son indistinguibles, $\text{card}\ g^{-1} = \text{card}\ (\text{card}\ f^{-1})$ es lo relevante ahora e interesa calcular cuantas funciones f son compatibles con una especificación dada de $\text{card}\ (\text{card}\ f^{-1})^{-1}$.

La siguiente tabla entrega las respuestas a varias preguntas, postergando su demostración. Con la notación

$$n! = n \times (n-1) \times \cdots \times 2 \times 1 \tag{A.2.1}$$

$$n^{[k]} = n \times (n-1) \times \cdots \times (n-k+1) \tag{A.2.2}$$

$$\binom{n}{k} = \frac{n^{[k]}}{k!} = \frac{n!}{k!\,(n-k)!}. \tag{A.2.3}$$

$$\binom{n}{n_1\ n_2\ \cdots\ n_r} = \binom{n}{\boldsymbol{n}} = \frac{n!}{n_1! \times n_2! \times \cdots \times n_r!} \tag{A.2.4}$$

los resultados son los siguientes:

Hay J^I funciones de $\mathcal{X}$ en $\mathcal{Y}$. (A.2.5)

Hay $J^{[I]}$ funciones inyectivas de $\mathcal{X}$ en $\mathcal{Y}$. (A.2.6)

Si $J = I$, hay $I!$ funciones biyectivas de $\mathcal{X}$ en $\mathcal{Y}$. (A.2.7)

Hay $\binom{J}{I}$ funciones $\text{card}\ f^{-1}$, cuando f se restringe a ser inyectiva (A.2.8)

Hay $\binom{I}{n_1\ n_2\ \cdots\ n_J}$ funciones de $\mathcal{X}$ en $\mathcal{Y}$, en que n_j elementos de $\mathcal{X}$ son asignados a y_j. (A.2.9)

Hay $\binom{I+J-1}{I}$ funciones $\text{card}\ f^{-1}$. (A.2.10)

La función f^{-1} se representa también como una partición ordenada de $\mathcal{X}$, en otras palabras, $(\{f^{-1}(y_j),\, j = 1, \ldots, J\}$. Por otro lado, los resultados de la tabla anterior se pueden especializar al caso $\mathcal{Y} = \mathcal{X}$. Como subproducto de esta tabla se obtiene:

Hay	$n!$	maneras de ordenar un conjunto de n elementos (por (A.2.7)).	(A.2.11)
Hay	$n!$	permutaciones sobre un conjunto de n elementos (por (A.2.7)).	(A.2.12)
Hay	$\binom{J}{I}$	subconjuntos de tama no I de un conjunto de tama no J (por (A.2.7)).	(A.2.13)
Hay	J^I	particiones ordenadas de $\mathcal{X}$, que constan de a lo más J términos (por (A.2.5)).	(A.2.14)
Hay	$\binom{n}{I_1\ I_2 \cdots\ I_r}$	particiones ordenadas de $\mathcal{X}$, con tama nos pre-especificados I_r, $r = 1, \ldots, R$ (por (A.2.9)).	(A.2.15)

Cuando $R = 2$ en (A.2.9), especificar la partición equivale a especificar a uno de los conjuntos que la conforma. En consecuencia el número de particiones en dos clases, de cardinalidades I_1 e I_2, coincide con el número de combinaciones de tama nos I_1 o I_2 de un conjunto de tama no $I_1 + I_2$. Comparando (A.2.7) y (A.2.9) se obtiene una demostración de la conocida identidad

$$\binom{m_1 + m_2}{m_1\ m_2} = \binom{m_1 + m_2}{m_1} = \binom{m_1 + m_2}{m_2}$$

Un ejemplo interesante es el de barajar un naipe de 52 cartas, las que se reparten equitativamente entre 4 jugadores y *se supone que el orden en que le hayan llegado las cartas a un jugador es irrelevante*. Se puede visualizar esto como una función que a cada carta le asigna un jugador, quedando pre-establecido que a cada jugador le tocan 13 cartas. Por (A.2.9), hay

$$\frac{52!}{13! \times 13! \times 13! \times 13!}$$

reparticiones posibles. Si se prefiere, podemos enumerar a los jugadores y pensar en una partición ordenada, donde cada clase en la partición es el conjunto de cartas asignadas a un jugador. De esta forma, el número de reparticiones posibles se obtiene de (A.2.15).

La función f puede construirse asignando secuencialmente valores a $x_1, x_2, \ldots$. De (A.1.3) se deducen (A.2.5) y (A.2.6). Eligiendo $J = I$ se obtiene (A.2.7). Para demostrar (A.2.9) utilizamos *permutaciones* sobre $\mathcal{X}$, o sea, biyecciones de $\mathcal{X}$ sobre si mismo. Sea $\mathcal{G}$ el conjunto de funciones que cumple la condición en (A.2.9). Dado $f_0 \in \mathcal{G}$, toda $f \in \mathcal{G}$ puede encontrarse mediante alguna permutación T, a través de

$$f(x) = f_0(T(x)).$$

Sin embargo, algunas permutaciones dejan f_0 invariante, o sea, $f = f_0$. Ellas son aquellas que reordenan los elementos dentro de cada conjunto $f^{-1}(y_j)$. Por hipótesis $\operatorname{card} f^{-1}(y_j) = n_j$, de modo

que hay $n_j!$ permutaciones dentro de cada conjunto. Por la regla multiplicativa, hay exactamente

$$\prod_{j=1}^{J} n_j!$$

que dejan invariantes a f_0, sin importar cual sea esta función. Dado que toda función en $\mathcal{G}$ es obtenible a partir de f_0 mediante alguna de las $I!$ permutaciones, se obtiene (A.2.9). Para demostrar (A.2.8), basta darse cuenta que $\text{card } f^{-1}$ es una función con dominio $\mathcal{Y}$, I de cuyos valores son 1 y el resto, es decir, $J-I$, son 0. Por (A.2.9) hay $\binom{J}{I\ J-I} = \binom{J}{I}$ tales funciones. Como $\text{card } f^{-1}$ está determinada por el conjunto $f(\mathcal{X})$, cuya cardinalidad es I, lo anterior demuestra (A.2.13). Una demostración alternativa descansa en el hecho que entre las $J!$ permutaciones sobre $\mathcal{Y}$, aquellas que dejan invariante una función f, son las que se reducen a permutar los $J-I$ elementos del complemento de $f(\mathcal{X})$. Esto demuestra que hay $\frac{J^{[I]}}{(J-I)!}$ tales funciones, lo que coincide con $\binom{J}{I}$.

Postergamos la demostración de (A.2.10).

A.3. Arreglos y combinaciones

A.3.1. Arreglos

Un *arreglo de largo* k es una k-tupla ordenada, $\boldsymbol{y} = (y_1, \ldots, y_I)$, o bien, un objeto semejante que sólo difiere en aspectos notacionales. Por ejemplo, (a, b, c) y abc son ambos llamados arreglos. Cuando y_i recorre Ω_i, el conjunto de todos los arreglos es el producto cartesiano

$$\text{Arr} = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_k.$$

Por (A.1.3) o (A.1.4) su cardinalidad es

$$\text{card Arr} = n_1 \times n_2 \times \cdots \times n_k, \tag{A.3.1}$$

donde $n_i = \text{ card } \Omega_i$. Nos restringimos en lo sucesivo al caso $\Omega_i = \mathcal{Y}$, donde $\text{card } \mathcal{Y} = J$, de modo que $\Omega = \mathcal{Y}^I$.

Decimos que $\boldsymbol{y}$ es un *arreglo de largo I con elementos (letras) en el conjunto (alfabeto) $\mathcal{Y}$*. Denotamos a $\mathcal{Y}^I$ por Arr_{IJ}. Un arreglo en Arr_{IJ} equivale a la función f con dominio $\mathcal{X} = \{1, 2, \ldots, I\}$ y recorrido $\mathcal{Y}$, definida por $f(j) = y_j$. Los resultados conocidos en el contexto funcional pueden aplicarse directamente. Como en problemas combinatoriales sólo importa la cardinalidad J de $\mathcal{Y}$, se habla de arreglos de largo I de un conjunto con J elementos. Por (A.2.5), la cardinalidad de Arr_{IJ} es J^I.

Cuando las letras no se repiten se habla de *arreglos sin repetición*. Un elemento del conjunto Arr-srep$_{IJ}$ de tales arreglos corresponde a una función inyectiva, de modo que hay $J^{[I]}$ tales arreglos (por (A.2.5)).

Comentario notacional: Los arreglos que no pertenecen a Arr-srep$_{IJ}$ se caracterizan por tener letras repetidas. Aunque es desafortunado desde el punto de vista lógico, se acostumbra usar el término *arreglos con repetición* para referirse a todos los elementos de Arr_{IJ}, incluyendo así a aquellos sin repetición, como caso particular. Por esta razón usamos indistintamente las notaciones Arr-crep$_{IJ}$ y Arr_{IJ}.

A.3.2. Combinaciones

La idea básica es no distinguir entre dos arreglos que difieran sólo en el orden en que aparecen sus letras. Identificando al arreglo con una función, esta idea corresponde a declarar equivalentes a las funciones g y f si existe una permutación T tal que $g(x) = f(T(x)) = f(x)$. Las clases de equivalencia correspondientes se denominan *combinaciones* y heredan el apelativo de con o sin repetición, según si se aplican a Arr-crep$_{IJ}$ o a Arr-srep$_{IJ}$. Denotamos los conjuntos de combinaciones por Comb-crep$_{IJ}$ y Comb-srep$_{IJ}$, respectivamente. *Si no se especifica el tipo, se subentiende que la combinación es sin repetición.* La cardinalidad de card Comb-crep$_{IJ}$ y card Comb-srep$_{IJ}$ se obtienen directamente de (A.2.10) y (A.2.8), respectivamente. Las fórmulas correspondientes se muestran en la Tabla A.3.1.

	arreglos	combinaciones
con repetición	card Arr-crep$_{IJ} = J^I$	card Comb-crep$_{IJ} = \binom{I+J-1}{I}$
sin repetición	card Arr-srep$_{IJ} = J^{[I]}$	card Comb-srep$_{IJ} = \binom{J}{I}$

Cuadro A.3.1: Fórmulas para arreglos y combinaciones

Aunque la formulación funcional genere las fórmulas, es interesante examinar cómo se obtiene la combinación asociada a un arreglo dado. Reordenamos las componentes del arreglo en *orden alfabético* (una enumeración arbitraria de $\mathcal{Y}$) generando un *arreglo estándar.* La combinación asociada es el conjunto de arreglos que se pueden obtener reordenando sus letras y se puede representar por el arreglo estándar. El número de arreglos estándar coincide con el número de combinaciones. Por ejemplo, si $\mathcal{Y} = \{c, a, d, e, b\}$ y se usa el orden alfabético habitual, el arreglo estándar correspondiente a (c, a, c, b, c, a) es (a, a, b, c, c, c). Suprimiendo comas y paréntesis, la combinación correspondiente a $cacbca$ se denota por $aabccc$. La combinación asociada a un arreglo queda determinada por el conjunto de letras que aparecen en el arreglo y por el número de veces que aparece cada una de esas letras. Si asignamos el número 0 a un elemento ausente del arreglo, basta conocer cuántas veces aparece cada elemento de $\mathcal{Y}$. En la notación funcional, esto corresponde a la función $\mathrm{card}\ f^{-1}$, que puede representarse por el arreglo $\boldsymbol{m} = (m_1, m_2, \ldots, m_J)$. Si los y_j fueran letras, el producto

$$\prod_{j=1}^{J} y_j^{m_j}$$

no depende del orden de los factores. Extendiendo esta notación exponencial a elementos y_j cualesquiera, de modo que el exponente indica cuantas veces se repite la letra correspondiente, se elimina implícitamente el orden de las letras y se evita enumerar los elementos de $\mathcal{Y}$. Por ejemplo, para $\mathcal{Y} = \{a, b, c, d, e\}$, el arreglo $cacbca$ genera la combinación $a^2bc^3d^0e^0 = a^2b^1c^3$. Las cardinalidades de

$$NCREP = \{\boldsymbol{m}/m_i \geq 0,\ i = 1, \ldots, n,\ \sum_{i=1}^{n} m_i = k\}, \tag{A.3.2}$$

y

$$NSREP = \{\boldsymbol{m}/m_i \in \{0,1\},\ i = 1,\ldots,n,\ \sum_{i=1}^{n} m_i = k\} \tag{A.3.3}$$

coinciden con las de Comb-crep$_{IJ}$ y Comb-srep$_{IJ}$ respectivamente.

Para calcular el número de combinaciones con repetición, asociamos a $\boldsymbol{m}$ un vector binario $\boldsymbol{n}$ más largo, el cual se puede poner en correspondencia con una combinación. La regla que asigna $\boldsymbol{n}$ a $\boldsymbol{m}$ es:

- Se parte de un arreglo con I ceros.
- Se intercalan m_j unos antes del j-ésimo cero.
- Se elimina el último elemento de este arreglo (que es siempre cero), con lo que se obtiene un vector binario de largo $I + J - 1$, que contiene I unos.

Esto demuestra (A.2.10).

Por ejemplo, si $\mathcal{Y} = \{a, b, c, d\}$, la combinación con repetición a^2bc^3 genera $\boldsymbol{m} = (2, 1, 3, 0)$, al cual se le asocia $\boldsymbol{n} = (1, 1, 0, 1, 0, 1, 1, 1, 0)$. Por su parte, $\boldsymbol{n} = (0, 1, 1, 1, 1, 0, 1, 1, 0)$ sólo puede provenir de $\boldsymbol{m} = (0, 4, 2, 0)$, o sea de la combinación b^4c^2.

A.4. Modelo de ocupación: bolas en casilleros

Interesa calcular cuántas *distribuciones* de I bolas en J casilleros hay, o sea, de cuántas maneras pueden I bolas ocupar J casilleros. Tomando $\mathcal{X}$ como el conjunto de bolas e $\mathcal{Y}$ como el conjunto de casilleros, una distribución parece corresponder a una función f. La función $f^{-1}(y)$ especifica *cuáles* bolas *caen* en un casillero, mientras que $\text{card } f^{-1}$ especifica *cuántas* bolas caen en cada uno. La primera es relevante cuando las bolas son *distinguibles* (bolas de distintos colores), la segunda cuando ellas son *indistinguibles* (bolas del mismo color). Otra consideración relevante es si se permiten o no **múltiples** bolas en algún casillero. Una respuesta negativa, que sólo es posible si $I \leq J$, corresponde a una función inyectiva. Se dice que la asignación de bolas a los casilleros es *con exclusión*. En el caso especial $I = J$, una distribución con exclusión corresponde a una biyección. Cuando las bolas son distinguibles, las distribuciones posibles difieren sólo en el orden en que aparecen las bolas en los casilleros. Esto ofrece una manera de modelar concretamente una permutación.

Para generalizar la noción de distinguibilidad y ponerla en un marco abstracto, conviene definir una función W sobre $\mathcal{X}$, que representa el *color* que tiene cada bola. Si sólo podemos distinguir colores distintos, lo relevante es cuántas fichas de cada color caen en cada uno de los casilleros. La versión abstracta es que el *invariante maximal* es

$$(\ \text{card } f^{-1}(\{y\} \cap \mathcal{X}_r),\ r = 1,\ldots,R),$$

donde $\mathcal{X}_r$ está formado por las bolas del r-ésimo color.

Consideramos ahora el caso donde el número de bolas coincide con el número de casilleros. Dos asignaciones son equivalentes si el color asignado a cada casillero (por la bola que lo ocupa) es el

mismo para ambas. Si I_r es el número de bolas de color w_r, i.e. $\operatorname{card} \mathcal{X}_r$, cada clase de equivalencia equivale a pintar los casilleros, de modo que I_r de ellos queden de color w_r, para $r = 1, \ldots, R$. Por (A.2.9), el número de clases de equivalencia es

$$\binom{I}{I_1\ I_2\ \cdots\ I_r} = \frac{I!}{I_1! \times I_2! \times \cdots \times I_R!}. \tag{A.4.1}$$

A.5. Modelo de Urna

La extracción de una muestra de I individuos de una población de tama no J equivale, combinatorialmente, a la extracción de I fichas de una urna que contiene J fichas. El proceso de selección de la muestra se denomina *muestreo*. El número de muestras depende tanto de la forma del muestreo, como de decisiones de carácter lógico. El muestreo se puede considerar como una función que a cada extracción le asigna el elemento de la urna que aparece en ella. Habitualmente, las extracciones están numeradas de 1 a I y se escribe el resultado del muestreo como un arreglo de largo I.

- Si se distingue entre dos muestras que sólo difieran en el orden, decimos que la muestra es *ordenada*. Cuando se extrae un pu nado de k fichas, el orden parece no existir, pero es concebible que el observador vea a unas primero y otras después. En este sentido el tomar en cuenta o no el orden es una decisión independiente del muestreo.
- Si una ficha extraida de la urna puede aparecer en futuras extracciones, decimos que el muestreo o la muestra es *con reposición*. En vez de reposición se suele usar palabras como devolución, restitución, o reemplazo. Una muestra sin reposición se puede describir indicando *cuáles* fichas aparecen en la muestra y *en qué orden* están.

Para $I = J$ aparecen todas las fichas, de modo que el número de muestras coincide con el número de maneras de permutar las I fichas de la urna.

Identificando cada extracción con una bola, y cada elemento de la población con un casillero, se establece una correspondencia entre dos problemas combinatoriales. También es posible comparar el modelo de urnas con el modelo de arreglos y combinaciones, o con el modelo abstracto. El concepto de orden es idéntico al usado con los arreglos y combinaciones y corresponde a la distinguibilidad de bolas en el modelo de bolas y casilleros. El muestreo sin reposición corresponde a la ausencia de letras repetidas, a no tener ocupaciones múltiples en los casilleros, o a la inyectividad de la función. Se establece así una correspondencia entre el problema de contar arreglos y combinaciones con el de contar muestras o distribuciones de bolas. Esto explica que las fórmulas obtenidas sean idénticas, difiriendo los enunciados de cada problema excepto por los encabezamientos de cada tabla.

A.6. Permutaciones y coeficientes multinomiales

A.6.1. Permutaciones

Una permutación se puede considerar como un caso particular de arreglo sin repetición o como una biyección T de un conjunto sobre si mismo. Se la puede especificar directamente o a través de T^{-1}. Es posible transformar una función inyectiva g de $\mathcal{X}$ en $\mathcal{Y}$, en una biyección agregando a $\mathcal{X}$ $J-I$ elementos para constituir un nuevo conjunto $\mathcal{X}'$. Si declaramos a dos biyecciones de $\mathcal{X}'$ en $\mathcal{Y}$ equivalentes cuando coinciden para $x \in \mathcal{X}$, la clase de equivalencia está en correspondencia biunívoca con el arreglo original. Dado que esta clase se genera permutando las últimas $J-I$ componentes del arreglo, ella tiene cardinalidad $(J-I)!$. Por (A.1.5), el número de clases es $\frac{J!}{(J-I)!}$, lo que coincide con (A.2.8). Para ilustrar la idea, tomemos $\mathcal{Y} = \{1,2,3,4,5,6,7\}$, $I = 4$ y el arreglo 2361745. Los arreglos equivalentes son 2361745, 2361754, 2361475, 2361457, 2361574 y 2361547 y están en correspondencia con 745, 754, 475, 457, 574 y 547 respectivamente. Estas últimas son las $3! = 6$ permutaciones de $\{4,5,7\}$. En general, los $7! = 5040$ arreglos de largo 7 se agrupan en clases de tama no 6, de modo que hay $\frac{5040}{6} = 840$ tales clases y 840 arreglos sin repetición de largo 4.

El resultado de barajar el naipe es un arreglo $(y_1, y_2, \ldots, y_I)$ que corresponde a una permutación de $(1,2,\ldots,I)$. Sea T la permutación sobre $\{1,2,\ldots,I\}$ definida por $T(y_j) = j$. Si nos concentramos, en cambio, en una carta particular i, y especificamos en qué posición queda ella, tenemos $j = T(i)$. En otras palabras, podemos trabajar tanto con la biyección T como con su inversa. Ambas son permutaciones.

A.6.2. Aplicaciones de los coeficientes multinomiales

Supongamos que se lanza 10 veces un dado, obteniéndose secuencialmente los números 2, 1, 4, 2, 2, 3, 5, 4, 2 y 4. Mirando esto como una función f de $\{1,2,\ldots,10\}$ en $\{1,2,\ldots,6\}$, el arreglo $\boldsymbol{m}$ correspondiente a $\text{card } f^{-1}$ es $\boldsymbol{m} = (1,4,1,3,1,0)$. Por (A.2.9), el número de resultados de los 10 lanzamientos del dado, para los cuales se obtiene este valor de $\boldsymbol{m}$, coincide con

$$\binom{10}{1\ 4\ 1\ 3\ 1\ 0} = 25200.$$

Esto se puede generalizar a I lanzamientos de un dado imaginario de J caras.

Una aplicación importante es el *Teorema del multinomio:*

$$\left(\sum_{j=1}^{k} x_i\right)^m = \sum_{m_1,\ldots,m_k} \binom{m}{m_1\ m_2\ \cdots\ m_k} \prod_{j=1}^{k} x_j^{m_j} \tag{A.6.1}$$

Para $k=2$, con $a = x_1$, $x_2 = b$, $m_1 = i$, $m_2 = j$ y $m = n$ se obtiene el *Teorema del Binomio:*

$$(a+b)^n = \sum_{i+j=n} \binom{n}{i\ j} a^i b^j = \sum_{i=1}^{n} \binom{n}{i} a^i b^{n-i} \tag{A.6.2}$$

Tomando los términos x_i, a y b iguales a 1 se obtienen las importantes identidades

$$\sum_{m_1,\ldots,m_k} \binom{m}{m_1\ m_2\ \cdots\ m_k} = k^m, \tag{A.6.3}$$

y

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n. \tag{A.6.4}$$

En vez de hacer la demostración formal, examinemos cómo se obtiene la expansión de $(x_1 + x_2 + x_3 + x_4)^5$. Se tiene

$$\begin{aligned}(x_1+x_2+x_3+x_4)^5 \ &= \ (x_1+x_2+x_3+x_4)\times(x_1+x_2+x_3+x_4)\times(x_1+x_2+x_3+x_4)\\ &\quad\times(x_1+x_2+x_3+x_4)\times(x_1+x_2+x_3+x_4)\end{aligned}$$

La expansión consiste en

- Elegir un término de cada paréntesis (4 opciones).
- Multiplicarlos.
- Simplificar el monomio resultante.
- Repetir para las 4^5 elecciones posibles.
- Escribir la suma de los 4^5 términos como una combinación lineal de los distintos monomios.
- Determinar el coeficiente de un monomio dado.

Sea x_{r_i} el término elegido en el i-ésimo paréntesis, y sea $\boldsymbol{r} = (r_1, r_2, r_3, r_4, r_5)$. Por ejemplo, $\boldsymbol{r} = (2, 1, 2, 3, 3)$ corresponde al producto $x_2x_1x_2x_3x_3 = x_1x_2^2x_3^2$. Aprovechando la invarianza de la multiplicación bajo permutaciones de los factores, podemos seguir un camino más corto para encontrar el coeficiente $c(m_1, \ldots, m_k)$ que acompa na a

$$\prod_{j=1}^{k} x_j^{m_j} \tag{A.6.5}$$

en la expansión de

$$\left(\sum_{j=1}^{k} x_i\right)^m$$

(el coeficiente de $x_1x_2^2x_3^2$ sería $c(1, 2, 2, 0, 0)$ bajo esta notación). En efecto, denotemos por A al conjunto del paréntesis y por A_j al subconjunto formado por aquellos en que se selecciona el término x_j. Para obtener el monomio (A.6.5), debemos elegir estos subconjuntos de tal forma que se satisfaga la restricción $\operatorname{card} A_j = m_j$. El número de elecciones posibles es el número de particiones ordenadas con clases de tama nos $m_1, m_2, \ldots, m_k$. Por (A.2.14) se obtiene (A.6.1).

Si partimos de los arreglos $\boldsymbol{r} = (r_1, \ldots, r_m)$, declaramos que dos arreglos son equivalentes si dan origen al mismo monomio. Si el monomio está especificado por (A.6.5), para valores dados de los m_i, la equivalencia queda caracterizada por

$$\text{card } \{i/r_i = x_j\} = m_j,\ j = 1, \ldots, k. \tag{A.6.6}$$

De esta forma, el coeficiente de $x_1 x_2^2 x_3^2$ es

$$\binom{5}{1\ 2\ 2\ 0\ 0} = \frac{120}{1 \times 2 \times 2 \times 1 \times 1} = 30.$$

Por (A.3.2), el número de términos distintos en la expansión del multinomio coincide con el número de combinaciones con repetición de largo k de un conjunto de m elementos, es decir hay

$$\binom{m+k-1}{k}$$

términos distintos en la expansión del multinomio.

A.7. Resumen: Equivalencia de modelos.

Se presentan acá las fórmulas obtenidas, poniendo en paralelo los diversos modelos discutidos.

- J^I
  Función arbitraria de $\mathcal{X}$ en $\mathcal{Y}$ con $\text{card } \mathcal{X} = I,\ \text{card } \mathcal{Y} = J$.
  Arreglo con repetición de largo I con un alfabeto de J letras.
  Distribuir I bolas distinguibles en J casilleros.
  Muestra con reposición, de tama no I, extraida de una población de tama no J.
  Lanzar I veces un dado de J caras.

- J^I
  Función inyectiva de $\mathcal{X}$ en $\mathcal{Y}$ con $\text{card } \mathcal{X} = I,\ \text{card } \mathcal{Y} = J$.
  Arreglo sin repetición de largo I con un alfabeto de J letras.
  Distribuir I bolas distinguibles en J casilleros, con exclusión.
  Muestra sin reposición, de tama no I, extraida de una población de tama no J.

- $\binom{J}{I}$
  $f(\mathcal{X})$ para una función inyectiva de $\mathcal{X}$ en $\mathcal{Y}$, con $\text{card } \mathcal{X} = I,\ \text{card } \mathcal{Y} = J$.
  Combinación sin repetición de largo I con un alfabeto de J letras.
  Distribuir I bolas indistinguibles en J casilleros, con exclusión.
  Muestra no ordenada, sin reposición, de tama no I, extraida de una población de tama no J.

- $\binom{I+J-1}{I}$
  $\text{card } f^{-1}$ para una función arbitraria de $\mathcal{X}$ en $\mathcal{Y}$, con $\text{card } \mathcal{X} = I,\ \text{card } \mathcal{Y} = J$.
  Particiones ordenadas de a lo más J clases, de un conjunto de tama no I.
  Combinación con repetición de largo I con un alfabeto de J letras.
  Distribuir I bolas indistinguibles en J casilleros, sin exclusión.
  Muestra no ordenada, con reposición, de tama no I, extraida de una población de tama no J.

- $n!$
  Biyecciones de un conjunto de tama no n sobre si mismo.
  Ordenes en un conjunto de tama no n.
  Palabras que usan exactamente una vez cada una de n letras.
  Distribuciones de n bolas de colores en n casilleros, con una bola en cada casillero.
  Maneras de barajar un naipe de n cartas.

- $\binom{\sum_r n_r}{n_1\ n_2\ \cdots\ n_k}$
  Funciones con n_r elementos asignados al r-ésimo elemento de $\mathcal{Y}$.
  Particiones ordenadas, con clases de tama nos $n_1, n_2, \ldots, n_k$ clases, de un conjunto de tama no I.
  Número de palabras distintas que usan n_r veces la r-ésima letra de un alfabeto de k letras.
  Distribuciones de bolas de colores, donde n_r representa el número de ellas que poseen el r-ésimo color.
  Muestras ordenadas de tama no n en que el r-ésimo elemento de una población de tama no k
  aparece r veces.
  Repartir un naipe de n cartas asignando n_r cartas al r-ésimo jugador.
  Coeficiente de $\prod_r x_r^{n_r}$ en la expansión de $(\sum_{r=1}^{k} x_r)^n$, con $n = \sum_r n_r$.