GEOMETRÍA E INFORMACIÓN

OPTATIVO

Mariela Adelina Portesi
Pedro Walter Lamberti
Steeve Zozor

Versión completa del 20 de agosto de 2019

Facultad de Ciencias Exactas

UNIVERSIDAD NACIONAL DE LA PLATA

Editorial de la Universidad de La Plata

Esto es una dedicatoria
del libro.

Agradecimientos

Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla. Este es el texto de agradecimiento, max una carilla.

Esto es un epígrafe con texto simulado.
Esto es un epgrafe con texto simulado.
AUTOR DEL EPÍGRAFE, TÍTULO DE LA OBRA

PRÓLOGO

Este libro surge de la experiencia de los autores en el dictado del curso semestral “Métodos de geometría diferencial en teoría de la información”, que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. **. . . acabar**

Los autores

ADVERTENCIA

Este libro surge de la experiencia de los autores en el dictado del curso semestral “Métodos de geometría diferencial en teoría de la información”, que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. **. . . acabar**

Mariela A. Portesi

Grenoble, Junio de 2016

Índice

CAPÍTULO 1
Elementos de teoría de probabilidades

While writing my book I had an argument with Feller.
He asserted that everyone said "random variable"
and I asserted that everyone said "chance variable."
We obviously had to use the same name in our books,
so we decided the issue by a stochastic procedure.
That is, we tossed for it and he won.
J. L. DOOB, STATISTICAL SCIENCE (1953)

Anadir por lo menos un poco

- **Las nociones de convergencia (aparece en casos limites de leyes, en el TCL). Ver (Ash & Doléans-Dade, 1999, Cap. 2.8, 6), (Billingsley, 2012, Cap. 5), (Athreya & Lahiri, 2006, Sec. 9, p. 287), (Brockwell & Davis, 1987, Cap. 6), (Jacob & Protters, 2003, Caps. 17, 18).**
- **Definicion de los cumulentes (ver momentos)**
- **Cotas de Chernoff con la MGF o PGF**
- **Descomposición de Bartlett caso matriz variado**
- **Dibujar en el plano complejo Φ_X? Unas curvas son lindas.**
- **Ejemplos: von Mises y vonMises-Fisher? Cantor (singular...)?**
- **Hablar de simulación (inversion OK; mezcla? rejección? multivariada via la condicional?) VEr Kotz vol 2 por ejemplo**

1.1 Introducción

A pesar de que las nociones de azar (que proviene del árabe *zahr* que significa dado, flor) o de aleatoriedad (del latín *alea* que es suerte, dado) son muy antiguas (Serrano Marugán, 2000), el matemático italiano y jugador de dados y cartas Gerolamo Cardano es "probablemente" uno de los primeros en tratar matemáticamente el concepto de probabilidad en el siglo XVI, en su libro sobre los juegos de azar escrito en 1564 pero publicado en 1663 (Cardano, 1663) (ver (Bellhouse, 2005) o (Hald, 1990, Cap. 4)). La denominación de probabilidad, ella, viene de Aristote y designaba una percepción de una idea. Tomó su sentido más actual solamente durante la edad media en Europa, por mala traducción de la escritura de Aristote. Después de la primeras semillas, debido a Cardano, hay que mencionar los franceses Pierre de Fermat y Blaise Pascal en el medio del siglo XVII (Pascal, 1679) o (Hald, 1990, Cap. 5), y el neerlandese C. Huygens (Huygens, 1657) o (Hald, 1990, Cap. 6)), que fueron claramente unos de los primeros a desarrollar la teoría de las probabilidades. Más tarde, pasos importante fureon debidos al suizo Jacob Bernoulli (miembro de una dinastía de matemáticos) (Bernoulli, 1713, en latín) o ((E. D. Sylla, Translator), 1713; Hald, 1990, 2006) y al franco-inglés Abraham de Moivre (de Moivre, 1756; Hald, 1990, 2006). Hasta la época de Bernoulli, el enfoque era puramente discreto, es decir que el conjunto de estados posibles era discreto de tamaño finito (6 caras de un dado, 32 tarjetad, 2 caras de una moneda,...). La meta de la mayoría de los estudios eran dedicados a los juegos (dados, cartas), problema de seguro/riesgo, o estudios sociales en problaciones.

El francés Pierre Simon Laplace (de Laplace, 1820) fue quizás uno de los primeros en proveer un aporte importante al desarrollo de la teoría de las probabilidades en los siglos XVIII-XIX, a través del punto de vista "frecuentista" y combinatorial (ver también (Hald, 1990, Caps. 13, 15 & 22) o (Hald, 2006)). En la misma época, hay que mencionar C. F. Gauss, matemático muy prolífico, quien trabajó, entre muchas cosas, en la predicciíon de la trayectoría del planetisimo Céres (Gauss, 1809, 1810) o (Hald, 2006, Cap. 7). Proponiendo un error cuadratico, apareció implicitamente la ley Normal, o Gausiana, que tiene su nombre, a pesar de que la desarollo más Laplace (aún que, sobre el mismo problema, propuso el un error tipo L^1, vínculado a la ley doble-exponencial o de Laplace) (Laplace, 1809a, 1809b; de Laplace, 1820). A veces la ley de Gauss, quizas la más importante en la teoría de las probabilidades, llamada también gausiana o normal, es conocida como ley de Laplace-Gauss.

Un paso muy importante, especialmente tratanto de aleatoriadidad continua (ej. medida de una velocidad, que puede tomar cualquier valor real si tomamos en cuenta la dirección), fue debido entre otros a Kolmogorov en 1933 que se apoyó sobre trabajos de Richard von Mises (von Mises, 1932) y también sobre la teoría de la medida y de la integración, debidas entre otros a Émile Borel y Henri Lebesgue (Borel, 1898, 1909; Lebesgue, 1904, 1918; Halmos, 1950), para formalizar analíticamente la teoria de las probabilidades (Kolmogorov, 1956; Barone & Novikoff, 1978; Jacob & Protters, 2003). Este punto de vista permite tratar formalmente el caso de variables discretas, continuas, o mezcal de ambas, que sean escalar o multivariada, en un marco único y muy poderoso, sin perder las intuición que lleva el punto de vista frecuencista.

1.2 Probabilidades

El concepto de *probabilidad* es importante en situaciones donde el resultado de un dado proceso o medición es incierto, cuando la salida de una experiencia no es totalmente previsible. La probabilidad de un evento es una medida que se asocia con cuán probable es el evento o resultado.

Una definición de probabilidad se puede dar en base a la enumeración exhaustiva de los resultados posibles de un experimento o proceso, suponiendo que el conjunto de posibilidades es completo en el sentido de que una de ellas debe ocurrir o debe ser verdad. Si el proceso tiene K resultados distinguibles, mutuamente excluyentes e igualmente probables (esto es, no se prefiere una posibilidad frente a otras), y si k de esos K resultados tienen un dado atributo, la probabilidad asociada a dicho atributo en un dado proceso es $\frac{k}{K}$. Por ejemplo, sorteando un número entre los naturales del 1 al 10, la probabilidad de "obtener un número par" es $\frac{5}{10} = \frac{1}{2}$.

Otra definición de probabilidad se basa en la frecuencia relativa de ocurrencia de un evento. Si en una cantidad K muy grande de procesos independientes cierto atributo aparece k veces, se identifica a la probabilidad asociada a un proceso o ensayo con la frecuencia relativa de ocurrencia $\frac{k}{K}$ del atributo (van Brakel, 1976; Hald, 1990; Shafer & Vovk, 2006, & Ref.).

Los axiomas de Kolmogorov proveen requisitos suficientes para determinar completamente las propiedades de la medida de probabilidad $P(A)$ que se puede asociar a un evento A entre un conjunto de resultados o eventos de un proceso.

Llamemos Ω al *espacio muestral* o *espacio fundamental*, que es el espacio de *muestras* (*outcomes*, en inglés) $\omega \in \Omega$. Se asocia $\mathcal{A}$ a una colección de sub-conjuntos de Ω, donde los elementos de $\mathcal{A}$ son llamados *eventos*. Por ejemplo, para un dado de 6 caras, Ω es el conjunto de caras que se pueden etiquetar con los números naturales del 1 al 6 (o también con las letras *a, b, c, d, e, f*, u otro etiquetado), y $\mathcal{A}$ tiene los eventos A "es un número natural par" y B "es un número natural impar". En el caso de analizar el tiempo de vida de un aparato, $\Omega \equiv \mathbb{R}_+$. El conjunto de resultados posibles se supone conocido, aún cuando se desconozca de antemano el resultado de una prueba.

Entre los eventos se pueden considerar operaciones y definiciones análogas a las de la teoría de conjuntos (ver entre otros (Spiegel, 1976; Brémaud, 1988; Mandel & Wolf, 1995; Sierpiński, 1975, 1976; Borel, 1898, 1909)):

- Combinación o unión de eventos: $A \cup B$ implica que se da A, ó B, o ambos (por ejemplo, para un dado de 6 caras, si A son los eventos "cara par" y B los eventos "cara menor o igual a 3", resulta $A \cup B = \{1\,;\,2\,;\,3\,;\,4\,;\,6\}$). Según la literatura, se denota a veces $A + B$ o $A \vee B$.

- Intersección de eventos: $A \cap B$ implica que se dan ambos A y B (en el ejemplo precedente, $A \cap B = \{2\}$). Se denota a veces (A, B) o $A \wedge B$.

- Complemento de un evento: $\bar{A}$ indica que no se da A. Se denota a veces $-A$ o $A^{\complement}$ (en el

ejemplo precedente, $\bar{A} = \{1\,;\,3\,;\,5\}$).

- Eventos *disjuntos* o *mutuamente excluyentes* o *incompatibles*: son aquellos que no se superponen, se anota $A \cap B = \emptyset$ donde $\emptyset = \bar{\Omega}$ denota el *evento nulo* (evento que no puede ocurrir, es el complemento de Ω). Por ejemplo los eventos "cara par" y "cara impar" son incompatibles.
- Denotaremos también $A \setminus B$ cuando el evento A se realiza pero no B. Se lo denota también $A - B$, que es también $A \cap \bar{B}$ (en el ejemplo precedente, $A \setminus B = \{4\}$).

Esto es ilustrado en la Fig. 1-1 empleando lo que se conoce como diagramas de Venn [1]. La unión y la intersección de eventos satisfacen las mismas reglas que en la teoría de conjuntos, es decir cada una es conmutativa $A \cup B = B \cup A$, $A \cap B = B \cap A$, asociativa $(A \cup B) \cup C = A \cup (B \cup C)$, $(A \cap B) \cap C = A \cap (B \cap C)$, distributiva con respecto a la otra $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (ver por ejemplo (Jeffrey, 1948, 1973; Halmos, 1950; Feller, 1971; Brémaud, 1988; Mandel & Wolf, 1995; Ibarrola, Pardo & Quesada, 1997; Lehmann & Casella, 1998; Athreya & Lahiri, 2006; Cohn, 2013; Hogg, McKean & Craig, 2013)).

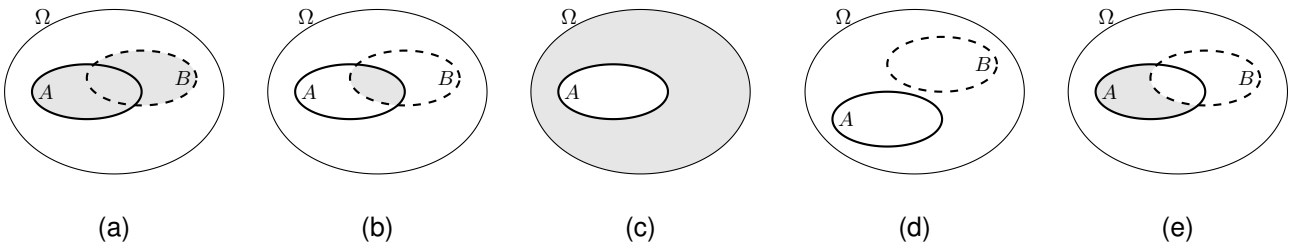


Figura 1-1: Ilustración de las operaciones entre eventos: (a) unión $A \cup B$, (b) intersección $A \cap B$, (c) complemento $\bar{A}$, (d) eventos excluyentes $A \cap B = \emptyset$, y (e) $A \setminus B$. A es representado en línea llena, B en línea discontinua; en (a)-(c) y (e), el resultado de la operación es la zona sombreada.

Formalmente, se define de manera abstracta un espacio medible $(\Omega, \mathcal{A})$ de la manera siguiente ((Halmos, 1950; Feller, 1968, 1971; Brémaud, 1988; Ibarrola et al., 1997; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013); ver también (Barone & Novikoff, 1978; Borel, 1898; Sierpiński, 1918, 1975, 1976, & Ref.) para notas históricas):

Definición 1-1 (Espacio medible)**.** $(\Omega, \mathcal{A})$*, formado por un espacio muestral* Ω *y una colección* $\mathcal{A}$ *de conjuntos de* Ω*, es llamado espacio medible si satisface los requisitos*

1. $\emptyset \in \mathcal{A}$*,*

[1] Este tipo de diagramas fue popularizado por el inglés John Venn en 1880, pero en su trabajo (Venn, 1880) da la paternidad al matemático suizo Leonhard Euler, uno de los primeros en usar tal representación en el siglo XVIII en sus famosas "Cartas a una princesa alemana, acerca de diversas cuestiones de física y filosofía" (ver (Euler, 1768, L 102-105, pp. 95-126)), o antes a Christian Weise y Johan Christian Langius (Langius, 1712); apareció aún en trabajos de Gottfried Wilhelm Leibniz en el siglo anterior.

2. *si* $A \in \mathcal{A}$*, entonces* $\bar{A} \in \mathcal{A}$*,*

3. *la unión numerable de conjuntos de* $\mathcal{A}$ *queda en* $\mathcal{A}$ *(*$\mathcal{A}$ *es cerrado por la unión numerable).*

Con estas propiedades, $\mathcal{A}$ *es llamada una* σ*-álgebra. Los elementos de* $\mathcal{A}$ *son dichos medibles.*

Es sencillo mostrar que Ω también está en $\mathcal{A}$, y que $\mathcal{A}$ es cerrado por la intersección numerable. Un ejemplo de σ-álgebra sobre $\Omega = \{1\,;\,2\,;\,3\,;\,4\,;\,5\,;\,6\}$ puede ser $\mathcal{A} = \big\{\emptyset\,;\,\Omega\,;\,\{1\,;\,2\,;\,3\}\,;\,\{4\,;\,5\,;\,6\}\big\}$.

A partir de $(\Omega, \mathcal{A})$, se asocia una noción de probabilidad P a un dado evento. Esta queda determinada por los siguientes requisitos llamados *Axiomas de Kolmogorov* (ver por ejemplo (Spiegel, 1976; Kolmogorov, 1956; Shafer & Vovk, 2006; von Plato, 2005)):

1. $P(A) \geq 0 \;\; \forall\, A \in \mathcal{A}$.
2. Si $A_1, \ldots, A_i, \ldots$ son eventos mutuamente excluyentes de $\mathcal{A}$, entonces $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$.
3. $P(\Omega) = 1$.

Más formalmente, se define un *espacio de probabilidad* o *espacio probabilístico* de la manera siguiente (Halmos, 1950; Feller, 1968, 1971; Brémaud, 1988; Ibarrola et al., 1997; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Jacob & Protters, 2003; Cohn, 2013):

Definición 1-2 (Espacio de medida y espacio probabilístico)**.** *Sea* $(\Omega, \mathcal{A})$ *un espacio medible. Una función* $\mu : \mathcal{A} \mapsto \mathbb{R}_+$ *tal que*

1. $\mu(\emptyset) = 0$*, y*
2. *para cualquier conjunto numerable* $\{A_i\}_{i \in I}$ *(*I *numerable) de elementos mutuamente excluyentes de* $\mathcal{A}$ *se tiene* $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$*,*

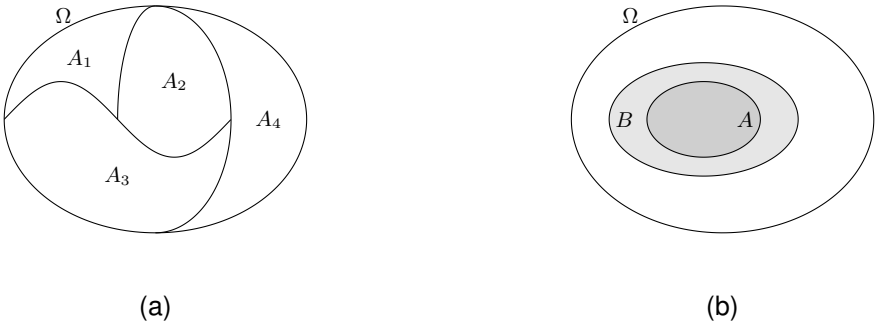
es llamada función medida o medida σ*-aditiva, y el espacio* $(\Omega, \mathcal{A}, \mu)$ *es llamado espacio de medida.*

- *Cuando* μ *es tal que existe un conjunto numerable* $\{A_i\}_{i \in I}$ *(*I *numerable) de elementos de* $\mathcal{A}$ *tal que* $\Omega = \cup_{i \in I} A_i$ *y* $\forall\, i \in I, \quad \mu(A_i) < +\infty$ *finito, la medida se dice* σ*-finita y el espacio de medida se dice* σ*-finito.*
- *Cuando* μ *está acotada por arriba,* $\mu(\Omega) < +\infty$*, la medida se dice finita y el espacio de medida también se dice finito.*
- *Además, si* $\mu(\Omega) = 1$*, la medida es dicha medida de probabilidad. En general, se la denota* P*. En este caso, el espacio* $(\Omega, \mathcal{A}, P)$ *es llamado espacio probabilístico.*

(ver también (Kolmogorov & Fomin, 1961, Cap. 5 & 6)). Es importante notar que una combinación lineal positiva de medidas es una medida, pero el producto de dos medidas no es una medida más.

A partir de los axiomas de Kolmogorov se pueden probar varios corolarios y propiedades:

- la probabilidad de un evento seguro o cierto es 1;
- la probabilidad de un evento que no puede ocurrir es 0: por ejemplo, $P(\emptyset) = 0$;
- el rango de las probabilidades está acotado: $0 \leq P(A) \leq 1 \ \forall \ A \in \mathcal{A}$;
- condición de normalización: si $\Omega = \bigcup_{i=1}^{n} A_i$, con A_i mutuamente excluyentes, entonces $\sum_{i=1}^{n} P(A_i) = 1$; el conjunto $\{A_i\}_{i=1}^{n}$ se dice *conjunto completo de eventos posibles excluyentes entre sí* y es ilustrado en la Figura 1-2;
- si A es subconjunto de B, lo que escribiremos $A \subset B$, es decir que si B se realiza, A se realiza también (pero no necesariamente al revés), entonces $P(A) \leq P(B)$; es ilustrado en la Figura 1-2.



(a) (b)

Figura 1-2: Ilustración de: (a) conjunto completo de eventos posibles excluyentes entre sí; (b) inclusión entre eventos, donde A está en gris oscuro mientras que B está en gris (claro y oscuro).

La probabilidad $P(A \cap B)$ del evento $A \cap B$ se llama también *probabilidad conjunta* de A y B. Se demuestra que

- $P(A \cap B)$ está acotada: $0 \leq P(A \cap B) \leq \min\{P(A), P(B)\}$ (viene de $A \cap B \subset A$ y $A \cap B \subset B$);
- si A y B son mutuamente excluyentes, entonces $P(A \cap B) = 0$ (viene de $A \cap B = \emptyset$);
- si $\{B_j\}_{j=1}^{m}$ es un conjunto completo de eventos posibles excluyentes entre sí, entonces $\sum_{j=1}^{m} P(A \cap B_j) = P(A)$ (viene de $\{A \cap B_j\}_j$ mutuamente excluyentes y $\bigcup_j (A \cap B_j) = A \cap \left(\bigcup_j B_j\right) = A \cap \Omega = A$).

En el caso de eventos no necesariamente mutuamente excluyentes, se prueba que la *ley de composición* o *fórmula de inclusión-exclusión* es

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

y que para n eventos resulta

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

La igualdad vale en el caso especial de eventos mutuamente excluyentes (recuperando el segundo axioma de Kolmogorov).

Se prueba también que si $\{A_i\}_{i=1}^{+\infty}$ es una secuencia *creciente* de eventos, *i. e.,* $\forall\, i \geq 1, \quad A_i \subset A_{i+1}$, entonces

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

Por otro lado, si $\{A_i\}_{i=1}^{+\infty}$ es una secuencia *decreciente* de eventos, *i. e.,* $\forall\, i \geq 1, \quad A_{i+1} \subset A_i$, entonces

$$P\left(\bigcap_{i=1}^{+\infty} A_i\right) = \lim_{i \to +\infty} P(A_i).$$

Podemos preguntarnos cuál es la probabilidad de un evento A, si sabemos que se da cierto evento B. Por ejemplo, para un dado de 6 caras equilibrado, cuál es la probabilidad de tener un número par sabiendo que tenemos un número menor o igual a 3. La respuesta se encuentra en la noción de *probabilidad condicional* (Hausdorff, 1901; Jeffrey, 1948, 1973; Brémaud, 1988; Mandel & Wolf, 1995; Ash & Doléans-Dade, 1999; Jacob & Protters, 2003; Shafer & Vovk, 2006):

Definición 1-3 (Probabilidad condicional)**.** *La probabilidad condicional de A dado B, denotado $P(A|B)$, se define como la razón entre la probabilidad del evento conjunto y la probabilidad de que se dé B (cuando éste es un evento de probabilidad no nula):*

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

En el ejemplo precedente, la probabilidad condicional va a ser $P(A|B) = \frac{P(A \cup B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

Claramente del hecho de que P es una medida de probabilidad se tiene

$$P(A|B) \geq 0.$$

Luego, de $A \cap B \subseteq B$ resulta $P(A \cap B) \leq P(B)$; es decir,

$$P(A|B) \leq 1.$$

Además, $P(\Omega \cap B) = P(B)$ dando

$$P(\Omega|B) = 1.$$

Para cualquier conjunto $\{A_i\}$ de eventos mutuamente excluyentes, los $(A_i \cap B)$ son también mutuamente excluyentes, así que $P\left(\left(\bigcup_i A_i\right) \bigcap B\right) = P\left(\bigcup_i \left(A_i \bigcap B\right)\right) = \sum_i P\left(A_i \bigcap B\right)$ dando

$$P\left(\bigcup_i A_i \middle| B\right) = \sum_i P\left(A_i | B\right).$$

Dicho de otra manera, $P(A|B)$ es una medida de probabilidad [2]. Diversas situaciones de probabilidades condicionales son ilustradas en la Fig. 1-3.

---

[2]Se puede definir un espacio de probabilidad $(\Omega_B, \mathcal{A}_B, P_B)$ donde $P_B(A) \equiv P(A|B)$. La notación $P_B(A)$ suele utilizarse en la literatura pero no la usaremos en esta obra para no confundirla con la medida de probabilidad de una variable aleatoria que definiremos en la sección siguiente.

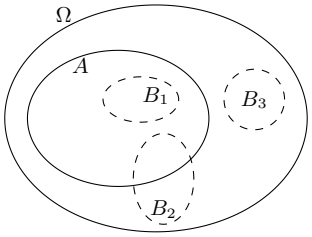


Figura 1-3: Ilustración de la probabilidad condicional con A interior de la curva en línea llena y unos B_i interiores de las curvas en líneas discontinuas. Se tiene $\omega \in B_1 \Rightarrow \omega \in A$ así que $P(A|B_1) = 1$; por otro lado, $\omega \in B_3 \Rightarrow \omega \notin A$ así que $P(A|B_3) = 0$. Entre estas situaciones extremas, si $P(\bar{A} \cap B_2) \neq 0$ y $P(A \cap B_2) \neq 0$ tenemos $0 < P(A|B_2) < 1$ (se puede tomar el ejemplo de probabilidad de un evento iguale a su superficia sobre la de Ω para ver estas propiedades en este caso particular).

Algunas propiedades interesantes son las siguientes:

- $P(A \cap B|B) = P(A|B)$ $\big($viene de $P(A \cap B \cap B) = P(A \cap B)\big)$;
- si A y B son mutuamente excluyentes, obviamente $P(A|B) = 0$;
- si $B \subseteq C$, entonces $P(A|B \cap C) = P(A|B)$ $\big($viene de $P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A \cap B)}{P(B)}$, pues $B \cap C = B\big)$;
- condición de normalización: si $\{A_i\}_{i=1}^n$ es un conjunto completo de resultados posibles mutuamente excluyentes, entonces $\sum_{i=1}^n P(A_i|B) = 1$;
- relación entre probabilidades condicionales inversas: $P(B|A) = \dfrac{P(B)}{P(A)} P(A|B)$, de donde $P(A|B)$ y $P(B|A)$ coinciden sólo cuando A y B tienen la misma probabilidad;

Las dos propiedades importantes son la fórmula de probabilidad total y la fórmula de Bayes (ver (Brémaud, 1988; Jacob & Protters, 2003; Bayes, 1763; Barnard, 1958) [3]). Les vamos a ver en forma de lemma:

Teorema 1-1 (Fórmula de probabilidad total)**.** *Sea $J \subseteq \mathbb{N}$ y $\{B_j\}_{j \in J}$ un conjunto completo de eventos no nulos mutuamente excluyentes, i. e., tal que $B_i \cap B_j = \emptyset$ si $i \neq j$ y $\bigcup\limits_{j \in J} B_j = \Omega$. Entonces*

$$P(A) = \sum_{j \in J} P(A|B_j) P(B_j)$$

Demostración. Escribimos $A = A \cap \Big(\bigcup_j B_j\Big) = \bigcup_j (A \cap B_j)$. Los $A \cap B_j$ son mutuamente excluyentes, y $P(A \cap B_j) = P(A|B_j)P(B_j)$ lo que cierra la prueba. □

[3] La obra del matemático y religioso inglés Thomas Bayes fue de hecho recopilada y publicada después de su muerte por Richard Price.

Lema 1-1 (Fórmula de Bayes). *Sea $J \subseteq \mathbb{N}$ y $\{B_j\}_{j\in J}$ un conjunto completo de eventos no nulos mutuamente excluyentes. Entonces*

$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j\in J} P(A|B_j)P(B_j)};$$

Demostración. Esta fórmula resuelta de la definición de la probabilidad condicional y de la fórmula de probabilidad total. □

Veamos ahora la noción de independencia entre dos eventos. Por ejemplo, si se tiran dos dados sobre sendas mesas, no hay ninguna razón para que la muestra de uno "influya" la del otro. Dicho de otra manera, dos eventos son independientes si el conocimiento de uno no lleva ninguna "información" sobre el otro (Brémaud, 1988; Mandel & Wolf, 1995; Ash & Doléans-Dade, 1999; Hausdorff, 1901; Jacob & Protters, 2003; Borel, 1909):

Definición 1-4 (Independencia estadística). *Dos eventos A y B se dicen estadísticamente independientes si la probabilidad condicional de A dado B es igual a la probabilidad incondicional de A:*

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

Es equivalente al hecho de que la probabilidad conjunta se factoriza:

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

Por inducción, la condición necesaria y suficiente para que n eventos $A_1, \ldots, A_n$ sean mutuamente estadísticamente independientes es que la probabilidad conjunta se factorice como

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Se deduce que los eventos mutuamente excluyentes no son estadísticamente independientes.

Es importante notar que la independencia mutua no es equivalente a la independencia por pares de eventos, como lo ilustra el ejemplo siguiente.

Ejemplo 1-1 (Independencia mutua vs por pares). *Tiramos 2 dados independientemente y consideramos los eventos: $A_i, i = 1, 2$ "el dado i es par" y A_3 "la suma de ambos dados es impar". Es claro que A_1 y A_2 son independientes y además para $i = 1$ o 2, $P(A_i \cap A_3) = \frac{1}{4} = P(A_i)P(A_3)$, mientras que $P(A_1 \cap A_2 \cap A_3) = 0 \neq \frac{1}{8}$: los eventos son independientes por pares, pero no son mutuamente independientes (Hogg et al., 2013).*

Definición 1-5 (Independencia condicional). *Dos eventos A y B se dicen estadísticamente independientes condicionalmente a un tercer evento C, si la probabilidad conjunta de A y B condicionalmente a C es igual al producto de la probabilidad de A condicionalmente a C por la de B condicionalmente a C:*

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

Si $P(B|C) \neq 0$, es equivalente a $P(A|B \cap C) = P(A|C)$.

Es importante notar que dos eventos pueden ser independientes, pero no serlo condicionalmente a un tercero, como lo ilustra el ejemplo siguiente.

Ejemplo 1-2 (Independencia incondicional pero no condicional)**.** *Teniendo dos monedas bien equilibradas y tirándolas de manera independiente, consideramos los eventos* A *"la primera faz es una cruz",* B *"la segunda faz es una cara",* C *"las faces son idénticas". Claramente* $P(A \cap B) = \frac{1}{4} = P(A)P(B)$*, mientras que* $P(A \cap B|C) = 0 \neq P(A|C)P(B|C) = \frac{1}{16}$*.*

Al revés, dos eventos pueden ser condicionalmente independientes a un tercero, pero ser dependientes.

Ejemplo 1-3 (Independencia condicional pero no incondicional)**.** *Sea Alice tirando una moneda bien equilibrada y denotamos* A *el evento "era una cruz". Claramente* $P(A) = \frac{1}{2}$*. Suponemos que Alice transmite el resultado a Bob a través de un intermediario Charlie con una probabilidad* ε *de mentir a Charlie, y llamamos* C *el evento "Alice dijo a Charlie que era una cruz". Tenemos que* $P(C) = P(C|A)P(A) + P(C|\overline{A})P(\overline{A}) = (1-\varepsilon)\frac{1}{2} + \varepsilon\frac{1}{2} = \frac{1}{2}$*. Suponemos ahora que Charlie transmite a Bob lo que le dijo Alice, con una probabilidad* ϑ *de mentir (independientemente de Alice) y llamamos* B *el evento "Charlie dijo a Bob que era una cruz". Es de nuevo sencillo ver que* $P(B) = \frac{1}{2}$*. Ahora,* $P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = 2\,P(A \cap B \cap C)$*. El evento* $A \cap B \cap C$ *es era una cruz y Alice no mintó y Charlie tampoco, es decir, por la independencia:* $P(A \cap B|C) = (1-\varepsilon)(1-\vartheta)$*. Inmediatamente* $P(B|C) = 1-\vartheta$ *y* $P(A|C) = 2\,P(A \cap C)$ *siendo* $A \cap C$ *el evento "era una cruz y Alice no mintió", i.e.* $P(A|C) = 1-\varepsilon$*. En conclusión,* $P(A \cap B|C) = P(A|C)P(B|C)$*:* A *y* B *son independientes condicionalmente a* C*. Ahora,* $P(A \cap B) = P(A \cap B \cap C) + P(A \cap B \cap \overline{C}) = \frac{1}{2}(1-\varepsilon)(1-\vartheta) + \frac{1}{2}\varepsilon\vartheta \neq \frac{1}{4} = P(A)P(B)$ *en general:* A *y* B *no resultan independientes. Este ejemplo es una instancia de lo que se llama un proceso de Markov, que vamos a ver un poco más en el cápitulo* **??***.*

1.3 Variables aleatorias y distribuciones de probabilidad

En un experimento o un dado proceso, los posibles resultados son típicamente números reales, siendo cada número un evento. Luego los resultados son mutuamente excluyentes. Se considera a esos números como valores de una *variable aleatoria* X a valores reales, que puede ser discreta, continua o mixta.

Formalmente, la noción de variable aleatoria se apoya sobre la noción de función medible. Por esta formalización, vamos a necesitar definir la integración de manera general, más allá del enfoque de Riemann ("a la Lebesgue"), así como la noción de derivada de una medida con respecto a otra para definir densidades de probabilidad, en analogía a la densidad de masa en mecánica por ejemplo (Lebesgue, 1904, 1918; Kolmogorov & Fomin, 1961; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013).

1.3.1 Consideraciones preliminaries: Teorías de la medida y de la integración.

La primera noción que subyace a la definición formal de variable aleatoria es la de función medible:

Definición 1-6 (Función medible)**.** *Sean* $(\Omega, \mathcal{A})$ *y* $(\Upsilon, \mathcal{B})$ *dos espacios medibles. Una función* $f : \Omega \mapsto \Upsilon$ *se dice* $(\mathcal{A}, \mathcal{B})$*-medible si*

$$\forall\, B \in \mathcal{B}, \quad A \equiv f^{-1}(B) = \left\{\omega \in A \;\middle|\; f(\omega) \in B\right\} \in \mathcal{A}.$$

Dicho de otra manera, la pre-imágen de un elemento dado de $\mathcal{B}$ *(elemento medible) pertenece a* $\mathcal{A}$ *(elemento medible). Por abuso de escritura, se dice más simplemente que* $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ *es medible.*

Además, a partir de un espacio de medida y una función f medible, se puede definir una medida imágen sobre el espacio de llegada (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

Teorema 1-2 (Teorema de la medida imágen)**.** *Sean* $(\Omega, \mathcal{A}, \mu)$ *un espacio de medida,* $(\Upsilon, \mathcal{B})$ *un espacio medible y* $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ *una función medible. Sea* μ_f *tal que*

$$\forall\, B \in \mathcal{B}, \quad \mu_f(B) = \mu\left(f^{-1}(B)\right).$$

Entonces, μ_f *es una medida sobre el espacio medible* $(\Upsilon, \mathcal{B})$*, i. e.,* $(\Upsilon, \mathcal{B}, \mu_f)$ *define un espacio de medida. Además,* $\mu(\Omega) = \mu_f(\Upsilon)$ *(posiblemente infinitas). Se dice que* μ_f *es la medida imágen de* μ *por* f*.*

Demostración. Por definición, claramente $\mu_f \geq 0$. Además, obviosamente $f^{-1}(\emptyset) = \emptyset$ dando $\mu_f(\emptyset) = \mu(\emptyset) = 0$. Luego, si para un conjunto numerable $\{B_j\}$ de elementos de $\mathcal{B}$ disjuntos entre sí, las pre-imágenes de los B_j también son disjuntos entre sí (para $k \neq j$ no se puede tener $\omega \in f^{-1}(B_j) \cap f^{-1}(B_k)$ sino ω tendría dos imágenes distintas por f). Entonces $f^{-1}\left(\bigcup_j B_j\right) = \bigcup_j f^{-1}(B_j)$. Esto implica que $\mu_f\left(\bigcup_j B_j\right) = \mu\left(f^{-1}\left(\bigcup_j B_j\right)\right) = \mu\left(\bigcup_j f^{-1}(B_j)\right) = \sum_j \mu\left(f^{-1}(B_j)\right) = \sum_j \mu_f(B_j)$. Finalmente, necesariamente $f^{-1}(\Upsilon) = \Omega$ (obviamente $f(\Omega) \subseteq \Upsilon$) lo que cierra la prueba [4] □

A continuación, necesitaremos tratar de funciones medibles teniendo una propiedad (P) salvo sobre un conjunto de medida μ igual a cero. Más generalmente viene acá la noción de propiedad *casi siempre*:

---

[4] De hecho, se puede probar sencillamente que la pre-imágen de una unión numerable (disjuntos o no) es la unión de las pre-imágenes; lo mismo occure para la intersección y además la pre-imágen del complemento es el complemento de la pre-imágen. Esto se conoce como *leyes de de Morgan* (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013) (ver también (Kolmogorov & Fomin, 1957, Cap. 1) y (Kolmogorov & Fomin, 1961, Caps. 5 & 6)).

Definición 1-7 (Propiedad (e igualdad) μ-casi siempre)**.** *Una funcion medible f se dice tener una propiedad (P) dada μ-casi siempre, si y solamente si la tiene excepto sobre un conjunto de medida nula,*

$$\mu\left(\{\omega \mid f(\omega) \text{ no satisface (P) }\}\right) = 0.$$

Por ejemplo, dos funciones medibles f_1 y f_2 $(\Omega, \mathcal{A}, \mu) \to (\Upsilon, \mathcal{B})$ son iguales μ-casi siempre,

$$f_1 = f_2 \quad (\mu\text{-c.s.})$$

si y solamente si son iguales excepto sobre un conjunto de medida nula,

$$\mu\left(\{\omega \mid f_1(\omega) \neq f_2(\omega)\}\right) = 0.$$

Un espacio que juega un rol particular es $\mathbb{R}^d$, al cual se puede asociar una σ-álgebra particular conocida como *σ-álgebra de Borel* (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a, 2007b; Cohn, 2013):

Definición 1-8 ($\mathbb{R}^d$ y Borelianos)**.** *Para cualquier $d \geq 1$ entero, llamamos Borelianos $\mathcal{B}(\mathbb{R}^d)$ de $\mathbb{R}^d$ a la σ-álgebra más pequeña generada por los productos cartesianos $\bigtimes_{i=1}^{d}(-\infty\,;\, b_i]$ (similarmente, por los abiertos de $\mathbb{R}^d$, o también para los productos cartesianos de intervalos $\bigtimes_{i=1}^{d}(a_i\,;\, b_i]$), i. e., uniones numerables, intersecciones numerables, complementos de estos intervalos. $\mathcal{B}(\mathbb{R}^d)$ es también llamado σ-álgebra de Borel de $\mathbb{R}^d$.*

Se necesita ahora definir la noción de integración de una función medible con respecto a una medida:

Definición 1-9 (Medida e integración)**.** *Para una medida cualquiera, sobre un espacio de medida $(\Omega, \mathcal{A}, \mu)$, se define la integración a partir de*

$$\forall\, A \in \mathcal{A}, \quad \int_A d\mu(\omega) = \int_\Omega \mathbb{1}_A(\omega)\, d\mu(\omega) = \mu(A),$$

donde $\mathbb{1}_A$ es la función indicadora (ver notaciones). es la función indicadora del conjunto A. $d\mu(\omega)$ se escribe a veces también $\mu(d\omega)$, medida de un "infinitésimo". Claramente, por propiedades de una medida, para A_i, A_j disjuntos $\mathbb{1}_{A_i} + \mathbb{1}_{A_j} = \mathbb{1}_{A_i \cup A_j}$, dando $\int_\Omega \left(\mathbb{1}_{A_i} + \mathbb{1}_{A_j}\right) d\mu(\omega) = \mu(A_i \cup A_j) = \mu(A_i) + \mu(A_j) = \int_\Omega \mathbb{1}_{A_i} d\mu(\omega) + \int_\Omega \mathbb{1}_{A_j} d\mu(\omega)$ y entonces, sin perdida de generalidad para un conjunto $\{A_j\}$ numerable y $\{a_j\}$ reales no negativos, la integral de la función escalonada $\sum_j a_j \mathbb{1}_{A_j}$ es dada por

$$\int_\Omega \left(\sum_j a_j \mathbb{1}_{A_j}(\omega)\right) d\mu(\omega) = \sum_j a_j \int_\Omega \mathbb{1}_{A_j}(\omega)\, d\mu(\omega).$$

Para los A_i disjuntos es la consecuencia directa de la propiedad precedente, y si A_i, A_j no son disjuntos. De hecho, sufice considerar $A_i \setminus A_j,\ A_j \setminus A_i, A_i \cap A_j$ con $A \setminus B = \{\omega \mid \omega \in A \ \wedge \ \omega \notin B\}$ y respectivamente los coefficientes $a_i,\ a_j,\ a_i + a_j$ para volver al caso de conjuntos disjuntos.

Antes de definir la integración de una función real, medible, cualquiera, el último paso que falta es el siguiente:

Teorema 1-3 (Función medible como límite)**.** *Sea $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, no negativa y medible. Existe una sucesión $\{g_n\}_{n\in\mathbb{N}}$ creciente de funciones escalonadas que converge simplemente (punto a punto) hacia g.*

Demostración. La sucesión $g_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\, \mathbb{1}_{g^{-1}\left(\left[\frac{k}{2^n}\,;\,\frac{k+1}{2^n}\right)\right)} + n\, \mathbb{1}_{g^{-1}([n\,;+\infty))}$ es escalonada, creciente y converge hacia g (notar que esta sucesión comparte la idea que subyace a la integración de Riemann). □

De este resultado, se puede generalizar la noción de integración de una función real:

Definición 1-10 (Integración de una función real)**.** *Sea $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, no negativa y medible, y $\{g_n\}_{n\in\mathbb{N}}$ una sucesión creciente de funciones escalonadas que converge simplemente hacia g. Por definición,*

$$\int_\Omega g(\omega)\, d\mu(\omega) = \lim_{n\to\infty} \int_\Omega g_n(\omega)\, d\mu(\omega).$$

Notar de que el límite puede ser infinita.

Sea ahora $g : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ medible cualquiera. Se verifica sencillamente que también $|g|$ (valor absoluto) es medible y, por definición, g se dice μ-integrable si la integral de $|g|$ es finita,

$$g \text{ es } \mu\text{-integrable} \quad \Leftrightarrow \quad \int_\Omega |g(\omega)|\, d\mu(\omega) < +\infty.$$

Además, se escribe $g = g_+ + g_-$ con $g_+ = \max(g, 0)$ y $g_- = \min(g, 0)$. Es sencillo ver de que si g es medible, g_+ y g_- son medibles. Si g es μ-integrable, necesariamente g_+ y g_- son μ-integrables, y, por definición

$$\int_\Omega g(\omega)\, d\mu(\omega) = \int_\Omega g_+(\omega)\, d\mu(\omega) - \int_\Omega (-g_-(\omega))\ d\mu(\omega).$$

A continuación, damos unos teoremas que serán muy útiles más adelante, sin detallar las pruebas. Por esto, el lector se puede referir a (Ash & Doléans-Dade, 1999; Lieb & Loss, 2001; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013).

Teorema 1-4 (Teorema de convergencia monótona)**.** *Sea $\{f_n\}_{n\in\mathbb{N}}$ una sucesión creciente de funciones medibles sobre $(\Omega, \mathcal{A}, \mu)$, positivas, convergiendo simplemente hacia una función f medible. Entonces*

$$\lim_{n\to+\infty} \int_\Omega f_n(\omega) d\mu(\omega) = \int_\Omega f(\omega) d\mu(\omega).$$

De hecho se prueba este teorema a partir de la definición de integración. Este teorema da una condición simple permitiendo intercambiar integración y límite.

Corolario 1-1. *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$*, positivas, tal que la serie* $\sum_n f_n$ *converge simplemente hacia una función* f*,* μ*-integrable. Entonces*

$$\int_\Omega \sum_{n\in\mathbb{N}} f_n(\omega) d\mu(\omega) = \int_\Omega f(\omega) d\mu(\omega).$$

Es una consecuencia del teorema de convergencia monótona, considerando la sucesión creciente $\{\sum_{k=0}^n f_k\}_{n\in\mathbb{N}}$.

Teorema 1-5 (Teorema de convergencia dominada)**.** *Sea* $\{f_n\}_{n\in\mathbb{N}}$ *una sucesión creciente de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$ *convergiendo simplemente hacia una función* f*, medible. Suponemos que existe una función* μ*-integrable* g *que domina la sucesión, i. e.,* $\forall\, \omega \in \Omega, \quad |f_n(\omega)| \leq g(\omega)$*. Entonces*

$$\lim_{n\to+\infty} \int_\Omega f_n(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\mu(\omega) \leq \int_\Omega g(\omega)\, d\mu(\omega).$$

Este teorema da una condición suficiente muy útil y muy usada para asegurarse de que se puede intercambiar límite e integración.

El último teorema que vamos a necesitar permite intercambiar dos integraciones. Antes, necesitamos definir la noción de espacio medible producto y medida producto.

Definición 1-11 (Espacio medible producto, medida producto)**.** *Sean dos espacios de medida* $(\Omega_1, \mathcal{A}_1, \mu_1)$ *y* $(\Omega_2, \mathcal{A}_2, \mu_2)$*. Llamamos espacio medible producto* $(\Omega, \mathcal{A})$ *al espacio del producto cartesiano* $\Omega = \Omega_1 \times \Omega_2$ *con la* σ*-álgebra* $\mathcal{A}$ *generada por los productos cartesianos* $A_1 \times A_2$ *donde* $A_i \in \mathcal{A}_i, i = 1, 2$*. Además, llamamos medida producto* μ *definida sobre* $\mathcal{A}$ *a la medida tal que* $\forall\, (A_1, A_2) \in \mathcal{A}_1 \times \mathcal{A}_2, \quad \mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$*.*

Teorema 1-6 (Teorema de Fubini)**.** *Sea* $(\Omega, \mathcal{A}, \mu)$ *espacio de medida producto de* $(\Omega_1, \mathcal{A}_1, \mu_1)$ *y* $(\Omega_2, \mathcal{A}_2, \mu_2)$ *donde* μ *es la medida producto. Sea* f *una función integrable sobre* $(\Omega, \mathcal{A}, \mu)$ *entonces*

- $\omega_1 \mapsto f(\omega_1, \omega_2)$ *es* μ_1*-integrable (*μ_2*-c.s.) y* $\omega_2 \mapsto f(\omega_1, \omega_2)$ *es* μ_2*-integrable (*μ_1*-c.s.),*
- $\omega_1 \mapsto \displaystyle\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$ *es* μ_1*-integrable y* $\omega_2 \mapsto \displaystyle\int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1(\omega_1)$ *es* μ_2*-integrable.*

Además,

$$\int_{\Omega_1\times\Omega_2} f(\omega)\, d\mu(\omega) = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega)\, d\mu_2(\omega_2) \right) d\mu_1(\omega_1) = \int_{\Omega_2} \left(\int_{\Omega_1} f(\omega)\, d\mu_1(\omega_1) \right) d\mu_2(\omega_2)$$

Teorema 1-7 (Integral a parámetro: continuidad y diferenciabilidad)**.** *Sea* I *un compacto de* $\mathbb{R}^d$ *y* $\{f(\cdot, t)\}_{t\in I}$ *una familia de funciones medibles sobre* $(\Omega, \mathcal{A}, \mu)$*, tal que* $t \mapsto f(\omega, t)$ *sea continua sobre* I *(*μ*-c.s.). Si existe una función* μ*-integrable* g *tal que*

$$\forall\, t \in I, \quad \forall\, \omega \in \Omega, \quad |f(\omega, t)| \leq g(\omega),$$

entonces $\omega \to f(\omega, t)$ *es* μ*-integrable y la función* $t \mapsto \displaystyle\int_\Omega f(\omega, t)\, d\mu(\omega)$ *es continua sobre* I*. Además, si* $f(\omega, \cdot)$ *es diferenciable sobre* I *y si existe una función* μ*-integrable* h *tal que*

$$\forall\, t \in I, \quad \forall\, \omega \in \Omega, \quad \|\nabla_t f(\omega, t)\| \leq h(\omega),$$

donde ∇_t *indica el gradiente, i. e., el vector de componentes* $\frac{\partial}{\partial t_1}, \ldots, \frac{\partial}{\partial t_d}$*, entonces la función* $t \mapsto \int_\Omega f(\omega, t)\, d\mu(\omega)$ *es diferenciable sobre* I*, y*

$$\nabla_t \int_\Omega f(\omega, t)\, d\mu(\omega) = \int_\Omega \nabla_t f(\omega, t)\, d\mu(\omega).$$

Básicamente, este teorema es consecuencia del teorema de convergencia dominada.

Seguimos esta sección con la noción de derivada de una medida con respecto a otra, dando una definición muy general de densidad:

Definición 1-12 (Densidad de una medida)**.** *Sean* μ *y* ν *dos medidas cualesquiera sobre un espacio medible* $(\Omega, \mathcal{A})$*. Si existe una función real no negativa* $p : \Omega \mapsto \mathbb{R}_+$ *medible tal que*

$$\forall\, A \in \mathcal{A}, \quad \nu(A) = \int_A p(\omega) d\mu(\omega),$$

p *es llamada densidad de* ν *con respecto a* μ*, denotada*

$$p = \frac{d\nu}{d\mu},$$

también llamada derivada de Radon-Nikodým.

Notar que dos funciones pueden cumplir esta definición, por ejemplo si son iguales μ-casi siempre. De hecho, si dos funciones $p_1 = p_2$ (μ-c.s.), y C es el conjunto donde no son iguales, siendo de medida nula, de $\int_A p_1(\omega) d\mu(\omega) = \int_{A\setminus C} p_1(\omega) d\mu(\omega) = \int_{A\setminus C} p_2(\omega) d\mu(\omega) = \int_A p_2(\omega) d\mu(\omega)$ se ve que dos funciones iguales casi siempre pueden ser densidad de una medida con respecto a una otra.

Es sencillo ver que si $\mu(A) = 0$, necesariamente $\nu(A) = 0$. De eso viene la noción de absoluta continuidad:

Definición 1-13 (Absoluta continuidad)**.** *Sean* μ *y* ν *dos medidas sobre un espacio medible* $(\Omega, \mathcal{A})$*. Se dice que* ν *es absolutamente continua con respecto a* μ*, denotado*

$$\nu \ll \mu,$$

si $\forall\, A \in \mathcal{A}, \quad \mu(A) = 0 \Rightarrow \nu(A) = 0.$

De hecho, se muestra la recíproca de la definición Def. 1-12 a través de lo que se conoce como teorema de Radon-Nikodým (Nikodym, 1930; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

Teorema 1-8 (Radon-Nikodým)**.** *Sean dos medidas* μ *y* ν*, entonces*

$$\nu \ll \mu \quad \Longleftrightarrow \quad \nu \text{ admite una densidad con respecto a } \mu.$$

Además, esta densidad $\frac{d\nu}{d\mu}$ *es única en el sentido de que si dos funciones cumplen la definición, son iguales* μ*-casi siempre.*

En todo lo que sigue, hablaremos de “la” densidad de una medida, salvo si se necesita explícitamente tener en cuenta esta sutileza.

A continuación, dos lemas van a ser muy útiles especialmente en el Capítulo **??**, tratando con dos (o más) medidas y densidades.

Lema 1-2. *Sean ν y μ dos medidas sobre $(\Omega, \mathcal{A})$ tales que $\nu \ll \mu$. Entonces, para cualquier función medible f,*

$$\int_\Omega f(\omega)\, \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\, d\nu(\omega)$$

Demostración. Tomando $f = \mathbb{1}_A$, de la definición Def. 1-12 se obtiene

$$\int_\Omega \mathbb{1}_A(\omega)\, \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega) = \nu(A) = \int_A d\nu(\omega)$$

Se cierra la prueba usande el teorema 1-3 y la definición 1-10, tratando f con su parte positiva y negativa separadamente. □

Lema 1-3. *Sean ν, μ y λ tres medidas sobre $(\Omega, \mathcal{A})$ y suponemos $\nu \ll \lambda$ y $\lambda \ll \mu$. Entonces*

- $\nu \ll \mu$;
- *equivalentemente, el soporte (ensemble de puntos que no anula la función) de $\frac{d\nu}{d\mu}$ está incluído (μ-casi siempre) en el soporte de $\frac{d\lambda}{d\mu}$;*
- $\frac{d\nu}{d\lambda}\frac{d\lambda}{d\mu} = \frac{d\nu}{d\mu}$ *(μ-c.s.).*

Demostración. El primer resultado viene de la definición de la absoluta continuidad $\mu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0 \quad \Rightarrow \quad \nu(A) = 0$. El segunndo resultado se obtiene escribiendo la medida en su forma integral. Adémas, por definición de la densidad, $\forall A \in \mathcal{A}, \quad \nu(A) = \int_A \frac{d\nu}{d\mu}(\omega)\, d\mu(\omega)$. Luego, aplicando el lema anterior a $f = \mathbb{1}_A \frac{d\nu}{d\mu}$ se obtiene que, también, $\nu(A) = \int_A \frac{d\nu}{d\lambda}(\omega)\, d\lambda(\omega) = \int_A \frac{d\nu}{d\lambda}(\omega)\, \frac{d\lambda}{d\mu}(\omega)\, d\mu(\omega)$, lo que cierra la prueba. □

Unas medidas que juegan un rol particular son las medidas de Lebesgue o medidas discretas.

Definición 1-14 (Medida de Lebesgue)**.** *La medida de Lebesgue μ_L sobre $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ se define tal que para cualquier producto cartesiano de intervalos,*

$$\mu_L\left(\bigtimes_{i=1}^d (a_i\,;\, b_i)\right) = \prod_{i=1}^d (b_i - a_i).$$

Acá, notamos dos hechos interesantes:

- μ_L es σ-finita. Viene de $\mathbb{R}^d = \cup_{i=1}^d \cup_{j_i \in \mathbb{Z}} \bigtimes_{i=1}^d (j_i\,;\, j_i + 1]$ conjuntamente a $\mu_L(\bigtimes_{i=1}^d (j_i\,;\, j_i + 1]) = 1 < +\infty$.

- $\forall\, A \in \mathcal{A}, \quad \mu_L(A) = |A|$ volumen de A.
- Para una función g suficientemente "suave", la integración con respecto a la medida de Lebesgue coincide naturalmente con la integración de Riemann.

La medida de Lebesgue es así natural para la integración. Luego, en lo que sigue, al mencionar igualdad μ_L-casi siempre, diremos simplemente "casi siempre" $(c.s.)$, entendiendo que es con respecto a la medida de Lebesgue. De la misma manera, hablando de densidad, sin precisiones, se entenderá que s con respecto a μ_L.

Al "contrario" de la medida de Lebesgue, medidas discretas son también particulares. La más "elemental" es conocida como *medida de Dirac*, dando lugar a medidas discretas:

Definición 1-15 (Medida de Dirac y medida discreta)**.** *La medida de Dirac al punto x_0, denotada δ_{x_0}, es tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad \delta_{x_0}(B) = \mathbb{1}_B(x_0)$$

Dado un conjunto $\mathcal{X} = \{x_i\}_i$ discreto (finito o infinito numerable), llamaremos medida discreta a la medida definida por

$$\mu_{\mathcal{X}} = \sum_i \delta_{x_i}$$

(en general, son definidas como combinaciones lineales positivas, siendo éste un caso particular).

Notar que,

- $\mu_{\mathcal{X}}$ es σ-finita (se muestra con el mismo enfoque que para la medida de Lebesgue).
- $\forall\, A \in \mathcal{A}, \quad \mu_{\mathcal{X}}(A) = |\mathcal{X} \cap A|$ cardenal de $\mathcal{X} \cap A$.
- Para una función g medible,

  $$\int_{\mathbb{R}^d} g(x)\, d\delta_{x_k}(x) = g(x_k) \qquad \text{y} \qquad \int_{\mathbb{R}^d} g(x)\, d\mu_{\mathcal{X}}(x) = \sum_{x \in \mathcal{X}} g(x),$$

  luego la integración se vuelve una suma. Se prueba saliendo de g de la forma $g = \mathbb{1}_C$ y del Teorema 1-3 conjuntamente con las definiciones Def. 1-10 y Def. 1-15.

Con esta serie de definiciones, tenemos todo lo necesario para introducir la definición de variables/vectores aleatorios reales y sus caracterizaciones.

1.3.2 Variables aleatorias y vectores aleatorios. Distribución de probabilidad.

Empezamos con la noción de variable aleatoria real, que corresponde como el resultado de un experimento o de un evento dado (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

Definición 1-16 (Variable aleatoria real)**.** *Una variable aleatoria real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$$

donde la medida P_X sobre $\mathcal{B}(\mathbb{R})$ es la medida imagen de P. P_X es frecuentemente llamada distribución de probabilidad o ley de la variable aleatoria X. En lo que sigue, escribiremos para cualquier $B \in \mathcal{B}(\mathbb{R})$ los eventos

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\},$$

así que, por definición,

$$P_X(B) = P(X \in B).$$

Para ilustrar esta definición, tomando el ejemplo de un dado, Ω es discreto y representa las caras, mientras que los números (se asocia a cada cara un número real) serán la imagen de Ω por X (ej. $X(\omega_j) = j, \quad j = 1, \ldots, 6$).

Notar que, por las propiedades de una medida sobre una σ-álgebra, para caracterizar completamente la distribución P_X es suficiente conocerla sobre los intervalos de la forma $(-\infty\,;\, b]$. Esto da lugar a la definición de función de repartición (a veces llamada función distribución por abuso de denominación) (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

Definición 1-17 (Función de repartición)**.** *La función de repartición F_X de una variable aleatoria X se define como*

$$F_X(x) = P_X((-\infty\,;\, x]) = P(X \leq x).$$

A veces, por abuso de lenguaje, se denomina a F_X ley de la variable aleatoria. Se encuentra también en la literatura la terminología función densidad cumulativa (cdf, por "cumulative density function" en inglés).

Naturalmente, de las propiedades de una medida de probabilidad se tiene:

- $0 \leq F_X(x) \leq 1$;
- $\lim_{x \to -\infty} F_X(x) = 0$ y $\lim_{x \to +\infty} F_X(x) = 1$ (viene de $P_X(\emptyset) = 0$ y $P_X(\mathbb{R}) = 1$);
- F_X es creciente $\big($viene de que $x_1 \leq x_2 \Leftrightarrow (-\infty\,;\, x_1] \subseteq (-\infty\,;\, x_2]\big)$;
- F_X no es necesariamente continua (lo vamos a ver más adelante); pero en cada punto x, F_X es continua por la derecha (ver inciso anterior).

Cuando se trabaja con $d \geq 2$ variables aleatorias es conveniente definir un *vector aleatorio* de dimensión d, y apelar para su estudio a nociones del álgebra lineal y a notación matricial. Se tiene el vector aleatorio d-dimensional $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$, caracterizado por d-uplas de variables aleatorias reales. Como en el caso univariado, se define este vector de la siguiente manera (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

Definición 1-18 (Vector aleatorio real)**.** *Un vector aleatorio real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X).$$

donde $\mathcal{B}(\mathbb{R}^d)$ *son los borelianos de* $\mathbb{R}^d$*,* σ*-álgebra generada por los productos cartesianos* $(-\infty\,;\, b_1] \times \cdots \times (-\infty\,;\, b_d]$ *y donde la medida* P_X *sobre* $\mathcal{B}(\mathbb{R}^d)$ *es la medida imagen de* P *llamada distribución de probabilidad de la variable aleatoria (o vector aleatorio)* X*. Como en el caso escalar, para cualquier* $B \in \mathcal{B}(\mathbb{R}^d)$

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \qquad \text{y} \qquad P_X(B) = P(X \in B).$$

Nota: a veces tenemos que considerar el caso de matrices aleatorias, o funciones medibles a valores matriciales. Dado que se puede poner en biyección una matriz con un vector (por ejemplo poniendo cada columna “debajo” de su columna antecedente), no desarrollaremos más este caso, a pesar de que a veces sea más conveniente trabajar con matrices en lugar de su forma en vector.

De las propiedades de una medida sobre una σ-álgebra, para caracterizar completamente la distribución P_X de nuevo es suficiente conocerla sobre los elementos de la forma $\bigtimes_{i=1}^{d}(-\infty\,;\, b_i]$, *i. e.,* la función de repartición multivariada (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

Definición 1-19 (Función de repartición multivariada)**.** *La función de repartición* F_X *de un vector aleatorio* X *es definida en* $x = (x_1, \ldots, x_d)$ *por*

$$F_X(x) = P_X\left(\bigtimes_{i=1}^{d}(-\infty\,;\, x_i]\right) = P\left(\bigcap_{i=1}^{d}(X_i \leq x_i)\right).$$

Por abuso de notación, escribiremos en lo que sigue

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

dando por entendido que $(X \leq x)$ *es el evento* $\bigcap_{i=1}^{d}(X_i \leq x_i)$*.*

De nuevo, de las propiedades de una medida de probabilidad surge:

- $0 \leq F_X(x) \leq 1$;
- $\lim_{\forall i, x_i \to -\infty} F_X(x) = 0$ y $\lim_{\forall i, x_i \to +\infty} F_X(x) = 1$;
- F_X es creciente con respecto a cada variable x_i.

Para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1\,;\ldots;\, d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ es obviamente un vector aleatorio k-dimensional. Es entonces sencillo ver que

$$F_{X_{I_k}}(x_{I_k}) = \lim_{\forall i \notin I_k, x_i \to +\infty} F_X(x)$$

$\Big($viene de que $\bigcap_{j=1}^{k}(X_{i_j} \leq x_{i_j}) = \Big(\bigcap_{j=1}^{k}(X_{i_j} \leq x_{i_j})\Big) \bigcap \Big(\bigcap_{i \notin I_k}(X_i \in \mathbb{R})\Big)\Big)$. Esta función se llama *función de repartición marginal* de F_X.

Cerramos estas generalidades con el caso de variables independientes:

Definición 1-20 (Independencia)**.** *Sean d variables aleatorias X_i y $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$. Las X_i son mutuamente independientes si y solamente si, para cualquier ensemble de conjuntos $B_i \in \mathcal{B}(\mathbb{R})$, los eventos $(X_i \in \mathcal{B}_i)$ son mutuamente independientes, i. e.,*

$$P_X\left(\bigtimes_{i=1}^{d} B_i\right) = \prod_{i=1}^{d} P_{X_i}(B_i).$$

Es equivalente a

$$F_X(x) = \prod_{i=1}^{d} F_{X_i}(x_i).$$

La ley del vector aleatorio se factoriza en este caso. Necesariamente, $\mathcal{X} = X(\Omega)$ es de la forma $\mathcal{X} = \bigtimes_{i=1}^{d} \mathcal{X}_i$ con $\mathcal{X}_i = X_i(\Omega)$, producto cartesiano.

Es importante notar que no es equivalente a tener la independencia por pares, como lo ilustramos al final de la sección precedente.

Más allá de este enfoque general, dos casos particulares de variables aleatorias son de interés: las variables discretas y las continuas. En el primer caso $X(\Omega)$ es discreto, finito o no. En las subsecciones siguientes estudiamos las particularidades de cada caso.

Para fijar notaciones, en todo lo que sigue escribiremos

$$\mathcal{X} = X(\Omega)$$

conjunto de llegada de X, o conjunto de valores que puede tomar la variable aleatoria. A veces, por simplicidad, se considera a $\mathcal{X}$ como el espacio muestral y se olvida que X sea una función medible entre espacios de probabilidades, *i. e.,* se trabaja en $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X)$ como en el espacio pre-imagen. Además, a veces, y por abuso, llamaremos frecuentemente a $\mathcal{X}$ *dominio de definición* de la medida de probabilidad P_X, siendo $P_X(\mathbb{R}^d \setminus \mathcal{X}) = 0$.

1.3.3 Variable aleatoria discreta

Definición 1-21 (Variable aleatoria discreta)**.** *Una variable aleatoria se dice discreta cuando $\mathcal{X} = X(\Omega)$ es discreto, finito o infinito numerable. aca saque la notacion del cardenal, porque ya esta en las notaciones En otras palabras, los posibles valores de una variable aleatoria discreta X consisten en un conjunto contable (finito o infinito numerable) de números reales, $\mathcal{X} = \{x_j\}$ y se puede escribir a X*

como una variable escalonada (ver ej. (Athreya & Lahiri, 2006; Hogg et al., 2013)),

$$X = \sum_j x_j \mathbb{1}_{A_j} \quad \textit{con} \quad A_j = X^{-1}(\{x_j\}).$$

Notar que Ω no es necesariamente discreto. Por ejemplo, si ω es la posición de un punto sobre una linéa, y se tiene $X(\omega) = 0$ si ω está a la izquierda de un umbral y $X(\omega) = 1$ si ω está a su derecha, luego $\mathcal{X} = \{0\,;\,1\}$ mientras que Ω no es discreto.

En el caso de una variable aleatoria discreta X, las probabilidades $P_X(\{x_j\}) = P(X = x_j),\ x_j \in \mathcal{X}$ caracterizan completamente esta variable aleatoria (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Hogg et al., 2013):

Definición 1-22 (Función de masa de probabilidad)**.** *Se define la función de masa de probabilidad de X, variable aleatoria discreta tomando sus valores sobre $\mathcal{X}$ por*

$$p_X(x) \equiv P(X = x) = P_X(\{x\}) \quad x \in \mathcal{X}.$$

Por abuso de denominación, llamaremos en este libro a p_X distribución de probabilidad. Además, usaremos también la notación

$$p_X = \begin{bmatrix} \cdots & p_X(x_i) & \cdots \end{bmatrix}^t$$

que llamaremos vector de probabilidad, de tamaño $|\mathcal{X}|$, posiblemente infinito.

En la Fig. 1-4-(a) se muestra una representación gráfica de una distribución de probabilidad discreta.

Notar que siendo P_X una medida de probabilidad, $p_X \geq 0$ y está obviamente normalizada en el sentido de que

$$\sum_{x \in \mathcal{X}} p_X(x) = 1.$$

Dicho de otra manera, en el caso finito $|\mathcal{X}| = \alpha < +\infty$, el vector de probabilidad p_X partenece al simplex estandar o simplex de probabilidad $p_X \in \Delta_{\alpha-1}$ (ver notaciones).

Volviendo a la medida de probabilidad P_X se nota que

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \sum_{x \in \mathcal{X} \cap B} p_X(x) = \int_B dP_X(x),$$

lo que da para la función de repartición

$$F_X(x) = \sum_{x_j \leq x} p_X(x_j).$$

De esta forma, se justifica la denominación *cumulativa* para F_X. También, se puede ver de inmediato que F_X es una función discontinua, con saltos finitos (en x_j, salto de altura $p_X(x_j)$). Esto es ilustrado figura 1-4-(b).

Un caso especial se tiene cuando un valor x_k es cierto o seguro, y no ocurre ninguno de los otros valores x_j $(j \neq k)$. La forma de la distribución es: $p_X(x) = 1$ si $x = x_k$ y 0 si no; el vector de probabilidad se escribirá $p_X = \mathbb{1}_k$ (ver notaciones; el vector posiblemente de dimensión infinita). See

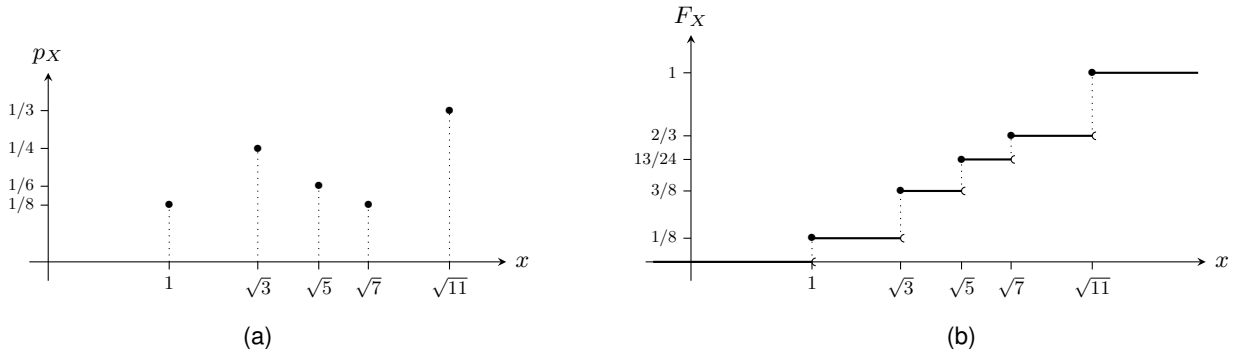


Figura 1-4: Ilustración de una distribución de probabilidad discreta (a), y la función de repartición asociada (b), con $\mathcal{X} = \{\, 1\,;\, \sqrt{3}\,;\, \sqrt{5}\,;\, \sqrt{7}\,;\, \sqrt{11}\,\}$ y $p_X = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{6} & \frac{1}{8} & \frac{1}{3} \end{bmatrix}^t$.

denota también con el *símbolo de Kronecker* $\delta_{jk} = 1$ si $j = k$ y 0 si no. En este libro, evitaremos usar este símbolo para no confundirlo con la medida de Dirac. Sin embargo, resuelta que P_X es precisamente la medida de Dirac en x_k (ver Def. 1-15).

Otra situación particular es la de *equiprobabilidad* o *distribución uniforme* cuando $|\mathcal{X}| = \alpha < +\infty$, $\alpha \in \mathbb{N}^*$. La forma de la distribución es: $p_X(x_j) = \frac{1}{\alpha} \quad \forall\, j = 1, \ldots, \alpha$, *i. e.,* $p_X = \begin{bmatrix} \frac{1}{\alpha} & \cdots & \frac{1}{\alpha} \end{bmatrix}^t$ o, en términos de medida, $P_X = \frac{1}{\alpha} \sum_{j=1}^{\alpha} \delta_{x_j}$. La función de repartición resulta una función escalonada, con saltos de altura $\frac{1}{\alpha}$ en cada $x_j, \quad j = 1\,,\, \ldots\,,\, \alpha$.

De manera general, la medida de probabilidad de una variable discreta se escribe como combinación convexa de medidas de Dirac,

$$P_X = \sum_j p_j \delta_{x_j}, \quad \text{con} \quad p_j = P(X = x_j) \geq 0, \quad \sum_j p_j = 1,$$

i. e., como una medida discreta.

Para comparar dos distribuciones discretas es útil reordenar cada vector de probabilidad permutando sus elementos hasta listarlos de forma decreciente. El vector reordenado a partir de p se anota $p^{\downarrow}$, de modo que $p_1^{\downarrow} \geq p_2^{\downarrow} \geq \ldots \geq p_\alpha^{\downarrow}$. En el ejemplo del caso con certeza se tiene $p^{\downarrow} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^t$, mientras que la distribución uniforme no varía. La comparación de dos vectores de probabilidad se puede apoyar sobre la noción de mayorización:

Definición 1-23 (Mayorización)**.** *Un vector de probabilidad (distribución) p mayorizado por un vector de probabilidad (distribución) q, denotado $p \prec q$, se define como:*

$$p \prec q \qquad \textit{sii} \qquad \sum_{i=1}^{k} p_i^{\downarrow} \leq \sum_{i=1}^{k} q_i^{\downarrow} \quad \forall\, k = 1\,,\, \ldots\, \alpha - 1, \qquad \textit{y} \qquad \sum_{i=1}^{\alpha} p_i^{\downarrow} = \sum_{i=1}^{\alpha} q_i^{\downarrow}$$

(sienco las últimas sumas iguales a 1). Si los alfabetos de definición de p y q son de tamaños diferentes, α es el tamaño más grande y la distribución sobre el alfabeto más corto es completada por estados de probabilidad 0 (sería equivalente a añadir estados ficticios de probabilidad nula).

Por ejemplo, $\begin{bmatrix}0{,}40 & 0{,}30 & 0{,}20 & 0{,}10\end{bmatrix}^t \prec \begin{bmatrix}0{,}50 & 0{,}30 & 0{,}15 & 0{,}05\end{bmatrix}^t$ (ver Fig. 1-5-(a)).

Es importante resaltar que la mayorización provee un *orden parcial* (no total) entre distribuciones, existiendo pares de distribuciones tales que ninguna mayoriza a la otra. Por ejemplo, $\begin{bmatrix}0{,}50 & 0{,}35 & 0{,}10 & 0{,}05\end{bmatrix}^t$ y $\begin{bmatrix}0{,}60 & 0{,}20 & 0{,}17 & ,03\end{bmatrix}^t$ no se comparan por mayorización (ver Fig. 1-5-(b)).

Es interesante notar que la siguiente propiedad es válida para toda distribución p de tamaño α (Marshall, Olkin & Arnold, 2011, p. 9, (6)-(8)):

$$\begin{bmatrix}\frac{1}{\alpha} & \frac{1}{\alpha} & \cdots & \frac{1}{\alpha}\end{bmatrix}^t \prec p \prec \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^t.$$

En este sentido, los casos particulares de equiprobabilidad y de certeza, se dice que son distribuciones extremas. Notamos que uno implica ignorancia máxima en el resultado de la variable mientras que el otro corresponde a conocimiento completo.

La relación de mayorización es ilustrada en la figura 1-5, donde se representan las sumas parciales en función de k, llamadas *curvas de Lorentz* [5] (Marshall et al., 2011; Lorenz, 1905). Gráficamente, $p \prec q$ es equivalente a tener la curva de Lorentz asociada a p por debajo de la asociada a q.

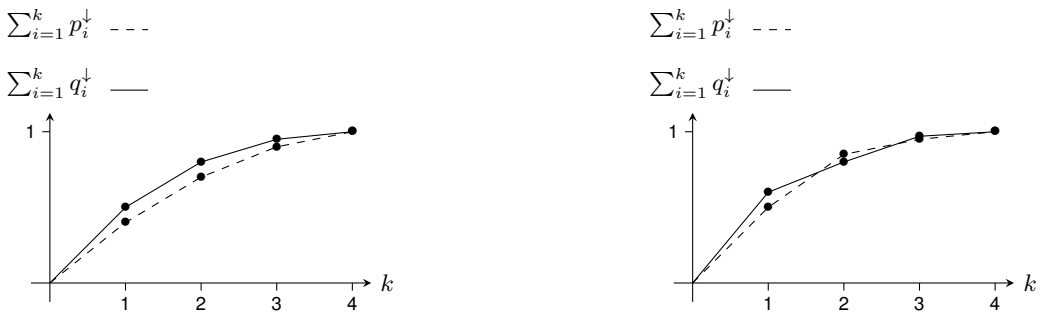


Figura 1-5: Orden parcial por mayorización: sumas parciales para $k = 1\,,\, 2\,,\, 3\,,\, 4$ (a) para los vectores de probabilidades $p = \begin{bmatrix}0{,}40 & 0{,}30 & 0{,}20 & 0{,}10\end{bmatrix}^t$ (línea punteada) y $q = \begin{bmatrix}0{,}50 & 0{,}30 & 0{,}15 & 0{,}05\end{bmatrix}^t$ (línea llena) y (b) para los vectores de probabilidad $p = \begin{bmatrix}0{,}50 & 0{,}35 & 0{,}10 & 0{,}05\end{bmatrix}^t$ (línea punteada) y $q = \begin{bmatrix}0{,}60 & 0{,}20 & 0{,}17 & 0{,}03\end{bmatrix}^t$ (línea llena). En el caso (a), $p \prec q$ mientras que en el caso (b), $p \not\prec q$ y $q \not\prec p$ (no están ordenadas por mayorización).

1.3.4 Variable aleatoria continua

[5] Se prueba sencillamente que estas curvas son crecientes y cóncavas.

En varios contextos, una variable aleatoria puede tomar valores en un conjunto no numerable, por ejemplo cualesquiera de los números en cierto intervalo de la recta real. Ya no es una variable discreta. En las variables que no son discretas, el caso particular de interés es el de variables continuas (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Hogg et al., 2013):

Definición 1-24 (Variable aleatoria continua)**.** *Una variable aleatoria X se dice continua si su función de repartición F_X es continua sobre $\mathbb{R}$.*

Cuando se puede, es conveniente asociar una *función densidad de probabilidad* (comúnmente anotada por su sigla en inglés: pdf, por *probability density function*). La definición de tal densidad se apoya en la definición 1-12 aplicada a la medida de probabilidad P_X:

Definición 1-25 (Variable aleatoria que admite una densidad de probabilidad)**.** *Sea X una variable aleatoria continua y P_X su medida de probabilidad. Por definición, se dice que X admite una densidad de probabilidad con respecto a una medida μ sobre $\mathbb{R}$ si $P_X \ll \mu$ (teorema de Radon-Nikodým 1-8).*
En general, nos enfocamos en la medida (llamada de referencia) $\mu = \mu_L$ de Lebesgue. Denotando $d\mu_L(x) \equiv dx$, la definición se reduce a: Si existe una función no negativa p_X medible sobre $\mathbb{R}$ tal que

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \int_B p_X(x)\, dx,$$

entonces se dice que X admite una densidad y p_X es llamada densidad de probabilidad de X (dando por entendido "con respecto a la medida de Lebesgue"). Notando que $P_X(B) = P_X(B \cap \mathcal{X})$, el soporte de p_X es necesariamente $\mathcal{X} = X(\Omega)$ (i. e., $p_X(\overline{\mathcal{X}}) = 0$ y $p_X(\mathcal{X}) \neq 0$), y

$$\forall\, B \in \mathcal{B}(\mathbb{R}), \quad P_X(B) = \int_{B \cap \mathcal{X}} p_X(x)\, dx.$$

Para la función de repartición F_X tenemos entonces

$$F_X(x) = \int_{-\infty}^{x} p_X(u)\, du$$

(esta expersión es valids por cualquier medida μ, densidad con respecto a esta medida de referencia, e integración sobre $(-\infty\,;\, x]$ con el "diferencial" $d\mu(x)$). Dicho de otra manera, si F_X es (continua y) derivable sobre $\mathbb{R}$, al menos por partes, X admite una densidad de probabilidad (con respecto a la medida de Lebesgue) y [6]

$$p_X(x) = \frac{dF_X(x)}{dx}.$$

Por abuso de terminología, en lo que sigue llamaremos a p_X también distribución de probabilidad, a pesar de que no tiene el mismo sentido que la masa de probabilidad del caso discreto.

La escritura integral de F_X justifica de nuevo la denominación *cumulativa* para F_X. Además, se puede ver por ejemplo que en este caso $P(a < X \leq b) = \int_a^b p_X(x)\, dx = F_X(b) - F_X(a)$ y que

[6] Recordar que, rigurosamente, la igualdad debe ser entendida "casi siempre".

claramente

$$\forall x \in \mathbb{R}, \quad P_X(\{x\}) = P(X = x) = 0,$$

esto es, $\{x\}$ es de medida P_X nula. Similarmente, cualquier conjuntos numerable de $\mathbb{R}$ es de medida P_X nula.

Notar aue aun cuando $0 \leq F_X \leq 1$, p_X puede ser mayor que uno. Por ejemplo, para $F_X(x) = 2\,x\,\mathbb{1}_{[0\,;\,\frac{1}{2})}(x) + \mathbb{1}_{[\frac{1}{2}\,;\,+\infty)}(x)$, que define correctamente una función de repartición, $p_X(x) = 2\mathbb{1}_{[0\,;\,\frac{1}{2})}(x)$. No es contradictorio en el sentido de que p_X no es una probabilidad, sino que $p_X(x)\,dx$ puede ser visto como la probabilidad de hallar a la variable con valores en el "intervalo infinitesimal entre x y $x + dx$". Finalmente, la condición de normalización se escribe

$$\int_{\mathcal{X}} p_X(x)\,dx = \int_{\mathbb{R}} p_X(x)\,dx = 1.$$

En la figura 1-6-(a) se muestra una representación gráfica de una función densidad de probabilidad para una variable continua que admite una densidad, y en la figura 1-6-(b) la función de repartición correspondiente.

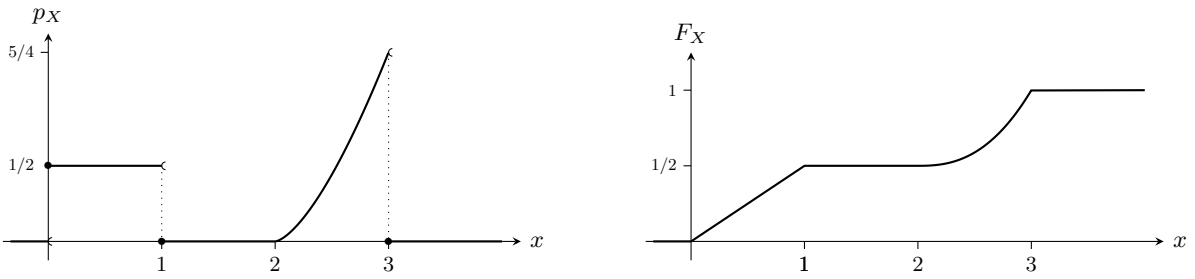


Figura 1-6: Ilustración de: (a) una distribución de probabilidad continua, y (b) la función de repartición asociada, con $\mathcal{X} = [0\,;\,1) \cup [2\,;\,3)$ y $p_X(x) = \frac{1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \frac{5(x-2)^{\frac{3}{2}}}{4}\mathbb{1}_{[2\,;\,3)}(x)$, *i. e.,* $F_X(x) = \frac{x}{2}\,\mathbb{1}_{[0\,;\,1)}(x) + \frac{1}{2}\mathbb{1}_{[1\,;\,2)}(x) + \frac{1+(x-2)^{\frac{5}{2}}}{2}\mathbb{1}_{[2\,;\,3)}(x) + \mathbb{1}_{[3\,;\,+\infty)}(x)$.

Notar que una variable aleatoria puede no ser ni continua, ni discreta, como se ilustra en el ejemplo siguiente:

Ejemplo 1-4 (Ejemplo de variable mixta)**.** *Sean U y V variables continuas, independientes, de densidad de probabilidad $p_U = p_V = \mathbb{1}_{[0\,;\,1)}$ (U y V son uniformes sobre $[0\,;\,1)$) y sea $X = V\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$, es decir $X(\omega) = V(\omega)$ si $U(\omega) < \frac{1}{2}$ y 1 si no. Entonces de la fórmula de probabilidades totales, $F_X(x) = P(X \leq x) = P\left((X \leq x)\,\middle|\,(U < \frac{1}{2})\right) P\left(U < \frac{1}{2}\right) + P\left((X \leq x)\,\middle|\,(U \geq \frac{1}{2})\right) P\left(U \geq \frac{1}{2}\right)$ i. e., $F_X(x) = \frac{1}{2}P\left((V \leq x)\,\middle|\,(U < \frac{1}{2})\right) + \frac{1}{2}P\left((1 \leq x)\,\middle|\,(U \geq \frac{1}{2})\right)$. Ahora, de la independencia de U y V, tenemos $F_X(x) = \frac{1}{2}F_V(x) + \frac{1}{2}\mathbb{1}_{[1\,;\,+\infty)}(x)$ es decir*

$$F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x).$$

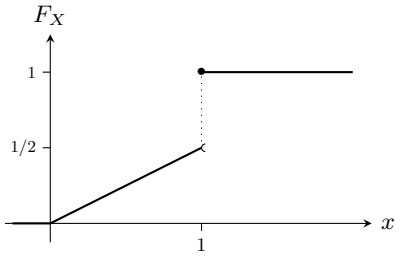


Figura 1-7: Función de repartición $F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x)$ asociada a $X = V\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con U y V variables independientes, continuas, uniformes sobre $\mathcal{X} = [0\,;\,1)$. F_X no es tipo escalón, así que X no es discreta. A pesar de que $\mathcal{X} = [0\,;\,1]$ es un intervalo, de la presencia del salto en $x = 1$ tampoco X es continua.

Esta función de repartición es representada en la figura 1-7: no es discreta ni continua. Entonces, a pesar de que $\mathcal{X} = [0\,;\,1]$ sea un intervalo real, X no es continua (y tampoco puede ser discreta).

Volvemos a las variables discretas X sobre $\mathcal{X} = \{x_j\}_j$, de medida de probabilidad de la forma $P_X = \sum_j p_j \delta_{x_j}$. Considerando la medida discreta $\mu_{\mathcal{X}} = \sum_j \delta_{x_j}$, es claro que $P_X \ll \mu_{\mathcal{X}}$. Entonces, formalmente, P_X admite una densidad con respecto a la medida discreta $\mu_{\mathcal{X}}$ y esta densidad, definida sobre $\mathcal{X}$ es $p_X(x) = P(X = x)$. A pesar de que sea una tautología, esto justifica que usemos la escritura p_X (minúscula) tanto en el caso discreto como en el caso continuo, y que hablemos (por abuso de terminología) de distribución de probabilidad en ambos casos.

Recordemos que cualquier medida de probabilidad (caso continuo o no) se escribe también con una integral $P_X(B) = \int_B dP_X(x)$ y que en el caso discreto cierto $X = x_k$, la medida de probabilidad P_X es la medida de Dirac. A veces, por abuso de escritura $dP_X(x)$ es denotado $\delta_{x_k}(x)\,dx$ o $\delta(x - x_k)\,dx$ donde ahora δ es llamada *distribución (delta) de Dirac*. Se puede ver este Dirac como una densidad de probabilidad $p_X(x)$ con respecto a la medida de Lebesgue pero no es una función "ordinaria" dado que P_X no es diferenciable con respecto a la medida de Lebesgue. Se la llama *función generalizada* o *distribución de Schwartz* [7]. En particular, $F_X(x) = \mathbb{1}_{\mathbb{R}_+}(x - x_k)$ ($\mathbb{1}_{\mathbb{R}_+}$ es conocido también como *función de Heaviside* [8]) y en el sentido de las distribuciones, $\frac{dF_X}{dx} = \delta_{x_k}$. Además, se usan en general las propiedades, para cualquier function f y real x_0,

$$f(x)\delta(x - x_0) = f(x_0)\delta(x - x_0) \qquad \text{y} \qquad \int_{\mathbb{R}} f(x)\delta(x - x_0)\,dx = f(x_0),$$

[7] La teoría de distribuciones valió a Laurent Schwarz la medalla Fields en 1950. Entre otros trabajos, se probó que el Dirac, visto como distribución de Schwartz o función generalizada, tiene una "representación integral" $\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\imath tx} dt$ o más rigurosamente transformada de Fourier de $x \mapsto 1$ en el sentido de las funciones generalizadas o distribuciones. Esto muestra claramente su carácter no ordinario (la integral siendo divergente en el sentido usual). Esto va más allá de la meta del capítulo y el lector se podrá referir a (Schwartz, 1966; Gel'fand & Shilov, 1964, 1968) por ejemplo.

[8] Viene del nombre del físico ingles Oliver Heaviside quien estudio las ecuaciones electromagneticas de Maxwell.

pero hay que entender la integración a través de la medida Dirac (insistimos en el hecho de que esta notación es un abuso de escritura, ej. (Gel'fand & Shilov, 1964)). Usando las distribuciones de Dirac, se puede unificar el tratamiento de las variables aleatorias discretas con las continuas en términos de densidad (con respecto a la medida de Lebesgue): si una variable aleatoria discreta toma los valores x_j con probabilidades $p_j = P(X = x_j)$ respectivamente, entonces formalmente se puede describir mediante una variable aleatoria continua X con "función densidad de probabilidad" $p_X(x) = \sum_j p_j\, \delta(x - x_j)$. Insistimos en el hecho de que rigurosamente debemos trabajar con medidas, como lo hemos formalizado al principio de este capítulo.

Terminamos mencionando el resultado siguiente, probado en (Athreya & Lahiri, 2006, Ec. (2.5) p. 47 & teorema 4.1.1) por ejemplo [9]:

Teorema 1-9 (Descomposición de una medida de probabilidad)**.** *Cualquier función de repartición $F_X(x)$ se descompone como combinación convexa de una función de repartición F_{d} discreta y una función de repartición F_{c} continua:*

$$\exists\, a \in [0\,;\,1] \quad \textit{tal que} \quad F_X(x) = aF_{\mathrm{d}}(x) + (1-a)F_{\mathrm{c}}(x)$$

En términos de medida, o como corolario de (Athreya & Lahiri, 2006, teorema 4.1.1), cualquier medida de probabilidad P_X se descompone como la combinación convexa de una medida discreta P_{d} y una continua P_{c},

$$\exists\, a \in [0\,;\,1],\ \widetilde{\mathcal{X}} \textit{ discreto} \quad \textit{tal que} \quad P_X = aP_{\mathrm{d}} + (1-a)P_{\mathrm{c}} \qquad \textit{con} \qquad P_{\mathrm{d}} \ll \mu_{\widetilde{\mathcal{X}}} \quad \textit{y} \quad P_{\mathrm{c}} \ll \mu_L$$

Entonces, $P_X \ll \mu_{\widetilde{\mathcal{X}}} + \mu_L$, i. e., admite una densidad con respecto a la medida σ-finita $\mu_{\widetilde{\mathcal{X}}} + \mu_L$.

Dicho de otra manera, cualquier variable aleatoria es mixta, como en el ejemplo 1-4. Del teorema, es discreta cuando $a = 1$, y continua cuando $a = 0$.

1.3.5 Vector aleatorio discreto

Un ejemplo de vector aleatorio discreto puede ser dado por un conjunto de dados (que podrían ser dependientes si están ligados por un hilo por ejemplo).

Definición 1-26 (Vector aleatorio discreto)**.** *Un vector aleatorio d-dimensional se escribe $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ y $\mathcal{X} = X(\Omega) \subset \bigtimes_{i=1}^{d} \mathcal{X}_i$ donde $\mathcal{X}_i = X_i(\Omega)$. Se dice que X es discreto cuando $\mathcal{X} \subseteq \mathbb{N}^d$, es discreto, finito o infinito numerable.*

---

[9] Básicamente, se muestra que $\widetilde{X} = \{x \in \mathcal{X} \mid \ p(x) = F_X(x) - \liminf_{u \to x} F(u) > 0\}$ es numerable. A continuación, $F(x) - \sum_{\widetilde{x} \in \widetilde{\mathcal{X}}} p(\widetilde{x}) \mathbb{1}_{(-\infty\,;\,\widetilde{x}]}(x)$ es continua, y se recupera la descomposición con $a = \sum_{\widetilde{x} \in \widetilde{\mathcal{X}}} p(\widetilde{x})$.

Obviamente, la medida de probabilidad en los $x = (x_1\,,\,\ldots\,,\,x_d) \in \bigtimes_{i=1}^{d} \mathcal{X}_i$ caracteriza completamente este vector aleatorio:

Definición 1-27 (Función de masa de probabilidad conjunta)**.** *Por definición, la función de masa de probabilidad de X, vector aleatorio discreto que toma sus valores sobre $\mathcal{X} \subset \bigtimes_i \mathcal{X}_i$, está dada por*

$$p_X(x) \equiv P(X = x) = P\left(\bigcap_{i=1}^{d} (X_i = x_i)\right) \quad \forall\, x_i \in \mathcal{X}_i,\ 1 \leq i \leq d.$$

Se la llama también función de masa de probabilidad conjunta de los X_i o, por abuso de denominación, llamaremos a p_X distribución de probabilidad (conjunta). Notar que $\mathcal{X}$ no es necesariamente igual al producto cartesiano de los $\mathcal{X}_i$.

En el caso multivariado, la notación vectorial es más delicada de usar: p_X sería un "tensor" (o tabla) d-dimensional (una matriz para $d = 2$, una "tabla" 3 dimensonal para $d = 3$,...). Pero es posible usar una notación vectorial, recordando que $\mathbb{N}^d$ puede ser puesto en biyección con $\mathbb{N}$, y dada una biyección usarla para etiquetar los componentes de p_X puestos en vector. En el caso finito $\mathcal{X}_i = \{x_{j_i}\}_{j=1}^{\alpha_i}$ con $\alpha_i = |\mathcal{X}_i| < +\infty$, se puede organizar los componentes tales que $p_X(x_{j_1}\,,\,\ldots\,,\,x_{j_d})$ sea la j-esima componente del vector p_X con $j = 1 + \sum_{i=1}^{d}(j_i - 1) \prod_{k=i+1}^{d} \alpha_k$.

De nuevo, se puede interpretar p_X como densidad con respecto a la medida discreta $\mu_{\mathcal{X}}$, Def. 1-15.

Similarmente al caso escalar $d = 1$, la función de repartición de un vector aleatorio discreto d-dimensional es tipo escalón d-dimensional, *i. e.,* compuesto de partes de hiperplanos d-dimensionales, F_X constante sobre $[x_{(j-1)_1}\,;\,x_{j_1}) \times \cdots \times [x_{(j-1)_d}\,;\,x_{j_d})$. Además, las componentes son mutuamente independientes si y solamente si la función de répartición se factoriza, o equivalentemente la función de masa se factoriza, *i. e.,*

$$X_i \text{ mutuamente independientes} \quad \Leftrightarrow \quad p_X = \prod_{i=1}^{d} p_{X_i}.$$

En notación tensorial, $p_X = p_{X_1} \otimes \cdots \otimes p_{X_d}$, producto de Kronecker (ver notaciones).

Al final, de la fórmula de calculo de función de repartición marginal vista en la sección 1.3.2, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1\,;\,\ldots\,;\,d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ tiene como probabilidad marginal o distribución marginal

$$p_{X_{I_k}}(x_{I_k}) = \sum_{\forall i \notin I_k, x_i \in \mathcal{X}_i} p_X(x).$$

1.3.6 Vector aleatorio continuo

Como para el caso de una variable, se puede considerar cualquier medida de referencia μ sobre $\mathbb{R}^d$ para definir una noción de densidad (d-variada), pero en general nos enfocamos en la medida de Lebesgue.

Definición 1-28 (Vector aleatorio continuo y densidad de probabilidad multivariada)**.** *Un vector aleatorio* $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ *se dice continuo si su función de repartición* F_X *es continuo sobre* $\mathbb{R}^d$*. Como en el caso escalar, por definición 1-12 (y la recíproca evocaca después de la definición), se dice que* X *admite una densidad de probabilidad* p_X *con respecto a una medida* μ *sobre* $\mathbb{R}^d$ *si* $P_X \ll \mu$*. De nuevo, nos enfocamos en la medida (llamada de referencia)* $\mu = \mu_L$ *de Lebesgue: si existe una función no negativa y medible* $p_X : \mathbb{R}^d \mapsto \mathbb{R}$ *tal que*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_X(B) = \int_B p_X(x)\, dx = \int_{B \cap \mathcal{X}} p_X(x)\, dx$$

con $\mathcal{X} = X(\Omega)$ *soporte de* p_X *y* $d\mu_L(x) \equiv dx = dx_1 \cdots dx_d$*, entonces se dice que* X *admite una densidad y* p_X *es llamada densidad de probabilidad de* X *(entendiendo "con respecto a la medida de Lebesgue"), o también densidad de probabilidad conjunta de los* X_i*. En particular,*

$$F_X(x) = \int_{\times_{i=1}^d (-\infty\,;\, x_i]} p_X(u)\, du$$

o, equivalentemente, para F_X *(continua y) derivable sobre* $\mathbb{R}^d$ *(con respecto a la medida de Lebesgue), por lo menos por partes,*

$$p_X(x) = \frac{\partial^d F_X(x)}{\partial x_1 \cdots \partial x_d}.$$

Usaremos la terminología (por abuso) de distribución de probabilidad.

Como en el caso escalar, $p_X \geq 0$ (no es necesario que sea menor que 1) y satisface la condición de normalización

$$\int_{\mathcal{X}} p_X(x)\, dx = \int_{\mathbb{R}^d} p_X(x)\; dx = 1.$$

El teorema 1-9 se cumple también en el caso d-dimensional: cualquier medida de probabilidad P_X (resp. función de repartición F_X) se descompone como combinación convexa de una medida de probabilidad (resp. función de repartición) discreta y una continua. En otros términos existe un $\widetilde{X}$ discreto tal que $P_X \ll \mu_{\widetilde{\mathcal{X}}} + \mu_L$ (σ-finita).

Mencionamos que las d variables aleatorias $X_1, \ldots, X_d$, componentes de un vector aleatorio X, son independientes si y solamente si la función de repartición se factoriza, lo que da, derivando esta,

$$X_i \text{ mutuamente independientes} \quad \Leftrightarrow \quad p_X(x) = \prod_{i=1}^{d} p_{X_i}(x_i).$$

Seguimos esta sección mencionando que, de la fórmula de cálculo de función de repartición marginal vista en la sección 1.3.2, para un subconjunto $I_k = (i_1\,, \ldots\,, i_k)$ de $1 \leq k \leq d$ elementos de $\{1\,; \ldots\,; d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ tiene la *densidad de probabilidad marginal*

$$p_{X_{I_k}}(x_{I_k}) = \int_{\times_{i \notin I_k} \mathcal{X}_i} p_X(x) \prod_{i \notin I_k} dx_i = \int_{\mathbb{R}^{d-k}} p_X(x) \prod_{i \notin I_k} dx_i.$$

En particular, la función densidad de probabilidad marginal que caracteriza a la variable aleatoria X_i es la ley que se obtiene integrando la densidad de probabilidad conjunta sobre todas las variables excepto la i-ésima.

Como en el caso discreto, se puede querer comparar dos distribuciones de probabilidad, y por eso reordenar o rearreglar una densidad de probabilidad. Como en el caso discreto, denotamos $p_X^{\downarrow}$ a la densidad con rearreglo simétrico. Primero, se necesita definir el rearreglo simétrico de un ensemble y luego de una densidad de probabilidad:

Definición 1-29 (Rearreglo simétrico de un conjunto)**.** *Sea $\mathcal{P} \subset \mathbb{R}^d$ abierto de volumen finito, $|\mathcal{P}| < +\infty$. El rearreglo simétrico $\mathcal{P}^{\downarrow}$ de $\mathcal{P}$ es la bola centrada en 0 de igual volumen que $\mathcal{P}$, i. e.,*

$$\mathcal{P}^{\downarrow} = \left\{ x \in \mathbb{R}^d \;\middle|\; \frac{2\pi^{\frac{d}{2}} \|x\|^d}{\Gamma\left(\frac{d}{2}\right)} < |\mathcal{P}| \right\} = \mathbb{B}_d \left(0\,,\, \frac{1}{\sqrt{\pi}} \left(\frac{|\mathcal{P}| \Gamma\left(\frac{d}{2}\right)}{2} \right)^{\frac{1}{d}} \right),$$

Esto se ilustra en la figura 1-8-a.

Definición 1-30 (Rearreglo simétrico de una densidad de probabilidad)**.** *Sea p_X una densidad de probabilidad y sean $\mathcal{P}_t = \{y \mid p_X(y) > t\} \subset \mathbb{R}^d$ para cualquier $t > 0$, sus conjuntos de niveles. La densidad de probabilidad con rearreglo simétrico $p_X^{\downarrow}$ de p_X se define como* [10]

$$p_X^{\downarrow}(x) = \int_0^{+\infty} \mathbb{1}_{\mathcal{P}_u^{\downarrow}}(x)\, du.$$

De $\forall\, t < \tau \Leftrightarrow \mathcal{P}_\tau \subseteq \mathcal{P}_t \Leftrightarrow \mathcal{P}_\tau^{\downarrow} \subseteq \mathcal{P}_t^{\downarrow}$, es sencillo ver que si $x \in \mathcal{P}_\tau^{\downarrow}$, entonces $x \in \mathcal{P}_t^{\downarrow}$, lo que conduce a $p_X^{\downarrow}(x) > \tau$, y vice-versa. Más allá, sobre $\mathcal{P}_{\tau+d\tau} \setminus \mathcal{P}_\tau$ la función p_X "vale" τ y sobre $\mathcal{P}_{\tau+d\tau}^{\downarrow} \setminus \mathcal{P}_\tau^{\downarrow}$ la función $p_X^{\downarrow}$ "vale" también τ, lo que da $\int_{\mathcal{P}_\tau^{\downarrow}} p_X^{\downarrow}(x)\, dx = \int_{\mathcal{P}_\tau} p_X(x)\, dx$ (ver (Lieb & Loss, 2001; Wang & Madiman, 2004) para une prueba más rigurosa). La representación de la definición es conocida como representación tarta en capas ("layer cake representation" en ingles). Esto es ilustrado en la figura 1-8-b para el caso escalar $d = 1$. Notar que, por construcción, $p_X^{\downarrow}$ cae el la familia de densidades esfericas (ver sección 1.10) (Lord, 1954; Fang, Kotz & Ng, 1990; Cambanis, Huang & Simons, 1981; Eaton, 1981).

A partir de esta definición del rearreglo, se puede ahora extender la noción de mayorización del caso discreto al caso continuo de la manera siguiente:

Definición 1-31 (Mayorización en el contexto continuo)**.** *Una densidad de probabilidad p se dice mayorizada por una distribución q sii:*

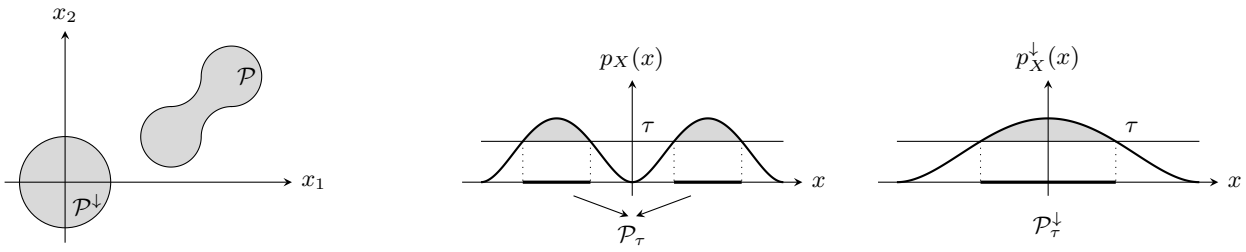
$$p \prec q \qquad \textit{sii} \qquad \int_{\mathbb{B}_d(0,r)} p^{\downarrow}(x)\, dx \leq \int_{\mathbb{B}_d(0,r)} q^{\downarrow}(x)\, dx \quad \forall\, r > 0, \qquad \textit{y} \qquad \int_{\mathbb{R}^d} p^{\downarrow}(x)\, dx = \int_{\mathbb{R}^d} q^{\downarrow}(x)\, dx,$$

(las últimas integrales son obviamente iguales a 1).

Nota: la función

$$\mathcal{L}_p(r) = \int_{\mathbb{B}_d(0,r)} p^{\downarrow}(x)\, dx$$

[10] Se prueba que esta función, positiva por definición, suma a 1. Además, por construcción, depende únicamente de $\|x\|$ y decrece con $\|x\|$.



(a) (b)

Figura 1-8: (a): Ilustración del rearreglo simétrico $\mathcal{P}^{\downarrow}$ de un conjunto $\mathcal{P}$, siendo $\mathcal{P}^{\downarrow}$ la bola centrada en 0 de mismo volumen que $\mathcal{P}$, en el caso bi-dimensional $d = 2$. (b) Construcción del rearreglo $p_X^{\downarrow}$ en un contexto escalar (ver ejemplo 1-5 para $d = 1,\ \nu = 5,\ m = 1$): dado un τ, se busca $\mathcal{P}_\tau$ (izquierda) y se deduce $\mathcal{P}_\tau^{\downarrow}$ (derecha); dado un x, se busca el mayor t tal que $x \in \mathcal{P}_t^{\downarrow}$, siendo entonces este t máximo igual a $p_X^{\downarrow}(x)$ (derecha); además, por construcción, las superficies en grise son iguales.

da el equivalente de la curba de Lorentz en el contexto continuo, así que la relación de mayorización se interpreta graficamente de la misma manera que en el caso discreto (excepto que, contrariamente al caso discreto, la curva no es necesariamente concava).

Ejemplo 1-5. *Consideramos la densidad de probabilidad d-dimensional mezcla* [11] **de Student-r (ver sección 1.10)**

$$p_X(x) = \alpha \left(\left(1 - \|x+m\|^2\right)^{\frac{\nu-d}{2}} \mathbb{1}_{\mathbb{B}_d(-m,1)}(x) + \left(1 - \|x-m\|^2\right)^{\frac{\nu-d}{2}} \mathbb{1}_{\mathbb{B}_d(m,1)}(x) \right)$$

con

$$\nu > d-2, \qquad m \in \mathbb{R}^d \setminus \mathbb{B}_d \qquad \textit{y} \qquad \alpha = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)} \quad \textit{coeficiente de normalización}$$

Esta densidad de probabilidad es dibujada figura 1-8-((b) izquierda) para $d = 1,\ \nu = 5, m = 1$ y figura 1-9-(a) para $d = 2,\ \nu = 6,\ m = \frac{1}{\sqrt{d}}\mathbb{1}$. El dominio de definición, el máximo, y la matriz de covarianza **(ver sección 1.10)** *son dados por*

$$\mathcal{X} = \mathbb{B}_d(-m,1) \cup \mathbb{B}_d(m,1), \qquad \max_{\mathcal{X}} p_X(x) = \alpha, \qquad \Sigma_X = \frac{1}{\nu+2} I + mm^t$$

Ahora, para $\tau > \alpha$, $\mathcal{P}_\tau = \emptyset$ y para cualquier $\tau \in [0\,;\,\alpha]$ buscando los x tal que $p_X(x) > \tau$ (notando que las bolas en $\mathcal{X}$ son disjuntas) conduce a

$$\mathcal{P}_\tau = \mathbb{B}_d(-m,\beta_\tau) \cup \mathbb{B}_d(m,\beta_\tau) \qquad \textit{con} \qquad \beta_\tau = \sqrt{1 - \left(\frac{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)\tau}{\Gamma\left(\frac{\nu}{2}+1\right)} \right)^{\frac{2}{\nu-d}}}$$

Notando que las esferas de $\mathcal{P}_\tau$ son disjuntas, queda claro que el volumen de $\mathcal{P}_\tau$ es dado por $|\mathcal{P}_\tau| = 2\,|\mathbb{B}_d(0,\beta_\tau)| = \left|\mathbb{B}_d\left(0, 2^{\frac{1}{d}}\beta_\tau\right)\right|$, que conduce al dominio rearreglado,

$$\mathcal{P}_\tau^{\downarrow} = \mathbb{B}_d\left(0, 2^{\frac{1}{d}}\beta_\tau\right)$$

[11] Una mezcla de ley es definida como combinación convexa $f = \sum_{i=1}^k \alpha_i f_i$ de leyes f_i con $\alpha = \begin{bmatrix}\alpha_1 & \cdots & \alpha_k\end{bmatrix}^t \in \Delta_{k-1}$.

Ahora, se muestra sencillamente que $x \in \mathcal{P}_u^{\downarrow}$ es equivalente a

$$u < \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)}\left(1-\frac{\|x\|^2}{4^{\frac{1}{d}}}\right)^{\frac{\nu-d}{2}}\mathbb{1}_{\mathbb{B}_d\left(0,2^{\frac{1}{d}}\right)}(x)$$

De la definición 1-30 obtenemos al final

$$p_X^{\downarrow}(x) = \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)}\left(1-\frac{\|x\|^2}{4^{\frac{1}{d}}}\right)^{\frac{\nu-d}{2}}\mathbb{1}_{\mathbb{B}_d\left(0,2^{\frac{1}{d}}\right)}(x)$$

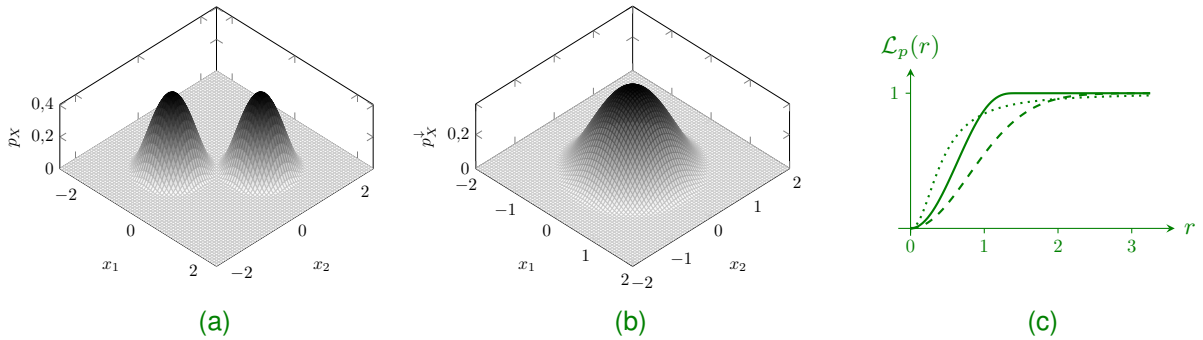
dibujada figura 1-8-((b) derecha) para $d=1,\ \nu=5$ y figura 1-9 para $d=2,\ \nu=6$. Finalmente, la curva de Lorentz es dada por

$$\begin{aligned}\mathcal{L}_{p_X}(r) &= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)}\int_{\mathbb{B}_d(0,r)}\left(1-\frac{\|x\|^2}{4^{\frac{1}{d}}}\right)^{\frac{\nu-d}{2}}\mathbb{1}_{\mathbb{B}_d\left(0,2^{\frac{1}{d}}\right)}(x)\,dx \\ &= \frac{\Gamma\left(\frac{\nu}{2}+1\right)}{2\,\pi^{\frac{d}{2}}\Gamma\left(\frac{\nu-d}{2}+1\right)}\frac{2\,\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}\int_0^r \rho^{d-1}\left(1-\frac{\rho^2}{4^{\frac{1}{d}}}\right)^{\frac{\nu-d}{2}}\mathbb{1}_{\left[0\,;\,2^{\frac{1}{d}}\right]}(\rho)\,d\rho\end{aligned}$$

pasando en coordenadas hiperesferica (ver (Lord, 1954; Fang et al., 1990; Cambanis et al., 1981) y sección 1.10). Por el cambio de variable $u=\frac{\rho^2}{4^{\frac{1}{d}}}$ obtenemos

$$\mathcal{L}_{p_X}(r) = \frac{\int_0^{\frac{r^2}{4^{\frac{1}{d}}}} u^{\frac{d}{2}-1}(1-u)^{\frac{\nu-d}{2}}\,du}{B\left(\frac{d}{2},\,\frac{\nu-d}{2}+1\right)}$$

conocida como función beta incompleta (Gradshteyn & Ryzhik, 2015, Ec. 8.392), tomada en $\frac{r^2}{4^{\frac{1}{d}}}$. La figura 1-9-(c) dibuja $\mathcal{L}_{p_X}(r)$ para $d=2,\ \nu=6$, la de la ley gausiana esferica g_X de covarianza $\frac{\mathrm{Tr}\,\Sigma_X}{d}I$ y la de la ley Student-t esferica s_X de covarianza $\frac{\mathrm{Tr}\,\Sigma_X}{d}I$ y con $\nu'=2{,}25$ grado de libertad (ver sección 1.10). Eso ilustra graficamente la relación de mayorización $g_X \prec p_X$ y que ambas $s_X \not\prec p_X$ y $p_X \not\prec s_X$.



(a) (b) (c)

Figura 1-9: (a) Densidad de probabilidad p_X del ejemplo 1-5 en el contexto bi-dimensional $d=2$ y para $\nu=6,\ m=\frac{1}{\sqrt{d}}\mathbb{1}$. (b) Densidad rearreglada $p_X^{\downarrow}$. (c) Equivalente continua de la curva de Lorentz para la densidad p_X (linea llena), la densidad gausiana esferica g_X con covarianza de misma traza que la de p_X (linea con guiones) y la densidad Student-t con covarianza de misma traza que la de p_X y $\nu'=2{,}25$ grado de libertad (linea punteada): $g_X \prec p_X$ pero $s_X \not\prec p_X$ y $p_X \not\prec s_X$.

1.4 Transformación de variables y vectores aleatorios

En esta sección nos interesamos en los efectos sobre una variable o un vector aleatorio. Por ejemplo, en un juego con dos dados, nos puede interesar la ley de la suma que daría el número de casilla que debemos adelantar en un juego de la oca.

Teorema 1-10 (Transformación medible de un vector aleatorio)**.** *Sea* $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *una variable aleatoria, y* $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$ *una función medible. Entonces,* $Y = g(X)$ *es una variable aleatoria* $(\Omega, \mathcal{A}) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$*. Además, la medida imagen* P_Y *está vinculada a* P_X *por*

$$\forall\, B \in \mathcal{B}(\mathbb{R}^{d'}), \quad P_Y(B) = P_X(g^{-1}(B)).$$

Demostración. Este resultado es obvio. Siendo g una función medible (recordar Def. 1-6), para todo $B \in \mathcal{B}(\mathbb{R}^{d'})$, por definición $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)$. Además, si P_X es la medida (de probabilidad) asociada al espacio de salida de g, el resultado es consecuencia del teorema de la medida imagen 1-2. □

(Ver ej. (Mukhopadhyay, 2000; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Bogachev, 2007b; Cohn, 2013)).

Es sencillo probar que cualquier combinación de funciones medibles queda medible, cualquier producto (adecuado) de funciones medibles queda medible, y que si $\{f_k\}_{k=1}^{d'}$ son $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}))$-medibles, entonces $f = (f_1, \ldots, f_{d'})$ es $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^{d'}))$-medible (Athreya & Lahiri, 2006).

Mencionamos que si $\mathcal{X} = X(\Omega)$ es discreto, entonces $\mathcal{Y} = g(\mathcal{X}) = Y(\Omega)$ será discreto también, y:

Teorema 1-11 (Función de masa por transformación medible)**.** *Sean* X*, vector aleatorio* d*-dimensional discreto,* $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'}))$ *una función medible, e* $Y = g(X)$ *necesariamente discreto* d'*-dimensional sobre* $\mathcal{Y} = g(\mathcal{X})$*. La distribución de* Y *está vinculada con la de* X *por la relación*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x).$$

Demostración. El resultado es inmediato. □

En particular, si g es inyectiva (necesariamente biyectiva de $\mathcal{X}$ en $\mathcal{Y}$), el vector de probabilidad queda invariante, $p_Y = p_X$; solamente cambian los estados.

Es importante mencionar que con $\mathcal{Y}$ discreto, $\mathcal{X}$ no es necesariamente discreto (Athreya & Lahiri, 2006). Por ejemplo, $Y = \mathbb{1}_{X>0}$ es tal que $\mathcal{Y} = \{0\,;\,1\}$ a pesar de que $\mathcal{X}$ puede no ser discreto.

Tratar con variables aleatorias continuas resulta más delicado. Vimos en el ejemplo precedente que el carácter continuo puede perderse por transformación. De la misma manera, en un ejemplo de la sección anterior, vimos que $Y = X_1 \mathbb{1}_{X_2>0}$ con X_i independientes uniformes no es continua ni discreta. En el enfoque de variables continuas, una clase importante de funciones en las cuales nos vamos a interesar son las funciones continuas (y diferenciables):

Lema 1-4 (Continuidad y carácter medible)**.** *Sea $g : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ continua. Entonces, g es $(\mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^{d'}))$-medible.*

Demostración. Por continuidad, la pre-imagen de un abierto de $\mathbb{R}^{d'}$ por g es un abierto de $\mathbb{R}^d$ y entonces es en $\mathcal{B}(\mathbb{R}^d)$. La prueba se cierra recordando la definición de $\mathcal{B}(\mathbb{R}^{d'})$, σ-álgebra generada por los abiertos de $\mathbb{R}^{d'}$. □

En lo que sigue, nos interesamos más especialmente en el caso de funciones $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. De hecho, si $d' < d$, es sencillo llegar al caso considerado añadiendo $d - d'$ transformaciones. Por ejemplo, con $d = 2$ si nos interesa $X_1 + X_2$, se puede considerar $\begin{bmatrix} X_1 + X_2 & X_2 - X_1 \end{bmatrix}^t$ y llegar a la variable de interés por cálculo de marginal. Si $d' > d$ la situación es más delicada, $g(Y)$ viviendo sobre una variedad d-dimensional de $\mathbb{R}^{d'}$.

En el caso de vectores aleatorios continuos X que admitend una densidad de probabilidad, una pregunta natural es entonces saber si se conserva la continuidad y la existencia de una densidad, así como su forma. La respuesta se da en el teorema siguiente (Brémaud, 1988; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013):

Teorema 1-12 (Densidad de probabilidad por transformación continua inyectiva diferenciable)**.** *Sean X, vector aleatorio d-dimensional continuo que admite una densidad de probabilidad p_X, y sea $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ una función continua inyectiva y diferenciable tal que $|\mathrm{J}_g| > 0$ (ver notaciones), Sea $Y = g(X)$. Entonces Y es continua y admite una densidad de probabilidad p_Y de soporte $\mathcal{Y} = g(\mathcal{X}) = Y(\Omega)$ tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = p_X(g^{-1}(y)) \left|\mathrm{J}_{g^{-1}}(y)\right|.$$

Demostración. Por definición, admitiendo X una densidad y siendo g medible,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_Y(B) = P_X(g^{-1}(B)) = \int_{g^{-1}(B)\cap\mathcal{X}} p_X(x)\, dx.$$

Por cambio de variables $x = g^{-1}(y)$ (siendo g inyectiva, el antecedente es único por definición) y notando que $g\left(g^{-1}(B) \cap \mathcal{X}\right) = B \cap \mathcal{Y}$,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^d), \quad P_Y(B) = \int_{B\cap\mathcal{Y}} p_X(g^{-1}(y)) \left|\mathrm{J}_{g^{-1}}(y)\right| \, dy$$

lo que cierra la prueba [12]. □

El caso escalar puede ser visto como caso particular, dando:

---

[12] La aparición del Jacobiano viene del mismo enfoque que el cambio de variables en la integración de Riemann. De hecho, como lo hemos visto, $\mu_L(B) = |B|$ es el volumen y de la definición misma del determinente, para cualquier matriz cuadrada el volumen se escribe $\mu_L(MB) = |MB| = |M||B| = |M|\mu_L(B)$ donde la misma escritura $|\cdot|$ representa el valor absoluto del determinente de una matriz. Esta notación se justifica precisamente por su significación de volumen, y el resultado es inmediato para $g(x) = Mx$. La forma para una transformación más general se obtiene a partir de un desarollo de Taylor al orden 1 de la transformación, haciendo aparecer el determinente del Jacobiano (Athreya & Lahiri, 2006; Cohn, 2013).

Corolario 1-2. *Sean X, variable aleatoria continua que admite una densidad de probabilidad p_X, $g : \mathbb{R} \mapsto \mathbb{R}$ una función continua, inyectiva y diferenciable, e $Y = g(X)$. Entonces Y es continua y admite una densidad de probabilidad p_Y tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = p_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|.$$

De hecho, se pueden ver estos resultados esquemáticamente como una "conservación" de probabilidad, $p_X(x)dx = p_Y(y)dy$, el volumen dy estando relacionado al dx a traves de lamatriz Jacobiana (ver nota de pie **??**).

Una forma alternativa de derivar este corolario consiste en salir de la función de repartición, notando que g es necesariamente monótona [13]: si $y \notin \mathcal{Y}$, necesariamente $p_Y = 0$ ($F_Y(y) = 1$ si $y > \sup \mathcal{Y}$ y $F_Y(y) = 0$ si $y < \text{ínf}\, \mathcal{Y}$) y para cualquier $y \in \mathcal{Y}$,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \begin{cases} P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) & \text{si} \quad g \quad \text{es creciente} \\ P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{si} \quad g \quad \text{es decreciente} \end{cases}.$$

El resultado se obtiene calculando las derivadas del primer y último términos respecto de la variable transformada y.

Si g no es inyectiva, g^{-1} es multivaluada o multiforme. En este caso, se puede todavía tratar el problema, particionando $\mathbb{R}^d$ en conjuntos donde g es inyectiva, dando

Teorema 1-13 (Densidad de probabilidad por transformación continua no inyectiva diferenciable)**.** *Sean X, vector aleatorio d-dimensional continuo que admite una densidad de probabilidad p_X, y sea $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ una función continua y diferenciable. Denotamos $\{\mathcal{X}_{[k]}\}_{k=0}^m$ la partición de $\mathcal{X}$ tal que $|\mathrm{J}_g(y)| = 0$ sobre $\mathcal{X}_{[0]}$, y para todo $k \geq 1$ se tiene $g : \mathcal{X}_{[k]} \mapsto \mathcal{Y}$ inyectiva y tal que $|\mathrm{J}_g(y)| > 0$. Suponemos que $\mathcal{X}_{[0]}$ es de medida de Lebesgue nula, notamos g_k^{-1} la función inversa de g sobre $g(\mathcal{X}_{[k]})$ (rama k-ésima de la función multivaluada g^{-1}), $\mathrm{J}_{g_k^{-1}}$ su matriz Jacobiana, e $I(y) = \{k \mid y \in g(\mathcal{X}_{[k]})\}$ los índices tales que y tiene un inverso por g_k. Esto es ilustrado en la figura 1-10 para $d = 1$. Entonces Y es continua y admite una densidad de probabilidad p_Y tal que*

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{k \in I(y)} p_X(g_k^{-1}(y)) \left| \mathrm{J}_{g_k^{-1}}(y) \right|.$$

En el caso escalar $d = 1$ esto se formula

$$\forall\, y \in \mathcal{Y}, \quad p_Y(y) = \sum_{k \in I(y)} p_X(g_k^{-1}(y)) \left| \frac{dg_k^{-1}(y)}{dy} \right|.$$

Esto se ilustra el la figura 1-10.

Demostración. Basta escribir $B = \bigcup_{k=0}^m \left(B \cap g(\mathcal{X}_{[k]})\right)$ unión de borelianos disjuntos, notar que por consecuencia $g^{-1}(B) = \bigcup_{k=0}^m g^{-1}\left(B \cap g(\mathcal{X}_{[k]})\right)$ unión de borelianos disjuntos y por linealidad escribir

[13] Notar que $P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x) + P(X = x)$, pero siendo X continua, $P(X = x) = 0$.

la integración sobre $g^{-1}(B)$ como la suma de integrales sobre $g^{-1}\left(B \cap g(\mathcal{X}_{[k]})\right)$. Se cierra la prueba notando que $g^{-1}\left(B \cap g(\mathcal{X}_{[0]})\right)$ es necesariamente de medida de Lebesgue nula, siendo la integral nula y que $g^{-1}\left(B \cap g(\mathcal{X}_{[k]})\right) = g_k^{-1}\left(B \cap g(\mathcal{X}_{[k]})\right)$. □

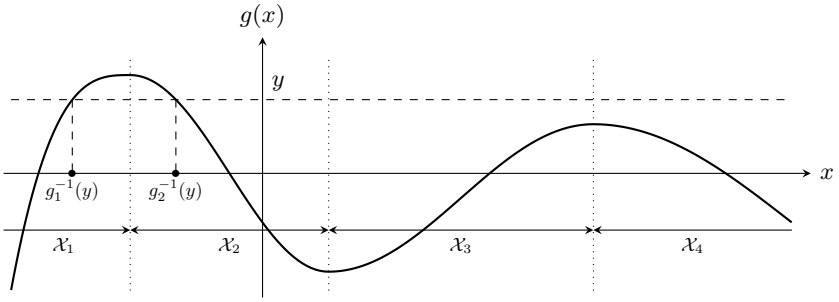


Figura 1-10: Ilustración de una transformación g no inyectiva, tal que $\mathcal{X}_{[0]} = \{x \big| g'(x) = 0\}$, representado por los valores de x en las lineas punteadas. Es de medida de Lebesgue nula. Se indican los dominios $\mathcal{X}_{[k]}$. La línea discontinua da un nivel y y los puntos en el eje x representan $g_k^{-1}(y),\ k \in I(y)$; en el ejemplo, $I(y) = \{1\,;\,2\}$ y, suponiendo que $\mathcal{X} = \mathbb{R},\ F_Y(y) = F_X(g_1^{-1}(y)) + 1 - F_X(g_2^{-1}(y))$.

Ejemplo 1-6 (Ejemplo de transformación no biyectiva)**.** *Sea X definida sobre $\mathcal{X} = \mathbb{R}$ y la transformación de variables $Y = X^2$. Se tiene $y = g(x) = x^2$, continua diferenciable de derivada nula sobre $\mathcal{X}_{[0]} = \{0\}$, de medida nula, cuyas inversas son $g_1^{-1}(y) = -\sqrt{y}$ sobre $\mathcal{X}_{[1]} = \mathbb{R}_-^*$ y $g_2^{-1}(y) = +\sqrt{y}$ sobre $\mathcal{X}_{[2]} = \mathbb{R}_+^*$; luego $p_Y(y) = \frac{p_X(\sqrt{y}) + p_X(-\sqrt{y})}{2\sqrt{y}}$, sobre $\mathcal{Y} = \mathbb{R}_+^*$.*

De nuevo, en el caso escalar, se puede salir de la función de repartición

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \sum_{k=1}^{m} P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\,y]\right)$$

(siendo $\mathcal{X}_{[0]}$ de medida nula, sobre este dominio la probabilidad es cero). Sea $\mathcal{Y}_{[k]} = g_k(\mathcal{X}_{[k]})$. Ahora, si $y \notin I(y)$,

$$P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\,y]\right) = \begin{cases} P(X \in \mathcal{X}_{[k]}) & \text{si} \quad y > \sup \mathcal{Y}_{[k]} \\ 0 & \text{si} \quad y < \text{ínf}\, \mathcal{Y}_{[k]} \end{cases}$$

dando una derivada nula. Si $y \in I(y)$,

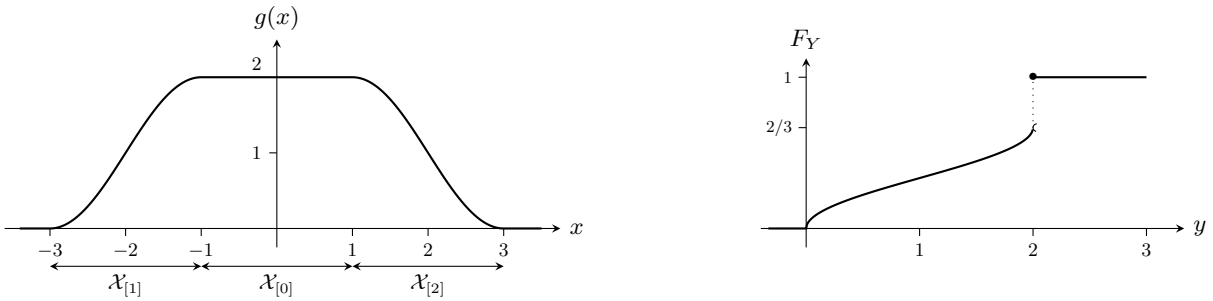
$$P\left(X \in \mathcal{X}_{[k]} \cap g_k^{-1}(-\infty\,;\,y]\right) = \begin{cases} F_X(g_k^{-1}(y)) - F_X(\text{ínf}\, \mathcal{Y}_{[k]}) & \text{si} \quad g_k \quad \text{es creciente} \\ F_X(\sup \mathcal{Y}_{[k]}) - F_X(g_k^{-1}(y)) & \text{si} \quad g_k \quad \text{es decreciente} \end{cases}.$$

El resultado sigue diferenciando estas expresiones. Se ilustra esto también en la figura 1-10.

Una tercera alternativa, a pesar de que sea delicado, es apoyarse en la teoría de distribuciones y expresar como $p_Y(y) = \int_{\mathcal{X}} p_X(x)\, \delta(y - g(x))\, dx$, donde se usa la expansión de la función delta en términos de sus ceros: $\delta(y - g(x)\,) = \sum_{k \in I(y)} \frac{1}{|g_k'(g_k^{-1}(y))|} \delta(x - g_k^{-1}(y))$ (Mandel & Wolf, 1995).

Es importante notar que la condición $\mathcal{X}_{[0]}$ de medida nula es importante. Aaso contrario, Y no resulta continua como se puede ver en el ejemplo siguiente.

Ejemplo 1-7 (Transformación con $\mu_L\left(\mathcal{X}_{[0]}\right) \neq 0$)**.** *Sea X uniforme sobre $\mathcal{X} = (3\,;\,3)$ e $Y = g(X)$ con $g(x) = \left(1 + \cos\left((|x| - 1)\frac{\pi}{2}\right)\right)\mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;\,1]}(|x|)$. Esta función se representa en la figura 1-11-(a). Claramente, g es continua y diferenciable sobre $\mathcal{X}$, pero con $\mathcal{X}_{[0]} = [-1\,;\,1]$ que no es de medida nula. Saliendo de $F_Y(y) = P(g(X) \leq y)$ se calcula sencillamente $F_Y(y) = \frac{2}{3}\left(1 - \frac{1}{\pi}\arccos(y-1)\right)\mathbb{1}_{[0\,;\,2)} + \mathbb{1}_{[2\,;\,+\infty)}(y)$, ilustrada en la figura 1-11-(b). Claramente F_Y es discontinua en $y = 2$: Y no es continua.*



(a) (b)

Figura 1-11: (a): Gráfica de $g(x) = \left(1 + \cos\left((|x| - 1)\frac{\pi}{2}\right)\right)\mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;\,1]}(|x|)$. Suponiendo que $\mathcal{X} = (-3\,;\,3)$, claramente $\mathcal{X}_{[0]} = [-1\,;\,1]$ no es de medida nula. (b): Para X uniforme sobre $\mathcal{X}$, la variable $Y = g(X)$ resulta con función de repartición F_Y no continua.

Un ejemplo de cambio de transformación puede sevir a calcular la densidad de probabilidad de una suma:

Ejemplo 1-8 (Distribución de la suma de vectores aleatorios)**.** *Sean X e Y dos vectores aleatorios d-dimensionales conjuntamente continuos, de densidad de probabilidad conjunta $p_{X,Y}$, y sea el vector*

$$V = X + Y.$$

Queremos calcular la densidad de probabilidad de V. Para esto, se puede considerar la transformación biyectiva

$$g : (x, y) \mapsto (u, v) = (x, x + y).$$

Entonces

$$g^{-1}(u, v) = (u, v - u)$$

y la matriz Jacobiana es

$$J_{g^{-1}} = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix}$$

donde la identidad I y la matriz nula 0 son en $\mathcal{M}_{d,d}(\mathbb{R})$ conjuntos de matrices $d \times d$ (ver notaciones). Claramente $\left|J_{g^{-1}}\right| = 1$ así que

$$p_{U,V}(u, v) = p_{X,Y}(u, v - u)$$

como lo podiamos intuir. Además, por marginalización, inmediatamente

$$p_V(v) = \int_{\mathbb{R}^d} p_{X,Y}(u, v-u)\, du.$$

Si X *e* Y *son independientes,* $p_{U,V}(u,v) = p_X(u)p_Y(v-u)$ *y la fórmula integral se escribe*

$$p_V(v) = \int_{\mathbb{R}^d} p_X(u)p_Y(v-u)\, du = \int_{\mathbb{R}^d} p_Y(u)p_X(v-u)\, du$$

(por cambio de variable en la secunda expresión). Esta fórmula es conocida como producto de convolución entre las funciones[14] p_X *y* p_Y *y como lo podemos ver, es conmutativo.*

1.5 Leyes condicionales

Al considerar un par de vectores aleatorios X e Y, una pregunta natural puede ser cómo caracterizar el vector Y si "observamos $X = x$". En otras palabras, la pregunta es describir la ley de Y "sabiendo (o observando) que $X = x$". En lo que sigue, para fijar notación, consideramos $(X,Y) : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}))$ tal que X es d_X-dimensional e Y es d_Y-dimensional (incluyendo los casos escalares). Escrimiremos de nuevo $\mathcal{X} = X(\Omega)$ e $\mathcal{Y} = Y(\Omega)$.

Caso X discreta: Un caso sencillo a estudiar es cuando $\mathcal{X}$ es discreto. En este caso, para cualquier $x \in \mathcal{X}$, tenemos $P_X(x) = P(X = x) \neq 0$ y de la definición de la probabilidad condicional Def. 1-3, $\forall\, B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad P(Y \in B | X = x) = \dfrac{P\big((Y \in B) \cap (X = x)\big)}{P(X = x)}$ define una medida de probabilidad que llamamos medida de probabilidad condicional. Siendo una medida de probabilidad, nos referimos a la subsección anterior para definir una función de repartición tomando $B = \prod_{i=1}^{d} (-\infty\,;\, y_i]$, caracterizando completamente la medida de probabilidad:

Definición 1-32 (Medida de probabilidad y función de repartición condicional (X discreto))**.** *Por cualquier $x \in \mathcal{X}$, la medida condicional de Y, condicionalmente a $(X = x)$, se define por*

$$\forall\, B \in \mathcal{B}\left(\mathbb{R}^{d_X}\right), \qquad P_{Y|X=x}(B) = P(Y \in B | X = x),$$

y la función de repartición condicional se define por,

$$\forall\, x \in \mathcal{X},\, y \in \mathcal{Y}, \qquad F_{Y|X=x}(y) = P(Y \leq y |\, X = x) = \frac{P\left((Y \leq y) \cap (X = x)\right)}{P\left(X = x\right)}.$$

---

[14] Un producto de convolution entre funciones se define entre cualquieras funciones, que sean densidades de probabilidad o no. Una condición suficiente para que existe tal producto es que las funciones que se convoluan sean L^1 (Golberg, 1961; Stein & Weiss, 1971; Pinsky, 2009) (es el caso de densidad de probabilidad).

Ahora, cuando Y también es discreta, se puede definir la función de masa discreta de probabilidad condicional, y si Y es continua y admite una densidad de probabilidad, se puede definir una densidad de probabilidad condicional:

Definición 1-33 (Función de masa o densidad de probabilidad condicional (X discreta))**.** *Por definición, cuando $\mathcal{Y}$ es discreta, la función de masa de probabilidad condicional de Y condicionalmente a $X = x$ es,*

$$\forall\, x \in \mathcal{X},\ y \in \mathcal{Y}, \quad p_{Y|X=x}(y) = P\left(Y = y \middle|\, X = x\right) = \frac{P\big((Y = y) \cap (X = x)\big)}{P(X = x)}.$$

Si Y es continua, es sencillo ver que $P_{Y|X=x} \ll P_Y$, i.e., para $B \in \mathcal{B}(\mathbb{R}^{d_Y})$, $P_Y(B) = 0 \Rightarrow P_{Y|X=x}(B) = 0$. Si Y admite una densidad con respecto a la medida de Lebesgue, $P_Y \ll \mu_L$ medida de Lebesgue, es claro que también $P_{Y|X=x} \ll \mu_L$, y por teorema de Radon-Nikodým 1-8, $P_{Y|X=x}$ admite una densidad de probabilidad (con respecto a la medida de Lebesgue) que denotaremos $p_{Y|X=x}$,

$$\forall\, B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad P_{Y|X=x}(B) = \int_B p_{Y|X=x}(y)\, dy.$$

A partir de la función de repartición, obtenemos

$$p_{Y|X=x}(y) = \frac{\partial^{d_Y} F_{Y|X=x}(y)}{\partial y_1 \dots \partial y_{d_Y}}.$$

Caso general: Cuando X es continua, el problema es más sutil porque $P(X = x) = 0$. Entonces, no se puede usar la definición de la probabilidad condicional, siendo el evento $(X = x)$ de probabilidad nula. Sin embargo, se pueden seguir los pasos de Rényi (Rényi, 2007, Cap. 5), de Feller (Feller, 1971, Cap. 10) o Ash & Doléans-Dade (Ash & Doléans-Dade, 1999, Sec. 5.3) por ejemplo para resolver el problema, llegando en el contexto continuo a un resultado intuitivo como en el caso discreto.

Sea $B \in \mathcal{B}(\mathbb{R}^{d_Y})$ tal que $P(Y \in B) \neq 0$ y definimos $\nu_B(A) = P\big((X \in A) \cap (Y \in B)\big)$ sobre $\left(\mathbb{R}^{d_X}, \mathcal{B}(\mathbb{R}^{d_X})\right)$. Es sencillo ver que siendo dado B, ν_B define una medida. Además, $\nu_B \ll P_X$, i.e., para $A \in \mathcal{B}(\mathbb{R}^{d_X})$, $P_X(A) = P(X \in A) = 0 \ \Rightarrow\ 0 = P\Big((X \in A) \cap (Y \in B)\Big) = \nu_B(A)$. Por teorema de Radon-Nikodým 1-8, ν_B admite una densidad $g_B = \frac{d\nu_B}{dP_X}$ con respecto a P_X,

$$\forall\, A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right), \quad P\big((X \in A) \cap (Y \in B)\big) = \int_A g_B(x)\, dP_X(x).$$

Claramente $g_B \geq 0$, y de $P(X \in A) = P\big((X \in A) \cap (Y \in B)\big) + P\big((X \in A) \cap (Y \in \overline{B})\big)$ para cualquier $A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right)$, se escribe $\forall\, A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right), \quad 0 \leq P\big((X \in A) \cap (Y \in \overline{B})\big) = \int_A dP_X(x) - \int_A g_B(x)\, dP_X(x)$, lo que permite concluire que $0 \leq g_B \leq 1$. Con el mismo razonamiento, se puede ver que $g_B(\mathbb{R}^{d_Y}) = 1$ y que g_B es σ-aditiva. En realidad, $g_B \leq 1$ P_X-casi siempre, pero olvidando esta subtilesa, g_B define una medida de probabilidad, que llamaremos *medida de probabilidad condicional*. Por continuación se define una función de repartición condicional de la misma manera que se definó la función de repartición. En resumen:

Definición 1-34 (Medida de probabilidad y función de repartición condicional)**.** *La medida de probabilidad condicional de $P_{Y|X=x}$ es definida tal que*

$$\forall\, (A, B) \in \mathcal{B}\left(\mathbb{R}^{d_X}\right) \times \mathcal{B}\left(\mathbb{R}^{d_Y}\right), \quad P\big((X \in A) \cap (Y \in B)\big) = \int_A P_{Y|X=x}(B)\, dP_X(x).$$

Tomando $B = \bigtimes_i (-\infty\,;\, y_i]$ *se obtiene la función de repartición condicional a partir de*

$$\forall\, A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right),\, y \in \mathcal{Y}, \quad P\big((X \in A) \cap (Y \leq y)\big) = \int_A F_{Y|X=x}(y)\, dP_X(x).$$

Adémas, si X *admite una densidad de probabilidad* p_X, $dP_X = p_X dx$ *y tomando* $A = \bigtimes_i (-\infty\,;\, x_i]$ *se obtiene*

$$F_{X,Y}(x,y) = \int_{\bigtimes_i (-\infty\,;\, x_i]} F_{Y|X=x}(y)\, p_X(x)\, dx$$

o, por diferenciación, para cualquier $y \in \mathcal{Y}$ *y* $x \in \mathcal{X}$ *(i. e., tal que* $p_X(x) \neq 0$*),*

$$F_{Y|X=x}(y) = \frac{\dfrac{\partial^{d_X} F_{X,Y}(x,y)}{\partial x_1 \ldots \partial x_{d_X}}}{p_X(x)}.$$

Claramente, $\left(\mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_Y}), P_{Y|X=x}\right)$ define un espacio de probabilidad y, a veces, por abuso de escritura, denotaremos $Y|X = x$ la variable aleatoria $(\Omega, \mathcal{A}, P) \to \left(\mathbb{R}^{d_Y}, \mathcal{B}(\mathbb{R}^{d_Y}), P_{Y|X=x}\right)$.

En el contexto de variable aleatorias independientes, intuitivamente conocer a X no va a llevar "información" sobre Y, lo que se formaliza de la manera siguiente:

Lema 1-5 (Probabilidad condicional e independencia)**.** *Sean* X *e* Y *vectores aleatorios independientes, entonces*

$$\forall\, x \in \mathcal{X},\, B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad P_{Y|X=x}(B) = P_Y(B)$$

A continuación, obviamente,

$$F_{Y|X=x}(y) = F_Y(y)$$

Demostración. Inmediatamente, de la independencia, tenemos

$$\begin{aligned} P\big((X \in A) \cap (Y \in B)\big) &= P(X \in A)\, P(Y \in B) \\ &= P_Y(B) \int_A dP_X(x) \\ &= \int_A P_Y(B)\, dP_X(x) \end{aligned}$$

lo que cierra la prueba. □

Ahora, tomando $A = \mathcal{X}$ en la primera fórmula que define la medida de probabilidad condicional se recupera el equivalente continuo de la fórmula de probabilidad total:

Teorema 1-14 (Fórmula de probabilidad total (caso general))**.**

$$\forall\, B \in \mathcal{B}\left(\mathbb{R}^{d_Y}\right), \qquad P(Y \in B) = \int_{\mathcal{X}} P_{Y|X=x}(B)\, dP_X(x)$$

lo que da en termino de función de repartición condicional

$$F_Y(y) = \int_{\mathcal{X}} F_{Y|X=x}(y)\, dP_X(x)$$

Si P_X *admite una densidad, se escribe todo notando que* $dP_X(x) = p_X(x) d\mu_L(x) \equiv p_X(x) dx$.

Por último, si (X, Y) admite una densidad, en sencillo ver que $P_{Y|X=x} \ll \mu_L$, y entonces $P_{Y|X=x}$ admite una densidad que llamaremos *densidad de probabilidad condicional*. Sean $A \in \mathcal{B}\left(\mathbb{R}^{d_X}\right)$ y $B \in \mathcal{B}\left(\mathbb{R}^{d_Y}\right)$,

$$\begin{aligned} P\Big((X \in A) \cap (Y \in B)\Big) &= \int_{A\times B} p_{X,Y}(x,y)\,dx\,dy \\ &= \int_B \left(\int_A \frac{p_{X,Y}(x,y)}{p_X(x)}\,dy\right) p_X(x)\,dx \end{aligned}$$

Entonces, si $p_X(x) \neq 0$, tenemos

$$P_{Y|X=x}(A) = \int_A \frac{p_{X,Y}(x,y)}{p_X(x)}\,dy.$$

Teorema 1-15 (Densidad de probabilidad condicional)**.** *Si (X, Y) admite una densidad de probabilidad, la medida de probabilidad condicional $P_{Y|X=x}$ admite una densidad, llamada densidad de probabilidad condicional definida por*

$$\forall\, x \in \mathcal{X}, \quad p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

definida sobre $\mathcal{Y}$. Claramente, a partir de la función de repartición condicional resulta que

$$p_{Y|X=x} = \frac{\partial^{d_Y} F_{Y|X=x}}{\partial y_1 \ldots \partial y_{d_Y}}.$$

De hecho, esta construcción rigurosa coincide con la intuición que podemos tener en este caso continuo. Por ejemplo, podemos pensar a $F_{Y|X=x}(y)$ como caso límite de $P\left(Y \leq y \,\middle|\, x \leq X \leq x + \delta x\right) = \frac{P\big((Y \leq y) \,\cap\, (x \leq X \leq x+\delta x)\big)}{P\,(x \leq X \leq x+\delta x)} = \frac{F_{X,Y}(x+\delta x, y) - F_{X,Y}(x,y)}{F_X(x+\delta x) - F_X(x)}$ cuando δx tiende a 0. En el caso escalar, se calcula por ejemplo haciendo un desarrollo de Taylor del numerador y del denominador a orden 1, o usando la regla de l'Hôpital [15] para re-obtener la función de repartición condicional de la definición **??**. En el caso multivariado, hace falta hacer los desarollos hasta el orden d_X para concluir.

Notar que:

- si X e Y son independientes,

$$p_{Y|X=x} = p_Y;$$

- por la expresión $p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$, por integración con respecto a y obtenemos la condición de normalización

$$\int_{\mathcal{Y}} p_{Y|X=x}(y)\,dy = 1.$$

También, se escribe la fórmula de probabilidad total a través de las densidades por la expresión $p_{X,Y}(x,y) = p_{Y|X=x}(y)\,p_X(x)$ y luego por integración con respecto a x:

[15] De hecho, esta regla es debido al suizo J. Bernoulli que tuvo un acuerdo financiero con el Guillaume François Antoine, marqués de l'Hôpital, permitiendolo de publicar unos resultados de Bernoulli bajo su nombre.

Teorema 1-16 (Fórmula de probabilidad total (caso con densidades)). *Si (X, Y) admite una densidad de probabilidad conjunta $p_{X,Y}$, entonces Y tiene una densidad de probabilida que se recupera a través de la fórmula*

$$p_Y(y) = \int_{\mathcal{X}} p_{Y|X=x}(y)\, p_X(x)\, dx.$$

De la expresión la densidad condicional, $p_{X,Y}(x, y) = p_{Y|X=x}(y)\, p_X(x) = p_{X|Y=y}(x)\, p_Y(y)$, y de la fórmula de probabilidad total se recupera sencillamente el equivalente continuo de la formula de Bayés:

Teorema 1-17 (Fórmula de Bayes (caso continuo)).

$$\forall\, y \in \mathcal{Y},\ x \in \mathcal{X}, \qquad p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)\, p_Y(y)}{\displaystyle\int_{\mathcal{Y}} p_{X|Y=y}(x)\, p_Y(y)\, dy}.$$

Volvemos ahora al ejemplo 1-8:

Ejemplo 1-9 (Distribución condicional de la suma de vectores aleatorios). *Sea $V = X + Y$, con X e Y vectores d-dimensionales. Introduciendo $U = X$ obtuvimos $p_{U,V}(u, v) = p_{X,Y}(u, v - u)$ dando también $p_V(v) = \displaystyle\int_{\mathbb{R}^d} p_{X,Y}(u, v - u)\, du$. Entonces, recordando que $U = X$, se obtiene*

$$p_{V|X=x}(v) = \frac{p_{X,Y}(x, v - x)}{p_X(x)} = \frac{p_{X,Y}(x, v - x)}{\displaystyle\int_{\mathbb{R}^d} p_{X,Y}(x, v - x)\, dv},$$

dando en el caso X e Y independientes

$$p_{V|X=x}(v) = p_Y(v - x).$$

Esto corresponde a la intución de que, con $V = X + Y$, fijando $X = x$ el vector aleatorio V es nada más que Y desplazado en x. Pero hay que tomar muchas precauciones con este razonamiento, valido únicamente cuando X e Y son independientes. En caso contrario, fijando X no coincide con un desplazamiento por la dependencia (esquematícamente, fijando X no sólo mueve Y sino que "cambia" su estadística).

1.6 Esperanza, momentos, identidades y desigualdades

Como lo hemos introducido, la noción formal de probabilidad nació en el contexto de juego (cartas, dados), bajo entre otros el impulso del matemático italiano y jugador de dados y cartas Gerolamo Cardano en el siglo XVI, y aún más bajo el trabajo muy profundo de la dinastía Bernoulli, y en particular Jacob Bernoulli (Bernoulli, 1713, en latín) o ((E. D. Sylla, Translator), 1713). En particular, Bernoulli se interesó no solamente al resultado de un sorteo, impredictible, pero en lo que pasa cuando se hace un gran numéro de sorteos. Así salió la idea de "resultado promedio". En realidad el trabajo de Bernoulli ha ido más adelante: probó la ley de gran número que vamos a ver más adelante. Sin

ir tan lejos, en el caso de una variable X aleatoria discreta (ej. un dado, que puede tomar valores en $\{1\,;\,\ldots\,;\,6\}$), haciendo varios sorteos, el promedio de estes sorteos va a ser $\sum_i \widetilde{p}_i x_i$ con x_i los valores que puede tomar la variable y $\widetilde{p}_i$ la proporción de x_i que se obtuvo en el sorteo. De hecho, intuitivamente (es la visión frecuencista), $\widetilde{p}_i$ va a tender a $p_i = P(X = x_i)$ cuando el número de sorteo tiende al infinito. La definición del valor promedio, o media, de una variable aleatoria cualquiera se formaliza mas rigorosamente en en marco de la teoria de la medida, pero coincide con la intuición.

1.6.1 Media de un vector aleatorio

Una variable aleatoria X tiene asociado un *promedio* o *media* (también llamado *valor esperado o de expectación o esperanza matemática*) que se obtiene pesando cada valor de X con la medida de probabilidad asociada a ese valor (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006),

Definición 1-35 (Media o valor/vector medio)**.** *Formalmente, la media de una variable aleatoria X integrable es definida por*

$$\mathrm{E}[X] = \int_\Omega X(\omega)\, dP(\omega).$$

Por el teorema de la medida imagen 1-2, pagina 25, esta media se escribe también a partir de la medida de probabilidad P_X como

$$\mathrm{E}[X] = \int_{\mathbb{R}} x\, dP_X(x).$$

En el caso vectorial d-dimensional, hay que entender la media, o vector medio, como un vector de componentes i-ésima la media $\mathrm{E}[X_i]$ de la componente i-ésima X_i de X, dando

$$\mathrm{E}[X] = \int_{\mathbb{R}^d} x\, dP_X(x).$$

A veces, se encuentra también la notación $\langle x\rangle$ o $\langle x\rangle_{p_X}$ para el valor medio, especialmente en la literatura de física.

La secunda formulación del valor medio se proba sencillamente, empezando por $X = \mathbb{1}_A$ para unos $A \in \mathcal{B}(\mathbb{R})$. Entonces $P_X = (1 - P(A))\delta_0 + P(A)\delta_1$. Luego $\int_\Omega \mathbb{1}_A(\omega) dP(\omega) = P(A) = (1 - P(A)) \times 0 + P(A) \times 1 = \int_{\mathbb{R}} x dP_X(x)$. Se cierra la prueba con el teorema 1-3 dando cualquier función medible como limite de funciones escalonadas, y por la definición 1-10 de la integral de cualquier función medible.

Luego, de la distribución marginal $P_{X_i}(B) = \int_{\mathbb{R}^{i-1}\times B\times\mathbb{R}^{d-i}} dP_X(x)$, se obtiene $\mathrm{E}[X_i] = \int_{\mathbb{R}^d} x_i\, dP_X(x)$, dando la última formulación en el caso vectorial.

Una variable aleatoria X se dice integrable cuando $E[|X|] < \infty$. De la misma manera, un vector aleatorio admite una media si y solamente si cada componente es integrable. Veremos más adelante que existen variables aleatorias que no admiten una media.

Más allá de la formulación matemática de la media $\mathrm{E}[X]$ representa la posición alrededor de la cual se "distribuye las probabilidades de occurencia". Es el equivalente probabilístico de centro de gravedad o barycentro en mecánica.

En el caso de variables aleatorias discretas, de soporte $\mathcal{X}$ discreto finito o numerable,inmediatamente

$$\mathrm{E}[X] = \sum_{x\in\mathcal{X}} x\, P(X = x) = \sum_{x\in\mathcal{X}} x\, p_X(x).$$

Fijense de que $\mathrm{E}[X]$ no partenece necesariamente a $\mathcal{X}$:

Ejemplo 1-10. *Sea X uniforme sobre $\mathcal{X} = \{1\,;\,3\,;\,7\}$, i. e., $\forall\, i \in \mathcal{X},\quad P(X = i) = \frac{1}{3}$. Se calcula $\mathrm{E}[X] = 1\times\frac{1}{3} + 3\times\frac{1}{3} + 7\times\frac{1}{3} = \frac{11}{3} \notin \mathcal{X}$. Tampoco es el promedio de los valores extremos.*

Cuando $|\mathcal{X}| = +\infty$, X no es necesariamente integrable:

Ejemplo 1-11. *Sea $\mathcal{X} = \mathbb{N}^*$ con $P(X = x) = \frac{6}{\pi^2\, x^2}$. Claramente, $\sum_x \frac{6}{\pi^2\, x}$ diverge, así que X no tiene una media.*

En el caso de vectores aleatorios continuos, obtenemos la expresión siguiente de la media (o vector medio):

$$\mathrm{E}[X] = \int_{\mathbb{R}^d} x\, p_X(x)\, dx.$$

Las mismas observaciones que hicimos en el caso discreto se encuentra en el caso continuo:

Ejemplo 1-12. *Sea X de densidad de probabilidad $p_X(x) = \frac{1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \frac{3\sqrt{x-2}}{4}\mathbb{1}_{[2\,;\,3)}(x)$ como ilustrado figura Fig. 1-6, pagina 39. Se calcula $\mathrm{E}[X] = \frac{31}{20} \notin \mathcal{X} = [0\,;\,1]\cup[2\,;\,3]$.*

Ejemplo 1-13. *Un ejemplo de vector aleatorio no teniendo media es dado en el caso de una distribución de Cauchy-Lorentz (ver más adelante) $p_X(x) = \dfrac{\alpha}{(1 + x^t x)^{\frac{d+1}{2}}}$ donde α es un factor de normalización.*

En el caso general, para calcular la media, hay que pasar por la distribución P_X, como en el ejemplo 1-4 pagina 39:

Ejemplo 1-14 (Continuación del ejemplo 1-4)**.** *Sea $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con U y V variables aleatorias independientes de distribución uniformas sobre $[0\,;\,1)$, i. e., $p_U(x) = \mathbb{1}_{[0\,;\,1)}(x)$. De $(X \in B) \;\Leftrightarrow\; \big((U < \frac{1}{2}) \cap (V \in B)\big) \cup \big((U \geq \frac{1}{2}) \cap (1 \in B)\big)$, del hecho de que los eventos de la unión son incompatibles y de la independencia de U y V (o saliendo de la función de repartición), se obtiene $P_X(B) = \frac{1}{2}P_V(B) + \frac{1}{2}\delta_1(B)$. A continuación, $\mathrm{E}[X] = \dfrac{1}{2}\displaystyle\int_{\mathbb{R}} dP_V(x) + \frac{1}{2}\int_{\mathbb{R}} d\delta_1(x) = \frac{1}{2}\int_{\mathbb{R}} p_V(x)\, dx + \frac{1}{2}\times 1 = \frac{1}{2}\int_0^1 dx + \frac{1}{2} = \frac{3}{4}$.*

Una nota intersante es de que, en el caso escalar, si $X \geq 0$ admitiendo una media, se obtiene

$$\mathrm{E}[X] = \int_{\mathbb{R}_+} P(X > t)\, dt = \int_{\mathbb{R}_+} (1 - F_X(t))\; dt.$$

Se proba saliendo de $x = \int_0^x dt = \int_{\mathbb{R}_+} \mathbb{1}_{(t\,;+\infty)}(x)\,dt$ dando $\mathrm{E}[X] = \int_{\mathbb{R}} \left(\int_{\mathbb{R}_+} \mathbb{1}_{(t\,;+\infty)}(x)\,dt \right) dP_X(x) = \int_{\mathbb{R}_+} \left(\int_{\mathbb{R}} \mathbb{1}_{(t\,;+\infty)}(x)\,dP_X(x) \right) dt$ por el teorema de Fubini Th. 1-6 pagina 28. Se cierra la prueba observando que la integral interior es nada más que $P(X > t)$. En el caso discreto con $\mathcal{X} = \mathbb{N}$, viene inmediatamente $\sum_{t\in\mathbb{N}} P(X > t)$ que podemos probar directamente saliendo de $P(X = t) = P(X > t) - P(X > t-1)$. En el caso de variable admitiendo una densidad, se lo obtiene también haciendo una integración por partes [16]

Esta fórmula se aplica al ejemplo 1-4 que tratamos:

Ejemplo 1-15 (Continuación del ejemplo 1-4)**.** *Sea $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ con U y V variables aleatorias independientes de distribución uniformas sobre $[0\,;1)$. Obtuvimos pagina 39 $F_X(x) = \frac{x}{2}\,\mathbb{1}_{[0\,;1)}(x) + \mathbb{1}_{[1\,;+\infty)}(x)$. A continuación, reobtenemos $\mathrm{E}[X] = \int_0^1 \left(1 - \frac{x}{2}\right) dx = \frac{3}{4}$.*

.

Terminamos esta sección con la propiedad de linealidad de la esperanza matemática E, como consecuencia de la linealidad de la integración y definición de la distribución marginal: para cualquier conjunto de vectores aleatorios $\{X_i\}$ integrables y cualquieras matrices $\{C_i\}$ dadas de dimensiones compatibles con las de X (incluyendo el caso escalar),

$$\mathrm{E}\left[\sum_i C_i X_i\right] = \sum_i C_i\,\mathrm{E}\left[X_i\right]$$

(la integrabilidad de la suma se proba a partir de la desigualdad triangular).

1.6.2 Momentos de un vector aleatorio

Si X es una variable aleatoria, para cualquier función medible f, $f(X)$ también lo es. Se puede entonces definir su valor medio, si existe. A pesar de necesitar evaluar la distribución de probabilidad de $Y = f(X)$, el valor medio se calcula a partir del de X:

Teorema 1-18 (Teorema de transferencia)**.** *Sea X un vector aleatorio d-dimensional y $f : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ una función medible tal que $f(X)$ sea integrable. Entonces*

$$\mathrm{E}\left[f(X)\right] = \int_\Omega f(X(\omega))\,dP(\omega) = \int_{\mathbb{R}^{d'}} f(x)\,dP_X(x).$$

En particular, en el caso $\mathcal{X} = X(\Omega)$ discreto se obtiene

$$\mathrm{E}\left[f(X)\right] = \sum_{x\in\mathcal{X}} f(x)\,P(X = x)$$

[16] El el casi discreto, hay que tener precauciones separando la series de una diferencia de terminos. En el caso X continuo admitiendo una densidad, hay que estudiar bien el coomportamiento de $t \mapsto t(1 - F_X(t))$ al infinito.

y para X continuo admitiendo una densidad de probabilidad

$$\mathrm{E}\left[f(X)\right] = \int_{\mathbb{R}^d} f(x)\, p_X(x)\, dx.$$

Demostración. Sea $B \in \mathcal{B}(\mathbb{R}^d)$ y consideramos $f(x) = \mathbb{1}_B(x)$. Entonces, $\mathcal{Y} = \{0\,;\,1\}$ y inmediatamente

$$P_Y = P_X(B)\,\delta_1 + (1 - P_X(B))\,\delta_0.$$

Entonces

$$\mathrm{E}[f(X)] = \int_{\mathbb{R}} P_X(B)\, d\delta_1 + \int_{\mathbb{R}} (1 - P_X(B))\, d\delta_0 = P_X(B) = \int_{\mathbb{R}^d} \mathbb{1}_B(x)\, dP_X(x).$$

En el caso $d' = 1$, para $f \geq 0$, se cierra entonces la prueba usando el teorema 1-3 pagina 27, escribiendo f como límite creciente de una sucesión de funciones escalonadas, y la definición Def. 1-10 de la integración real. El caso $d' > 1$ es nada mas que $d' = 1$, componente a componente. □

De manera general, estas medias son llamadas *momentos* de la variable aleatoria X. Los momentos relevantes usuales son los siguientes:

- para el "monomio" $f(x) = x^{\otimes K}$ producto tensiorial de x K veces [17] siendo $K \in \mathbb{N}^*$, se obtiene el tensor de los K*-ésimo momentos (ordinarios)* de X:

  $$m_K \equiv \mathrm{E}\left[X^{\otimes K}\right] = \int_{\mathbb{R}^d} x^{\otimes K}\, dP_X(x)$$

  que tiene unidades de $\prod_j X_{i_j}$ (de X_i^K si los componentes de X tienen la misma "unidad"). Se escribe también los momentos de orden K como

  $$m_{k_1,\ldots,k_d} = \mathrm{E}\left[\prod_{i=1}^{d} X_i^{k_i}\right] \quad \text{con} \quad \sum_i k_i = K.$$

  Se puede incluir el caso $K = 0$ con la convención $x^{\otimes 0} = 1$, que corresponde a la condición de normalización: $m_0 = \int_{\mathbb{R}} dP_X(x) = 1$. La media es el primer momento: $m_1 = \mathrm{E}[X] = m_X$. Típicamente, los primeros momentos son más relevantes que los de órdenes mayores, para la caracterización de una distribución. Para $K = 2$, en el caso escalar, el momento de orden 2 es el análogo del momento de inercia de la mecánica.

  Por ejemplo, para la distribución uniforme $p_X(x) = \frac{1}{b-a}$ en el intervalo $[a\,;\,b]$, resulta $m_K = \frac{b^{K+1} - a^{K+1}}{(K+1)(b-a)}$. En particular, $m_1 = \frac{a+b}{2}$, valor medio del intervalo.

  Fijense de que $X^{\otimes K}$ no es siempre integrable, por ejemplo, en el caso con densidad, si $p_X(x)$ tiene soporte (semi)infinito, necesariamente la función p_X debe tender a 0 cuando

[17] Recuerdense de que $x \otimes x$ es una matriz teniendo como componentes $x_i x_j$; entonces $x^{\otimes K}$ en un tensor K-dimensional teniendo como componentes $\left[x^{\otimes K}\right]_{i_1,\ldots,i_K} = \prod_j x_{i_j}$.

$\|x\| \to \infty$. Si $p_X(x)$ es *de largo alcance*, en el sentido de que no cae a 0 suficientemente rápido con x para x grandes, algunos momentos pueden no existir. Por ejemplo, la distribución de probabilidad de Cauchy–Lorentz (o función de Breit–Wigner), dada por $p_X(x) = \alpha \left(1 + (x - x_0)^t R^{-1}(x - x_0)\right)^{-\frac{d+1}{2}}$ sobre $\mathbb{R}^d$, con la matriz cuadrada $R > 0$, $x_0 \in \mathbb{R}^d$ y $\alpha > 0$ coeficiente de normalización, no tiene momentos finitos de orden $K \geq 1$.

- Frecuentemente (especialmente en el caso de variables discretas X sobre $\mathcal{X} = \mathbb{N}$), resulta útil introducir los K-ésimo *momento factorial* de X, $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathbb{N}^d$, $K = \sum_{i=1}^d k_i$ mediante

$$f_{k_1,\ldots,k_d} = \mathrm{E}\left[\prod_{i=1}^d (X_i)^{\underline{k_i}}\right]$$

con $(x)^{\underline{n}}$ factorial decreciente (ver notaciones). Se puede ver que cuando $\mathcal{X} = \{0\,;\,\ldots\,;\,n\}$, $n \in \mathbb{N}$, $f_K = 0$ cuando $K > n$.

- Los *momentos centrales* se definen alrededor de la media $\mathrm{E}[X]$, *i. e.*,, como el tensor de los K-ésimo momentos de la *desviación* $\Delta X \equiv X - \mathrm{E}[X]$:

$$\zeta_K \equiv \mathrm{E}\left[(X - \mathrm{E}[X])^{\otimes K}\right].$$

Se escribe también

$$\zeta_{k_1,\ldots,k_d} = \mathrm{E}\left[\prod_{i=1}^d (X_i - \mathrm{E}[X_i])^{k_i}\right] \quad \text{con} \quad \sum_i k_i = K.$$

Se deduce que si la distribución de probabilidad satisface a una simetría central con respecto a la media, *i. e.*, $X - m_X \overset{\mathrm{d}}{=} -(X - m_X)$ donde $\overset{\mathrm{d}}{=}$ significa que los vectores aleatorios tiene la misma distribución de probabilidad, entonces todos los momentos centrales impares son nulos. Los momentos (centrales) brindan medidas que caracterizan la distribución.

1. El primer momento, o media:

$$m_X = \mathrm{E}[X].$$

2. El segundo momento central se conoce como *matriz de covarianza*. En el caso escalar, hablamos de *varianza*, o *dispersión* o también *desviación cuadrática media*.

$$\Sigma_X \equiv \mathrm{Cov}[X] \equiv \zeta_2 = \mathrm{E}\left[(X - m_X)\,(X - m_X)^t\right].$$

En el caso escalar, la varianza se escribe en general

$$\mathrm{Var}[X] \equiv \sigma_X^2 = \mathrm{E}\left[(X - m_X)^2\right]$$

y es una medida del cuadrado del ancho efectivo de una densidad de probabilidad (o vector de probabilidad). Para dos vectores aleatorios X e Y respectivamente d y d'-dimensonal (con d' no necesariamente igual a d), hablamos de *covarianza entre X e Y*, que se escribe

$$\Sigma_{X,Y} = \equiv \mathrm{Cov}[X, Y] = \mathrm{E}\left[(X - m_X)\,(Y - m_Y)^t\right].$$

Esta matriz partenece a $\mathcal{M}_{d,d'}(\mathbb{R})$ y

$$\Sigma_{X,Y} = \Sigma_{Y,X}^t.$$

.

Se puede notar que, para dos componentes $i \neq j$ de X, Σ_X tiene como (i,j)-ésima componente la covarianza $\mathrm{Cov}[X_i, X_j] = \mathrm{E}\left[(X_i - m_{X_i})\left(X_j - m_{X_j}\right)\right]$ entre las variables X_i y X_j y tiene las varianzas de los X_i en su diagonal. Además, es sencillo ver que $\Sigma_X \in P_d(\mathbb{R})$, *i. e.,* $\mathrm{Cov}[X]$ es simétrica, por construcción, y $\mathrm{Cov}[X] \geq 0$ ($\forall\, \mu \in \mathbb{R}, \quad \mu^t \Sigma_X \mu \geq 0$; en el caso escalar la varianza es no negativa), con igualdad sólo cuando $P_X = \delta_{x_0}$ para un x_0 dado, esto es, cuando no hay incerteza sobre el resultado. De la desigualdad de Cauchy-Bunyakovsky-Schwarz (ver corolario **??**, pagina **??**) se proba sencillamente de que

$$|\mathrm{Cov}[X_i, X_j]|^2 \leq \sigma_{X_i}^2 \sigma_{X_j}^2,$$

así que se define también el *coeficiente de correlación* que es adimensional y toma valores entre -1 (variables completamente anticorrelacionadas) y 1 (variables completamente correlacionadas) como:

$$\rho_{ij} = \rho_{ji} \equiv \frac{\mathrm{Cov}[X_i, X_j]}{\sigma_{X_i} \sigma_{X_j}}.$$

Como ejemplo, dadas X_1 y $X_2 = aX_1 + b$ que fluctúan en fase ($a > 0$) o al revés ($a < 0$), se tiene la relación entre desviaciones $\Delta X_2 = a \Delta X_1$, conduciendo a $\rho_{12} = \frac{a}{|a|} = \pm 1$.
También, se puede ver que

$$\mathrm{Var}\left[\|\Delta X\|\right] = \mathrm{Tr}\, \Sigma_X.$$

La covarianza está bien definida si $\|X\|$ es una variable aleatoria de cuadrado integrable, esto es, cuando $E\left[\|X\|^2\right] < \infty$. Se proba sencillamente (desallorando el "producto" y usando la linealidad de la esperanza) que

$$\mathrm{Cov}[X, Y] = \mathrm{E}\left[XY^t\right] - m_X m_Y^t$$

conocido como *teorema de König-Huygens*. En el caso escalar y $X = Y$, es el equivalente del teorema de Huygens de la mecánica relacionando el momento de inertia de un solido con respeto al origen en función del momento de inertia con respeto al centro de masa. Además, inmediatamente,

$$\forall\, A \in \mathcal{M}_{n,d}(\mathbb{R}),\ B \in \mathcal{M}_{n',d'}(\mathbb{R}),\ a \in \mathbb{R}^n,\ b \in \mathbb{R}^{n'}, \quad \mathrm{Cov}[AX + a, BY + b] = A\, \mathrm{Cov}[X, Y] B^t.$$

En el caso escalar, $d = 1$, lo que es conocido también como el *ancho* de una distribución está dado por la *desviación estándar*

$$\sigma_X = \sqrt{\mathrm{Var}[X]}$$

tiene las mismas unidades de X, y se usa para normalizar los momentos centrales de orden superior. El *ancho relativo* es otra medida que caracteriza la distribución, dado por

$\frac{\sigma_X}{m_X} = \sqrt{\frac{\mathrm{E}[X^2]}{m_X^2} - 1}$ cuando $m_X \neq 0$.

Dado un vector aleatorio X, teniendo en cuenta que los dos primeros momentos dan las características más importantes de la distribución de probabilidad, puede resultar conveniente hacer una transformación de variable aleatoria a la llamada *variable estándar*: $Y \equiv \Sigma_X^{-\frac{1}{2}} (X - m_X)$, donde $\Sigma^{-\frac{1}{2}}$ es la única matriz simétrica definida positiva tal que su cuadrado es igual a Σ^{-1} (Horn & Johnson, 2013; Magnus & Neudecker, 1999) que entonces tiene media igual a 0 y una matriz de covarianza igual al identidad I (en el caso escalar, desviación estándar igual a 1).

3. En el caso escalar, el tercer momento central permite definir el *coeficiente de sesgo* o más sencillamente *sesgo* (o skewness en ingles) (Spiegel, 1976; Pearson, 1905):

$$\mathrm{Ses}[X] \equiv \gamma_X \equiv \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^3\right] = \frac{\zeta_3}{\sigma_X^3},$$

momento de orden 3 de la variable estandar, que resulta adimensional y puede tener signo positivo o negativo, anulándose para distribuciones que son simétricas respecto del valor medio.

4. En el caso escalar, el cuarto momento central da lugar a la *curtosis* (Pearson, 1905; Westfall, 2014):

$$\mathrm{Curt}[X] \equiv \kappa_X \equiv \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^4\right] = \frac{\zeta_4}{\sigma_X^4},$$

momento de orden 4 de la variable estandar, que posibilita diferenciar entre distribuciones altas y angostas. Veremos más adelante de que para la densidad Gausiana $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\,\sigma^2}\right)$, $m_X = m$, $\sigma_X = \sigma$, $\gamma_X = 0$, $\kappa_X = 3$. Se dice de que p_X es alta y angosta, o sub-gausiana, o *con colas livianas* o también platicúrtica cuando $\kappa_X < 3$, y se dice bajas y anchas o sobre-gausiana, o *con colas pesadas* o también leptocúrtica cuando $\kappa_X > 3$ (para $\kappa_X = 3$ la distribución es a veces dicha mesocúrtica). A veces, se define entonces la *curtosis por exceso*

$$\mathrm{ExCurt}[X] \equiv \bar{\kappa}_X = \mathrm{Curt}[X] - 3 = \mathrm{E}\left[\left(\frac{X - m_X}{\sigma_X}\right)^4\right] - 3 = \frac{\zeta_4}{\sigma_X^4} - 3.$$

Más que el pico de distribución, la curtosis (por excesso) describe las colas de una distribución (pesadas para el curtosis por exceso o livianas en el caso contrario) (Westfall, 2014).

Fijense de que, en el contexto escalar $d = 1$, se vinculan los cumulantes y los momentos ordinarios directamente de las definiciones:

$$\zeta_k = \sum_{l=0}^{k} \binom{k}{l} (-m_X)^{k-l}\, m_l$$

para cualquier $k \in \mathbb{N}$, siendo $m_0 = \zeta_0 = 1$. Por ejemplo, $\zeta_2 = m_2 - m_1^2$ que es nada más que la relación de König-Huyggens, mientras que $\zeta_3 = m_3 - 3m_1m_2 + 2m_1^3$. En el contexto multivariado,

la relación momentos–momentos centrales toma la expresión

$$\zeta_{k_1,\ldots,k_d} = \sum_{l_1=0}^{k_1} \cdots \sum_{l_d=0}^{k_d} \left(\prod_{i=1}^{d} \binom{k_i}{l_i} \left(-m_{X_i}\right)^{k_i - l_i} \right) m_{l_1,\ldots,l_d}.$$

1.6.3 Independencia, identidades y desigualdades

Una primera relación interesante concierna el caso de variables independientes y como se comporta la covarianza de estas:

Lema 1-6 (Independencia y covarianza)**.** *Sean X e Y dos vectores aleatorios integrables. Si son independientes, entonces*

$$\mathrm{E}[XY^t] = \mathrm{E}[X]\,\mathrm{E}[Y]^t \qquad \textit{i. e.,} \qquad \mathrm{Cov}[X,Y] = 0.$$

En particular, para X con componentes independientes, $\mathrm{Cov}[X]$ es una matriz diagonal.

Demostración. Sean $X = \sum_j x_j \mathbb{1}_{A_j}$ e $Y = \sum_k y_k \mathbb{1}_{B_k}$ dos variables escalonadas. Entonces, $A_j = (X = x_j)$ y $B_k = (Y = y_k)$. Luego

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{j,k} x_j y_k \,\mathrm{E}\left[\mathbb{1}_{A_j}\mathbb{1}_{B_k}\right] \\
&= \sum_{j,k} x_j y_k \,\mathrm{E}\left[\mathbb{1}_{A_j \cap B_k}\right] \\
&= \sum_{j,k} x_j y_k P(A_j \cap B_k) \\
&= \sum_{j,k} x_j y_k P(X = x_j) P(Y = y_k) \quad \text{(de la independencia)}
\end{aligned}$$

dando el resultado para variables escalonadas. Se cierra la prueba para variables positivas como límite de crecientes de funciones escalonadas, y variables reales tratando las partes positivas y negativas aparte. El caso vectorial se deduce trabajando con pares de componentes. □

Fijense de que la reciproca es falsa en general:

Ejemplo 1-16 (Uniforme sobre el disco unitario)**.** *Sea $X = (X_1, X_2)$ uniforme sobre el disco unitario o bola unitaria 2-dimensional $\mathbb{B}_2$ (ver notaciones), i.e., $p_X(x) = \frac{1}{\pi}\mathbb{1}_{\mathbb{S}^2}(x)$. Claramente, los X_i no pueden ser independientes del hecho que $\mathcal{X}_i = [-1\,;\,1]$ y $\mathcal{X} \neq \mathcal{X}_1 \times \mathcal{X}_2$ (es estrictamente incluido en el producto cartesiano). Por simetría central de p_X, es sencillo ver que $\mathrm{E}[X_1 X_2] = 0$ y similarmente $\mathrm{E}[X_i] = 0$: a pesar de que los X_i no sean independientes, $\mathrm{Cov}[X_1, X_2] = 0$.*

La consecuencia de la independencia sobre la covarianza facilita frecuentemente los calculos de media. Volviendo al ejemplo 1-4 de la pagina 39:

Ejemplo 1-17 (Continuación del ejemplo 1-4)**.** *Tratando de la media de* $X = V\,\mathbb{1}_{U<\frac{1}{2}} + \mathbb{1}_{U\geq\frac{1}{2}}$ *con* U *y* V *variables independientes de distribución uniformas sobre* $(0\,;\,1)$*, se calcula gracia a la linealidad y a la independencia,* $\mathrm{E}[X] = \mathrm{E}[V]\,\mathrm{E}\left[\mathbb{1}_{U<\frac{1}{2}}\right] + \mathrm{E}\left[\mathbb{1}_{U\geq\frac{1}{2}}\right] = \frac{1}{2}\times\frac{1}{2} + \frac{1}{2} = \frac{3}{4}$ *como lo hemos obtenido usando* P_X *en la pagina 58 o la positividad en la pagina 59.*

Una otra consecuencia de esta proposición trata de un conjunto de vectores aleatorios $\{X_i\}$ y un conjunto de matrices de dimensiones adecuadas,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i \Sigma_{X_i} A_i^t + \sum_{j\neq i} A_i\,\mathrm{Cov}[X_i, X_j] A_j^t.$$

En particular, en el caso escalar,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i^2\,\mathrm{Var}[X_i] + \sum_{j\neq i} A_i A_j\,\mathrm{Cov}[X_i, X_j].$$

Si los X_i son independientes, entonces las covarianzas conjuntas son nulas así que, respectivamente,

$$\mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i \Sigma_{X_i} A_i^t \qquad \text{y} \qquad \mathrm{Cov}\left[\sum_i A_i X_i + B\right] = \sum_i A_i^2 \sigma_{X_i}^2.$$

Si el teorema da una implocación de la independencia, de hecho existe una reciproca que toma la forma siguiente:

Teorema 1-19 (Independencia y momentos)**.** *Sean* X *e* Y *dos vectores aleatorios. Son independientes si y sólo si* $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ *para todo par de funciones* f *y* g*, medibles y acotadas de dimensiones adecuadas.*

Demostración. Se puede referirse a (Feller, 1971; Jacob & Protters, 2003) para unas pruebas rigurosa. En el caso escalar, el principio consiste a ver f y g como limites de funciones escalonadas. Para $f(x) = \sum_i a_i \mathbb{1}_{A_i}(x)$ y $g(y) = \sum_j b_j \mathbb{1}_{B_j}(y)$ se obtiene $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ si y sólo si $\sum_{i,j} a_i b_j \left(P((X\in A_i)\cap(Y\in B_j)) - P(X\in A_i)P(Y\in B_j)\right) = 0$. Básicamente, eso debe valer para cualquieras A_i, B_j y a_i, b_j, así que el término entre parentesis debe ser cero, lo que es nada más de la definición de la independencia de X e Y. El caso vectorial se entiende por pares de componentes. □

Relaciones tambíen muy útiles son conocidas como *Desigualdades de Chebyshev* (Bienaymé, 1853; Tchébichev, 1867; Markov, 1884; Olkin & Pratt, 1958; Ferentinos, 1982; Navarro, 2013; Stellato, Van Parys & Goulart, 2017). Estas desigualdades dan una cota superior a la probabilidad de que una cantidad que fluctúa aleatoriamente exceda cierto valor umbral, aún sin conocer detalladamente la forma de la distribución de probabilidad.

Teorema 1-20 (Desigualdades de Chebyshev)**.** *Sea un vector aleatorio* d*-dimensional* X *y una función* $g : \mathbb{R}^d \mapsto \mathbb{R}_+$ *medible tal que* $g(X)$ *sea integrable. Entonces,*

$$\forall\, a > 0, \quad P(g(X)\geq a) \;\leq\; \frac{\mathrm{E}[g(X)]}{a}.$$

Demostración. Sea $\mathcal{D}_a = \{x \in \mathcal{X} \mid g(x) \geq a\} \subset \mathcal{X}$. Entonces, g siendo non negativa,

$$\mathrm{E}[g(X)] = \int_{\mathcal{X}} g(x)\, dP_X(x) \geq \int_{\mathcal{D}_a} g(x)\, dP_X(x) \geq \int_{\mathcal{D}_a} a\, dP_X(x) = aP(X \in \mathcal{D}_a).$$

Se cierra la prueba notando de que $(X \in \mathcal{D}_a) = (g(X) \geq a)$. □

Existen varias formas similares, que son de hecho casos particulares de estas desigualdades.

Corolario 1-3 (Bienaymé–Chebyshev)**.** *Sea X un vector aleatorio d-dimensional admitiendo una esperanza m_X y una covarianza Σ_X. Entonces,*

$$\forall\, \varepsilon > 0, \qquad P\left(\left\|\Sigma_X^{-\frac{1}{2}}(X - m_X)\right\| > \varepsilon\right) \leq \frac{d}{\varepsilon^2}.$$

Viene del teorema incial aplicado a $\Sigma_X^{-\frac{1}{2}}(X - m_X)$, $g(x) = \|x\|^2$ y $a = \varepsilon^2$.

Corolario 1-4 (Markov)**.** *Sea X un vector aleatorio y $\varphi \geq 0$ una función no decreciente tal que $\varphi(\|X\|)$ sea integrable. Entonces,*

$$\forall\, \varepsilon \geq 0, \quad \textit{tal que} \quad \varphi(\varepsilon) \neq 0, \qquad P\left(\|X\| > \varepsilon\right) \leq \frac{\mathrm{E}\left[\varphi\left(\|X\|\right)\right]}{\varphi(\varepsilon)}.$$

La versión inicial des esta desigualdad trataba de funciones $\varphi(u) = u^r$, $r > 0$. Viene del teorema incial aplicado a $g(x) = \varphi\left(\|x\|\right)$ y $a = \varphi(\varepsilon)$, notando de que $\left(\varphi\left(\|X\|\right) \geq \varphi(\varepsilon)\right) = \left(\|X\| \geq \varepsilon\right)$ por la no decrecencia de φ. El caso anterior (una vez la variable centrada) es nada más que un caso especial.

Estas relaciones afirman que cuanto más chica es la varianza, más se concentra la variable en torno a su media. Ambas cotas son en general débiles, como se lo puede ver en el ejemplo siguiente

Ejemplo 1-18. *La desigualdad de Bienaymé–Chebyshev indica que la probabilidad de encontrar una fluctuación superior a $\varepsilon = 3\sigma_X$, tres desviaciones estándar alrededor de la media, está por debajo de $1/9$; el cálculo para una distribución típica como la Gaussiana, $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x-m_X)^2}{2}\right)$ ajusta dicha probabilidad por debajo de $0{,}003$.*

Una desigualdad muy importante que usaremos frecuentemente en el cápitulo siguiente, trata de funciones convexa, y del efecto sobre la media de una vector aleatorio.

Definición 1-36 (Función convexa)**.** *Por definición, una función $\phi : \mathcal{X} \subset \mathbb{R}^d \mapsto \mathbb{R}$ con $\mathcal{X}$ un convexo, es convexa si para cualquier $\pi_1 \in [0\,;\,1], \pi_2 = 1 - \pi_1$ y $x_1, x_2 \in \mathbb{R}^d$,*

$$\phi(\pi_1 x_1 + \pi_2 x_2) \leq \pi_1 \phi(x_1) + \pi_2 \phi(x_2).$$

ϕ es dicha estrictamente convexa si la desigualdad es estricta, salvo si $x_2 = x_1$.

Se puede ver de que si ϕ es dos veces diferenciable, su matriz Hessiana es simetrica no negativa, $\mathcal{H}\phi \in P_d(\mathbb{R})$.

Se muestra por recurrencia au, para cualquier conjunto $\{x_i\}_i$ numerable de elementos de $\mathcal{X}$ y reales positivos $\{\pi_i\}_i$ tales que $\sum_i \pi_i = 1$,

$$\phi\left(\sum_i \pi_i x_i\right) \leq \sum_i \pi_i \phi(x_i).$$

Dicho con palabras, la funcción del barycentro (combinación convexa) de los x_i es debajo del barycentro de los $\phi(x_i)$. Eso es ilustrado el la figura Fig. 1-12.

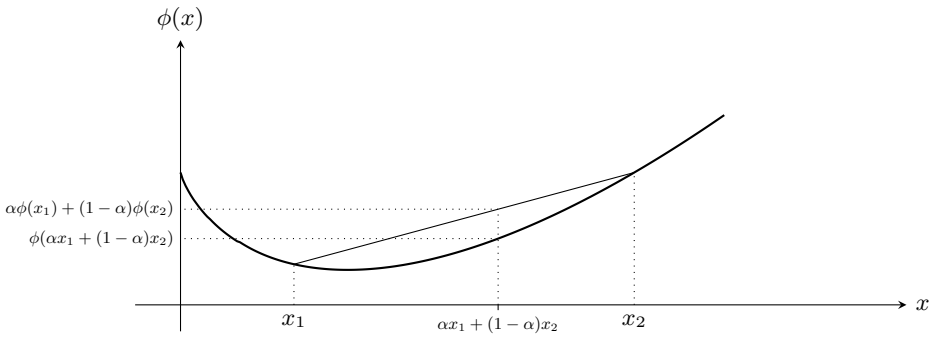


Figura 1-12: Ejemplo de función ϕ convexa: la cuerda, conteniendo los barycentros de $\{\phi(x_1)\,;\,\phi(x_2)\}$, es siempre arriba de la curba, *i. e.,* función de los barycentros de $\{x_1\,;\,x_2\}$.

Intuitivamente, la media teniendo un sabor de barycentro, se intuye de que la media de $\phi(X)$ va a ser arriva de la función de la media de X. Es precisamente el teorema de Jensen [18] (Jensen, 1906; Feller, 1971; Brémaud, 1988; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Cohn, 2013):

Teorema 1-21 (Desigualdad de Jensen)**.** *Sea X integrable y definida sobre $\mathcal{X} \subset \mathbb{R}^d$, convexo y $f : \mathcal{X} \mapsto \mathbb{R}$. Entonces*

$$\mathrm{E}[\phi(X)] \geq \phi\left(\mathrm{E}[X]\right).$$

Si ϕ es estrictamente convexa, la igualdad se alcanza si y solamente si X es determinista casi siempre.

Demostración. Sea $X = \sum_i x_i \mathbb{1}_{A_i}$ variable escalonada. Entonces $\phi(X) = \sum_i \phi(x_i)\mathbb{1}_{A_i}$, dando

$$\mathrm{E}[\phi(X)] = \sum_i P(A_i)\phi(x_i) \geq \phi\left(\sum_i P(A_i)x_i\right) = \phi\left(\mathrm{E}[X]\right)$$

con igualdad (cuando la convexidad es estricta) si y solamente si todos los x_i son iguales. Se cierra la prueba tomando $X \geq 0$ como limite de sucesión de funciones escalonadas (teorema 1-3, pagina 27), y cualquier X tratando de la parte positiva y negativa (ver pagina 27). El caso vectorial se trata componente a componente para X en termino de limite. Tomando el limite, la condición x_i todos iguales vuelve "casi todos" los x_i deben ser iguales, *i. e.,* X debe ser constante casi siempre. □

Terminamos esta sección con una desigualdad también muy útil, y conocida en los espacios de Hilbert, conocida como *desigualdad de Hölder* (Hölder, 1889; Shohat, 1929):

[18] En (Jensen, 1906) se trata del en el caso discreto y integral; en (Hölder, 1889; Hadamard, 1893) se encuentran las primeras semillas de esta desigualdad, y entre otros (Jessen, 1931a, 1931b; Perlman, 1974; Rudin, 1991) para versiones más generales.

Teorema 1-22 (Desigualdad de Hölder). *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales y* $r > 1$ *real.* $r^* > 1$ *tal que* $\frac{1}{r} + \frac{1}{r^*}$ *es llamado conjugado de Hölder de* r*, y*

$$\left|\mathrm{E}\left[X^t Y\right]\right| \leq \mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}} \mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}}.$$

Se obtiene la igualda si y solamente si existe un λ *tal que* $X = \lambda Y$ *casi siempre.*

Demostración. Obviamente, $|\mathrm{E}[X^t Y]| \leq \mathrm{E}[|X^t Y|]$. Luego, de la convexidad de la función $-\log$ se obtiene la desigualdad $\log(|ab|) = \frac{1}{r}\log|a|^r + \frac{1}{r^*}\log|b|^{r^*} \leq \log\left(\frac{|a|^r}{r} + \frac{|b|^{r^*}}{r^*}\right)$ con igualdad si y solamente si a es proporcional a b. Aplicado a las componentes de dos vectores a y b se obtiene la desigualdad de Young $|a^t b| \leq \frac{\|a\|_r^r}{r} + \frac{\|b\|_{r^*}^{r^*}}{r^*}$ con igualdad si y solamente si los vectores son proporcional. A continuación, denotando

$$\widetilde{X} = \frac{X}{\mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}}} \qquad \text{y} \qquad \widetilde{Y} = \frac{Y}{\mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}}}$$

tenemos

$$\mathrm{E}\left[\left|X^t Y\right|\right] = \mathrm{E}\left[\|X\|_r^r\right]^{\frac{1}{r}} \mathrm{E}\left[\|Y\|_{r^*}^{r^*}\right]^{\frac{1}{r^*}} \mathrm{E}\left[\left|\widetilde{X}^t \widetilde{Y}\right|\right].$$

De la desigualdad de Young, se obtiene entonces

$$\mathrm{E}\left[\left|\widetilde{X}^t \widetilde{Y}\right|\right] \leq \frac{\mathrm{E}\left[\left\|\widetilde{X}\right\|_r^r\right]}{r} + \frac{\mathrm{E}\left[\left\|\widetilde{Y}\right\|_{r^*}^{r^*}\right]}{r^*} = \frac{1}{r} + \frac{1}{r^*} = 1,$$

lo que cierra la prueba. □

Un corolario es conocido como desigualdad de Cauchy-Bunyakovsky-Schwarz [19] para $p = \frac{1}{2}$:

Corolario 1-5 (Desiguladad de Cauchy-Bunyakovsky-Schwarz). *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales. Entonces*

$$\left|\mathrm{E}\left[X^t Y\right]\right|^2 \leq \mathrm{E}\left[\|X\|^2\right] \mathrm{E}\left[\|Y\|^2\right].$$

Se obtiene la igualda si y solamente si existe un λ *tal que* $X = \lambda Y$ *casi siempre.*

Nota: se puede probar esta desigualdad considerando el polinomio $\mathrm{E}\left[\|\lambda X + Y\|^2\right] \geq 0$, del secundo orden en λ. Siendo no negativa para cualquier λ el discriminente debe ser no positivo, conduciendo a la desigualdad.

De hecho, se puede ver $\mathrm{E}[X^t Y]$ como un producto escalar entre variables aleatorias. La sola subtileza es que $\mathrm{E}\left[\|X\|^2\right] = 0$ conduce a $X = 0$ casi siempre, *i. e.,* se puede tener $X \neq 0$ pero con medida de probabilidad igual a cero (ej. puntos ω "aislados" en el contexto continuo).

Un otro corolario de la desiguldad de Hölder concierna el comportamiento de $s \mapsto \mathrm{E}\left[\|X\|_s^s\right]^{\frac{1}{s}}$ dado X:

---

[19] Esta desigualdad, fue probada por Cauchy para sumas en 1821 (Cauchy, 1821), para integrales por Bunyakovsky en 1859 (Bouniakowsky, 1859) y más elegamente por Schwarz en 1888 (Schwarz, 1888) en un enfoque más general. Ver también (Steele, 2004).

Corolario 1-6 (Crecencia de $s \mapsto \mathrm{E}\left[\,\|X\|_s^s\,\right]^{\frac{1}{s}}$). *Sean X vectores aleatorios d-dimensionales. Entonces*

$$s \mapsto \mathrm{E}\left[\,\|X\|_s^s\,\right]^{\frac{1}{s}} \quad \textit{es creciente}$$

Demostración. Aplicando la desigualdad de Hölder a $\|X\|_s^s$ y 1 se obtiene

$$\forall\, r > 1, \quad \mathrm{E}\left[\|X\|_s^s\right] \leq \mathrm{E}\left[\|X\|_{rs}^{rs}\right]^{\frac{1}{r}}$$

Se cierra la prueba elevando la desigualdad a la potencia $\frac{1}{s}$ y notando que $t = rs > s$. □

Varias otras desigualdades se encuentran en la literatura (ver por ejemplo en (Shohat, 1929, y notas debidas a Pearson) para unas de las más antiguas), así que no se puede ser exaustivo. Presentamos en esta sección las principales.

1.7 Esperanza condicional

Vimos en la sección 1.5 que una pregunta natural era de, dados dos vectores aleatorios X e Y, caracterizar el vector Y si "observamos X". Más adelante, nos podemos interesar a la media de Y cuando observamos X. Una manera intuitiva es de definir tal media cuando "sabemos" que $X = x$ a partir de la ley condicional $P_{Y|X=x}$ (Feller, 1968, 1971; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Spiegel, 1976; Kolmogorov, 1956; Jacob & Protters, 2003; Billingsley, 2012):

Definición 1-37 (Esperanza condicional). *Sean X e Y dos vectores aleatorios respectivamente d_X y d_Y-dimensionales, y sea la función*

$$f(x) \equiv \mathrm{E}[Y|X = x] = \int_{\mathbb{R}^{d_Y}} y\, dP_{Y|X=x}(y)$$

Se define la esperanza condicional de Y condicionalment a X como siendo la variable aleatoria

$$\mathrm{E}[Y|X] = f(X)$$

Claramente, $\mathrm{E}[Y|X = x]$ siendo una esperanza vínculado al espacio de probabilidad $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_{Y|X=x}\big)$, heride de las propiedades de la media. Entre otros, es lineal, lo que da

$$\mathrm{E}\left[\sum_i a_i Y_i | X\right] = \sum_i a_i\, \mathrm{E}[Y_i|X]$$

satisface la desiguadad de Jensen, para ϕ convexa

$$\mathrm{E}\left[\phi(Y)|X\right] \geq \phi\big(\,\mathrm{E}[Y|X]\,\big)$$

entre otros.

Como en el caso de medida de probabilidad, cuando dos variables son independientes, condicionar no cambia la esperanza:

Lema 1-7 (Esperanza condicional e independencia)**.** *Cuando X e Y son independientes, la esperanza condicional coincide con la de Y,*

$$X \text{ e } Y \text{ independientes} \quad \Rightarrow \quad \mathrm{E}[Y|X] = \mathrm{E}[Y]$$

Demostración. El resultado viene inmediatamente de $P_{Y|X=x} = P_Y$ (ver lema 1-5). □

La media condicional se revela muy útil y poderoso para evaluar esperanzas de variables aleatorias por ejemplo gracia a la formula de la esperanza total, equivalente de las formulas de probabilidad total lema 1-1, lema 1-14 y lema 1-16.

Teorema 1-23 (Media total)**.** *La media (total) del vector aleatorio Y concide con la media de la esperanza condicional, i. e.,*

$$\mathrm{E}[Y] = \mathrm{E}\left[\,\mathrm{E}\,[Y|X]\,\right]$$

Más generalmente, para cualquier función medible f,

$$\mathrm{E}[f(Y)] = \mathrm{E}\left[\,\mathrm{E}\,[f(Y)|X]\,\right]$$

Demostración. De la fórmula de probabilidad total tenemos para cualquier $B \in \mathcal{B}(\mathbb{R}^{d_Y})$,

$$\begin{aligned}\int_B dP_Y(y) &= P(Y \in B)\\ &= \int_{\mathbb{R}^{d_X}} P_{Y|X=x}(B)\, dP_X(x)\\ &= \int_{\mathbb{R}^{d_X}} \left(\int_B dP_{Y|X=x}(y)\right) dP_X(x)\end{aligned}$$

es decir

$$\int_{\mathbb{R}^{d_Y}} \mathbb{1}_B(y)\, dP_Y(y) = \int_{\mathbb{R}^{d_X}} \left(\int_{\mathbb{R}^{d_Y}} \mathbb{1}_B(y) dP_{Y|X=x}(y)\right) dP_X(x)$$

i. e.,

$$\mathrm{E}[\mathbb{1}_B(Y)] = \mathrm{E}\left[\,\mathrm{E}[\mathbb{1}_B(Y)|X]\right]$$

Ahora, se usa la linealidad para cualquier función escalonada f, y luego por el teorema de convergencia monótona 1-4, para cualquier función medible,

$$\mathrm{E}[f(Y)] = \mathrm{E}\left[\,\mathrm{E}[f(Y)|X]\right].$$

□

Un otro resultado importante, permitiendo frecuentemente simplificar la evaluacíon de momentos a partir de esperanza condicional es el siguiente:

Teorema 1-24. *Para cualquier funciónes medibles f, g, tenemos*

$$\mathrm{E}\,[\,f(X)\, g(Y) \mid X] = f(X)\ \mathrm{E}\,[g(Y)|X]$$

Más generalmente, para h medible

$$\mathrm{E}\left[h(X,Y) \mid X = x\right] = \mathrm{E}\left[h(x,Y)|X = x\right]$$

lo que se simplifica si además X e Y son independientes:

$$X \text{ e } Y \text{ independientes} \quad \Rightarrow \quad \mathrm{E}\left[h(X,Y) \mid X = x\right] = \mathrm{E}\left[h(x,Y)\right]$$

Demostración. De la definición de la medida de probabilidad condicional tenemos para cualquier $A \in \mathcal{B}(\mathbb{R}^{d_X}), \quad B \in \mathcal{B}(\mathbb{R}^{d_Y}), \quad C \in \mathcal{B}(\mathbb{R}^{d_X})$,

$$P\big((X \in A) \cap (Y \in B) \cap (X \in C)\big) = \int_C P_{X,Y|X=x}(A \times B)\, dP_X(x)$$

pero, también

$$P\big((X \in A) \cap (Y \in B) \cap (X \in C)\big) = \int_{A \cap C} P_{Y|X=x}(B)\, dP_X(x)$$

Entonces,

$$P_{X,Y|X=x}(A,B) = \mathbb{1}_A(x)\, P_{Y|X=x}(B)$$

A continuación,

$$\int_{\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}} \mathbb{1}_A(u)\, \mathbb{1}_B(v)\, dP_{X,Y|X=x}(u,v) = \mathbb{1}_A(x) \int_{\mathbb{R}^{d_Y}} \mathbb{1}_B(v)\, dP_{Y|X=x}(v)$$

Entonces, por linealidad, aplicando este resultado a funciones escalonadas, y luego por el teorema de convergencia monótona, para cualquieras f, g medibles,

$$\int_{\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}} f(u)\, g(v)\, dP_{X,Y|X=x}(u,v) = f(x) \int_{\mathbb{R}^{d_Y}} g(v)\, dP_{Y|X=x}(v)$$

es decir, por definición de la esperanza condicional,

$$\mathrm{E}[f(X)\, g(Y)|X = x] = f(x)\ \mathrm{E}[g(Y)|X = x]$$

lo que cierra la prueba de la primera identidad. Las otras se prueban con exactamente los mismos pasos. □

Un resultado que sirve a veces como definición, en el contexto de variable de cuadrado integrable, se víncula con la idea de aproximar una variable por una funcción de una otra:

Teorema 1-25. *Sea Y de cuadrado integrable, la esperanza condicional $E[Y|X]$ es la única variable $Z = f(X)$, función de X de cuadrado integrable, minimizando el error promedio cuadratico $\mathrm{E}[\|Y - Z\|^2]$. Dicho de otra manera, con el criterio de error cuadratico promedio mínimo, $E[Y|X]$ es la "mejor" función de X (en el sentido de la distancia inducida por el producto escalar) aproximando Y.*

Demostración. Usando la fórmula de esperanza total, y el teorema 1-24, se escribe

$$\begin{aligned} \mathrm{E}\left[\|Y - f(X)\|^2\right] &= \mathrm{E}\left[\mathrm{E}\left[\, \|Y - f(X)\|^2 \,\middle|\, X\right]\right] \\ &= \mathrm{E}\left[f(X)^2 - 2f(X)\,\mathrm{E}[Y|X] + \mathrm{E}\left[Y^2|X\right]\right] \end{aligned}$$

Ahora, buscando $\lambda \equiv f(x)$ minimizando $\|\lambda\|^2 - 2\lambda^t \operatorname{E}[Y|X=x] + \operatorname{E}\left[\|Y\|^2|X=x\right]$ para cualquier $x \in \mathcal{X}$, se minimizará el promedio en X. Inmediatamente, notando que buscamos el mínimo de un paraboliode de concavivad por arriba, anulando el gradiente en λ so obtiene $\lambda \equiv f(x) = \operatorname{E}[Y|X=x]$, el único mínimo, lo que cierra la prueba. □

Este resultado es muy conocido en el mundo de la estimación donde se quiere aproximar una variable minimizar el error cuadratico promedio (Kay, 1993; Robert, 2007).

Corolario 1-7. *Sea Y de cuadrado integrable, La esperanza condicional $E[Y|X]$ es la única variable $Z = f(X)$ de cuadrado integrable tal que para cualquier función medible g tal que $g(X)$ es de cuadrado integrable, $\operatorname{E}[g(X)^t Y] = \operatorname{E}[g(X)^t Z]$.*

Demostración. Como lo hemos visto en la sección anterior, $(U,V) \mapsto \operatorname{E}[U^t V]$ define un producto escalar. Según el teorema de proyección ortogonal (ver figura 1-13 y (Jacob & Protters, 2003; Athreya & Lahiri, 2006; ?, ?)), el $f(X)$ único, que minimiza $\operatorname{E}\left[\|Y - f(X)\|^2\right]$, es la proyección ortogonal de Y sobre el espacio $\sigma^2(X) = \{g(X) \mid g \text{ es medible con } \wedge \ g(X) \text{ es de cuadrado integrable}\}$. En otros terminos, la desviación $Y - \operatorname{E}[Y|X]$ entre Y y $\operatorname{E}[Y|X]$ es ortogonal a cualquier $g(X) \in \sigma^2(X)$, y reciprocamente si $Y - f(X)$ es ortogonal a cualquier $g(X)$, $f(X)$ es necesariamente la proyección ortogonal de Y sobre $\sigma^2(X)$, lo que cierra la prueba. □

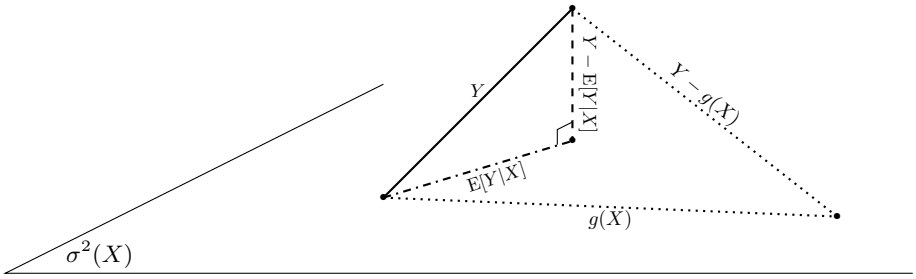


Figura 1-13: Ilustración del teorema de proyección ortogonal. El espacio $\sigma^2(X) = \{g(X) \mid g \text{ es medible con } \wedge \ g(X) \text{ es de cuadrado integrable}\}$ es representado por el plano y el vector representa Y. La linea punteada representa la desviación $Y - g(X)$ entre Y y $g(X)$ dado $g(X) \in \sigma^2(X)$, siendo su cuadrado promedio es mínimo cuando (linea con guiones) $g(X) = \operatorname{E}[Y|X]$ corresponde a la proyección ortogonal de Y (linea mixta).

A veces, este resultado sirve también a veces, como definición de la esperanza condicional.

1.8 Funciones generadoras

Como lo hemos visto, un vector aleatorio es completamente definida por su medida de probabilidad P, o equivalemente por la medida imagen P_X, o a través de la función de repartición F_X. Sin embargo,

bajo el impulso de Laplace en el siglo XVII (entre otros), se introdujo caracterizaciones alternativas a través de transformaciones de la medida de probabilidad, conocidas como *funciones generadoras* o *funciones generatrices* [20] (de Laplace, 1820). Existen varias funciones, cuyas tienes propiedades particulares que vamos a ver en las subsecciones siguientes. Entre otros, estas funciones dadas como valores de expectación de funciones de la variable aleatoria (discreta o continua), con un parámetro real o complejo, permiten hallar fácilmente los distintos momentos de una distribución de probabilidad.

1.8.1 Función generadora de probabilidad

De manera general, siguiendo el enfoque de A. de Moivre (ver nota de pie 20) dada una sucesión $a_n, \quad n \in \mathbb{N}$, se define la función generadora dicha *ordinaria* de la sucesión como $G(\{a_n\}_{n\in\mathbb{N}}, z) = \sum_{n\in\mathbb{N}} z^n$. A veces, esta serie es conocida como transformada en z de la sucesión $\{a_n\}_{n\in\mathbb{N}}$. Tratando de variables aleatorias discretas sobre $\mathbb{N}$, con $p_n = P_X(n) = P(X = n)$, se puede definir así la función generadora asociada a la sucesión p_n y se puede ver que no es nada más que el momento $\mathrm{E}\left[z^X\right]$. De manera general, la función generadora de probabilidad se define de la manera siguiente (Feller, 1968; Johnson, Kotz & Balakrishnan, 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

Definición 1-38 (función generadora de probabilidad o de momentos factoriales)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ *vector aleatorio* d *dimensional definido sobre* $\mathcal{X} \subset \mathbb{R}^d$*. La función definida por*

$$G_X(z) = \mathrm{E}\left[\prod_{i=1}^{d} z_i^{X_i}\right] \quad \textit{con} \quad z = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix}^t \in \mathbb{C}^d$$

es conocida como función generadora de probabilidad *o* función generadora de momentos factoriales *de* X*.*

Esta función est definida sobre un producto cartesiano de anillos [21] en el plano complejos, $r_i \leq |z_i| \leq R_i$ con $r_i \leq 1$ y $R_i \geq 1$.

---

[20] De hecho, de manera general, se introdujeron tales funciones en un marco más general, asociado a sucesiones de números, bajo el impulso de A. de Moivre (de Moivre, 1730); ver también (Stirling, 1730; Euler, 1741, 1750; de Moivre, 1756) o (Knuth, 1997, Sec. 1.2.9).

[21] Si para $|z_i| = R_i$ la integral converge uniformamente, para cualquier $z \in \mathbb{C}^d$, $r_i \leq |z_i| \leq R_i$ tenemos por ejemplo $\left|\int_{\mathbb{R}_+^d} \prod_{i=1}^{d} z_i^{x_i}\, dP_X(x)\right| \leq \int_{\mathbb{R}_+^d} \left|\prod_{i=1}^{d} z_i^{x_i}\right| dP_X(x) \leq \int_{\mathbb{R}_+^d} \left|\prod_{i=1}^{d} R_i^{x_i}\right| dP_X(x) < +\infty$. El mismo enfoque se usa para todos los hipercuadrantes $\otimes_i \mathbb{R}_{\pm}$. Además, claramente, para $|z_i| = 1$ tenemos $\left|\int_{\mathbb{R}^d} \prod_{i=1}^{d} z_i^{x_i}\, dP_X(x)\right| \leq \int_{\mathbb{R}^d} \left|\prod_{i=1}^{d} z_i^{x_i}\right| dP_X(x) \leq \int_{\mathbb{R}^d} dP_X(x) = 1$, lo que prueba que $r_i \leq 1 \leq R_i$.

La denominación *generadora de probabilidad* (pgf para *probability generating function* en ingles) se entiende sencillamente del hecho siguiente:

Lema 1-8. *Cuando* $\mathcal{X} = \mathbb{N}^d$ *para cualquier* $x = \begin{bmatrix} x_1 & \dots & x_d \end{bmatrix}^t \in \mathbb{N}^d$, *con* $k = \sum_{i=1}^d x_i$

$$\frac{1}{\prod_{i=1}^d x_i!} \left. \frac{\partial^k G_X}{\partial z_1^{x_1} \dots \partial z_d^{x_d}} \right|_{z=0} = P_X(x) = P(X = x)$$

Demostración. Se puede escribir la función G_X bajo su forma de generadora ordinaria $G_X(z) = \sum_{x \in \mathbb{N}^d} \left(\prod_{i=1}^d z_i^{x_i} \right) P(X = x)$ con $x = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}^t$. A continuación, se nota que la serie converge uniformamente por lo menos en la bola $\mathbb{B}_d \equiv \mathbb{B}_d(1)$, probando que G_X es diferenciable en $\mathbb{B}_d$, así que se puede ver esta series como el desarollo de Taylor de G_X (o, equivalentemente, diferenciar bajo la suma y tomar la derivada en $z = 0$), lo que cierra la prueba. □

De este resultado, se puede notar que, en el caso discreto, hay una relación uno-a-uno entre la medida de probabilidad P_X y la función generadora de probabilidad G_X. En el caso general, veremos en la subsección **??** que para z_j de la forma $z_j = e^{iu_j}$ con $u_j \in \mathbb{R}$ la transformación se inversa, de manera que se puede recuperar la medida de probabilidad P_X a partir de G_X. Dicho de otra manera, como la medida P_X, la función G_X caracteriza completamente el vector aleatorio X.

Aparece que la función generadora G_X se vincula también con los momentos factoriales, justificando su secunda denominación, *generadora de momentos factoriales* (fmgf para *factorial moments generating function* en ingles):

Lema 1-9. *Para cualquier* $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathbb{N}^d$ *con* $K = \sum_{i=1}^d k_i$, *derivando* G_X *se proba que, cuando existen* [22]

$$\left. \frac{\partial^K G_X}{\partial z_1^{k_1} \cdots \partial z_d^{k_d}} \right|_{z=1} = \mathrm{E}\left[\prod_{i=1}^d (X_i)_{k_i} \right]$$

momento factorial [23] *de* X.

De este resultado, se ve por ejemplo que, cuando existen, se recuperan los momentos de X a través de las derivadas de G_X:

- $G_X(1) = 1$, condición de normalización.
- $\nabla_z G_X(1) = \mathrm{E}[X]$.

[22] En el caso extremo, el rayo de convergencia de la serie dando G_X es igual a 1, así que no hay garantia que las derivadas en $z = 1$ existen.

[23] Recuerdense que $(x)_n = \prod_{i=0}^{n-1}(x - i), \quad n > 0$ símbolo de Pochhammer, con la convención $(x)_0 = 1$; ver pagina **??**

- $\mathcal{H}_z G_X(1) + \operatorname{diag}(\nabla_z G_X(1)) = \mathrm{E}[XX^t]$ donde $\mathcal{H}_z$ es la matrice Hessiana y $\operatorname{diag}(a)$ es une matriz diagonal de componentes (i,i)-esima a_i (vector a sobre la diagonal). Entonces la matriz de covarianza es dada por $\operatorname{Cov}[X] = \mathcal{H}_z G_X(1) + \operatorname{diag}(\nabla_z G_X(1)) - \nabla_z G_X(1) \nabla_z^t G_X(1)$.

La función G_X tiene unas propiedades permitiendo por ejemplo de manejar sencillamente distribuciones de probabilidades de combinaciones lineales de vectores aleatorios independientes, como lo vamos a ver a través del teorema siguiente.

Teorema 1-26. *Sean X e Y dos vectores aleatorios d-dimensionales independientes, $a = \begin{bmatrix} a_1 & \cdots & a_d \end{bmatrix}^t \in \mathbb{R}^d$ y $b = \begin{bmatrix} b_1 & \cdots & b_d \end{bmatrix}^t \in \mathbb{R}^d$. Entonces para cualquier $z = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix} \in \mathbb{C}^d$ (donde existen las funciones):*

$$G_{\operatorname{diag}(a)X+b}(z) = \prod_{i=1}^{d} z_i^{b_i} G_X\left(z_1^{a_1}, \ldots, z_d^{a_d}\right),$$

$$G_{X+Y}(z) = G_X(z)\, G_Y(z)$$

y para $z \in \mathbb{C}$

$$G_X\left(z^{a_1}, \ldots, z^{a_d}\right) = G_{a^t X}(z)$$

Demostración. El primer resultado es inmediato, escribiendo $z_i^{a_i X_i + b_i} = z_i^{b_i} \left(z_i^{a_i}\right)^{X_i}$. El secundo viene de $z_i^{X_i+Y_i} = z_i^{X_i} z^{Y_i}$ conjuntamente con el teorema 1-19 con $f(X) = \prod_{i=1}^d z_i^{X_i}$ y $g(Y) = \prod_{i=1}^d z_i^{Y_i}$. El tercer resultado es consecuencia de $\prod_{i=1}^d \left(z^{a_i}\right)^{X_i} = z^{\sum_{i=1}^d a_i X_i}$. □

Estos resultados permiten manejar sencillamente la medida de probabilidad de combinaciones lineales de vectores aleatorios independientes y de marginales a través esta función generadora.

De la tercera identidad, se puede hacer un paso más tratando de sumas aleatorias de vectores aleatorios:

Teorema 1-27. *Sea $X_n, \quad n \in \mathbb{N}$ una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de probabilidad) P_X (resp. G_X) y N una variable definida sobre $\mathbb{N}$, independiente de los X_n. Sea el vector aleatorio $S_N = \sum_{n=0}^N X_n$. Entonces*

$$G_{S_N}(z) = G_N\big(G_X(z)\big),$$

Demostración. Usando la formula de esperanza total del teorema. 1-23, se escribe

$$\begin{aligned} G_{S_N}(z) &= \mathrm{E}\left[z^{\sum_{n=0}^N X_n}\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[z^{\sum_{n=0}^N X_n} \middle| N\right]\right] \\ &= \mathrm{E}\left[G_X(z)^N\right] \end{aligned}$$

□

1.8.2 Función generadora de momentos

Como lo hemos visto, la función generadora de probabilidad permite recucperar los momentos de un vector aleatorio a través de combinaciones de sus derivadas. Con una pequaña modificación, se puede definir una función generada permitiendo recuperar más directamente los momentos, de manera siguiente (Feller, 1968; Johnson et al., 1997; Mukhopadhyay, 2000; Athreya & Lahiri, 2006):

Definición 1-39 (función generadora de momentos)**.** *La* función generadora de momentos *(mgf para* moment generating function *en ingles) de un vector aleatorio* d*-dimensional se define como*

$$M_X(u) = \mathrm{E}\left[e^{z^t X}\right]$$

para $u \in \mathbb{C}^d$*.*

De esta definición se nota inmediatamente que

$$M_X(u) = G_X\left(e^u\right) \qquad \text{donde} \qquad e^u = \begin{bmatrix} e^{u_1} & \cdots & e^{u_d} \end{bmatrix}^t$$

Entonces, como G_X, la generadora de los momentos caracteriza completamente el vector aleatorio X. Además, de este vínculo entre G_X y M_X, y del dominio de definición de G_X, queda claro que M_X es definida sobre un producto cartesiano de franjas del plano complejo, $v_i \leq \Re e\{u_i\} \leq V_i$ donde $v_i \leq 0 \leq V_i$ llamamos *indices de convergencia*. En el caso de variables escalares admitiendo una densidad de probabildad p_X, denotando $s = -u$, esta función se interpreta como la transformada (bilateral) de Laplace de p_X.

La generadora de los momentos permite recuperar directamente los momentos a través de derivadas, sin hacer combinaciones:

Lema 1-10. *Para cualquier* $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathbb{N}^d$ *con* $K = \sum_{i=1}^d k_i$*, derivando* M_X *se proba que, cuando existen*

$$\left.\frac{\partial^K M_X}{\partial u_1^{k_1} \cdots \partial u_d^{k_d}}\right|_{u=0} = \mathrm{E}\left[\prod_{i=1}^d X_i^{k_i}\right] = m_{k_1,\ldots,k_d}$$

momento de orden k *de* X*.*

En particular, se recuperan

- $M_X(0) = 1$, condición de normalización.
- $\nabla_u M_X(0) = \mathrm{E}[X]$ promedio,
- $\mathcal{H}_u M_X(0) = \mathrm{E}\left[XX^t\right]$, *i. e.,* $\mathrm{Cov}[X] = \mathcal{H}_u M_X(0) - \nabla_u M_X(0) \nabla_u^t M_X(0)$ matriz de covarianza.

Como la función G_X, la generadora de los momentos tiene unas propiedades similares a las de los teoremas 1-26 y 1-27:

Teorema 1-28. *Sean* X *e* Y *dos vectores aleatorios* d*-dimensionales independientes,* A *una matriz de* $\mathbb{R}^{d'\times d}$ *y* $b = \begin{bmatrix} b_1 & \cdots & b_{d'} \end{bmatrix}^t \in \mathbb{R}^{d'}$*. Entonces para cualquier* $u = \begin{bmatrix} u_1 & \cdots & u_{d'} \end{bmatrix}^t \in \mathbb{C}^{d'}$ *(donde la función existe):*

$$M_{AX+b}(u) = e^{u^t b} M_X\left(A^t u\right),$$

y para cualquier $u = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}^t \in \mathbb{C}^d$ *(donde la función existe):*

$$M_{X+Y}(u) = M_X(u)\, M_Y(u)$$

Además, para $X_n, \quad n \in \mathbb{N}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de momentos)* P_X *(resp.* M_X*)* y N *una variable aleatoria definida sobre* $\mathbb{N}$*, independiente de los* X_n*, y* $S_N = \sum_{n=0}^{N} X_n$*,*

$$M_{S_N}(u) = G_N\big(M_X(u)\big),$$

Demostración. Las pruebas siguen punto a punto los mismos pasos que las de los teoremas 1-26 y 1-27. □

De nuevo, se puede hacer un paso más tratando de sumas aleatorias de vectores aleatorios como en el teorema 1-27:

Teorema 1-29. *Sea* $X_n, \quad n \in \mathbb{N}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de probabilidad)* P_X *(resp.* M_X*)* e N *una variable aleatoria definida sobre* $\mathbb{N}$*, independiente de los* X_n*. Sea el vector aleatorio* $S_N = \sum_{n=0}^{N} X_n$*. Entonces*

$$M_{S_N}(u) = G_N\big(M_X(u)\big),$$

Demostración. El resultado es consecuencia directa del teorema 1-27. □

1.8.3 Función característica

Si la función generadora de momentos permite recuperar los momentos de un vector aleatorio, no es definida sobre todo $\mathbb{C}^d$. Sin embargo, cuando $\Re e\{u_i\} = 0$, esta función es siempre definida. Entonces, una función generadora muy útil que se usa frecuentemente es la de momentos para este tipo de argumentos, lo que es conocida como función característica y que es al final definida sobre $\mathbb{R}^d$ de manera siguiente (Lukacs, 1961; Golberg, 1961; Feller, 1968; Stein & Weiss, 1971; Johnson et al., 1997; Mukhopadhyay, 2000; Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Sasvári, 2013):

Definición 1-40 (función característica)**.** *La* función característica *(cf para* characteristic function *en ingles) de un vector aleatorio* d*-dimensional se define como*

$$\Phi_X(\omega) = \mathrm{E}\left[e^{\imath\omega^t X}\right]$$

para $\omega \in \mathbb{R}^d$.

De esta definición se nota inmediatamente que

$$\Phi_X(\omega) = M_X(\imath\omega) = G_X\left(e^{\imath\omega}\right) \qquad \text{donde} \qquad e^{\imath\omega} = \begin{bmatrix} e^{\imath u_1} & \cdots & e^{\imath u_d} \end{bmatrix}^t$$

De hecho, se puede definir esta función para un argumento complejo, pero es equivalente a volver a la definición de la generadora de momentos.

En su forma general, la función característica se escribe

$$\Phi_X(\omega) = \int_{\mathbb{R}^d} e^{\imath\omega^t x}\, dP_X(x)$$

y es relacionada a la transformada de Fourier-Stieltjes de la medida P_X (Pinsky, 2009, Chap. 5). Cuando P_X admite una densidad p_X, la función es una transformada de Fourier usual de la densidad p_X, introducida bajo el impulso de Fourier en 1822 para estudiar la difusión del calor (Fourier, 1822).

Insistamos sobre el hecho que la importancia de esta función reside en que siempre existe y está bien definida, dado que $\int_{\mathbb{R}^d} \left|e^{\imath\omega^t x}\right| dP_X(x) = \int_{\mathbb{R}^d} dP_X(x) = 1$.

Como para las generadoras ya introducidas, la función característica permite recuperar directamente los momentos a través de derivadas:

Lema 1-11. *Para cualquier* $k = \begin{bmatrix} k_1 & \cdots & k_d \end{bmatrix}^t \in \mathbb{N}^d$ *con* $K = \sum_{i=1}^d k_i$*, derivando* Φ_X *se proba que, cuando existen*

$$(-\imath)^K \left.\frac{\partial^K \Phi_X}{\partial\omega_1^{k_1}\cdots\partial\omega_d^{k_d}}\right|_{\omega=0} = \mathrm{E}\left[\prod_{i=1}^d X_i^{k_i}\right] = m_{k_1,\ldots,k_d}$$

momento de orden k *de* X.

En particular, se recuperan

- $\Phi_X(0) = 1$, condición de normalización.
- $-\imath\nabla_\omega M_X(0) = \mathrm{E}[X]$ promedio,
- $-\mathcal{H}_\omega M_X(0) = \mathrm{E}\left[XX^t\right]$, *i. e.,* $\mathrm{Cov}[X] = -\mathcal{H}_\omega M_X(0) + \nabla_\omega M_X(0)\nabla_\omega^t M_X(0)$ matriz de covarianza.

Fijense de que Φ_X no es siempre diferencial en $\omega = 0$; Por ejemplo, en el caso de la distribución de Cauchy–Lorentz univariada [24] $p_X(x) = \frac{\gamma}{\pi(\gamma^2+(x-x_0)^2)}$ con $\gamma > 0$, resulta $\Phi_X(\omega) = e^{-\imath x_0\omega - \gamma|\omega}$. Esta función está definida para todo ω, pero no es derivable en $\omega = 0$, lo que coincide con el hecho de que no están definidos los momentos para esta densidad de probabilidad.

Resumimos algunas otras propiedades importantes de la función característica:

Teorema 1-30 (Propiedades principales de la función característica)**.**

[24] Lo mismo occure en la extensión multivariada (Samorodnitsky & Taqqu, 1994).

1. *Φ_X es una función medible y continua en $\mathbb{R}^d$ (Pinsky, 2009, Prop. 5.2.1). Eso es una consecuencia del teorema de convergencia dominada (ver teorema 1-5 pagina 28).*

2. *$\Phi_X(0) = 1$: Eso es inmediato escribiendo la integral, siendo P_X una medida de probabilidad.*

3. *$|\Phi_X(\omega)| \leq 1 = \Phi_X(0)$: $|\Phi_X(\omega)|$ es máxima en $\omega = 0$. Eso viene directamente de $\left|e^{\imath\omega^t x}\right| = 1$.*

4. *$\Phi_X(-\omega) = \Phi_X^*(\omega)$: Φ_X tiene una símetria hermitica.*

5. *Φ_X es una no negativa definida, i. e., para un conjunto arbitrario de $n \geq 1$ números complejos $a_1, \ldots, a_n$ y n vectores $w_1, \ldots, w_n$ de $\mathbb{R}^d$, se cumple*

$$\sum_{k,l=1}^{n} a_k^* a_l \Phi_X(w_l - w_k) \geq 0$$

Dicho de otra manera, la matriz de componente (k,l)-esima $\Phi_X(w_l - w_k)$ es a hermitica (símetria hermítica dada por la propiedad anterior, y no negativa definida). Esta positividad viene de $\sum_{k,l} a_k^* a_l e^{\imath(w_l - w_k)^t x} = \left|\sum_l a_l e^{\imath w_l^t x}\right|^2 \geq 0$.

De hecho, existe una reciproca de este teorema, debido a S. Bochner [25] (Bochner, 1932, 1959; Golberg, 1961; Pinsky, 2009; Sasvári, 2013)

Teorema 1-31 (Bochner)**.** *Una función $\Phi : \mathbb{R}^d \mapsto \mathbb{C}$ es continua, definida no negativa (con $\Phi(0) = 1$) si y solamente existe una medida μ (de probabilidad) sobre $\mathcal{B}(\mathbb{R}^d)$ tal que*

$$\forall\, \omega \in \mathbb{R}^d, \quad \Phi(\omega) = \int_{\mathbb{R}^d} e^{\imath\, \omega^t x} d\mu(x)$$

Dicho de otra manera, cualquier función continua, definida positiva con $\Phi(0) = 1$ es la función característica de un vector aleatorio.

En el teorema, vimos que la transformada de Fourier-Stieljes de una medida de probabilidad P_X es medible, continua, definida no negativa, con $\Phi(0) = 1$. La reciproca es más dificil a probar y necesita lemas adicionales. Se puede encontrar una linda prueba en (Sasvári, 2013, Sec. 1.7) adonde dejamor el lector.

Como lo hemos notado en las subsecciones anteriores, la función característica define completamente el vector aleatorio. En particular, hay una relación uno-uno (casi siempre) entre Φ_X y la medida P_X. En particular, existe una fórmula de inversión permitiendo volver a la medida P_X a partir de Φ_X (Ash & Doléans-Dade, 1999; Sasvári, 2013):

[25] De hecho, lo probó Bochner en el caso escalar $d = 1$, pero se extiende al caso multivariado.

Teorema 1-32 (Fórmula de inversión). *Sea X vector aleatorio d-dimensional de función característica Φ_X. Sea $A = \bigtimes_{i=1}^{d}(a_i\,;\,b_i) \in \mathcal{B}(\mathbb{R}^d)$ y $\partial A = \times_{i=1}^{d}[a_i\,;\,b_i] \setminus A$ su borde. Entonces* [26],

$$\lim_{w_1\to+\infty,\ldots,w_d\to\infty} \frac{1}{(2\pi)^d}\int_{\times_{j=1}^{d}[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^{d}\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega$$
$$=$$
$$\int_{\mathbb{R}^d}\prod_{j=1}^{d}\left(\mathbb{1}_{(a_j\,;\,b_j)}(x_j)+\frac{1}{2}\mathbb{1}_{\{a_j\,;\,b_j\}}(x_j)\right)dP_X(x)$$

En particular, cuando P_X vale 0 sobre el borde de A, es decir $P_X(\partial A)=0$, se obtiene

$$\lim_{w_1\to+\infty,\ldots,w_d\to\infty} \frac{1}{(2\pi)^d}\int_{\times_{j=1}^{d}[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^{d}\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega \;=\; P_X(A).$$

Nota: el límite $\lim_{T\to+\infty}\int_{-T}^{T}$ se nota a veces $\mathrm{vp}\int_{\mathbb{R}}$, integral *en valor principal*.

Demostración. Por definición de la función característica, tenemos

$$\int_{\times_{j=1}^{d}[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^{d}\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega = \int_{\times_{j=1}^{d}[-w_j\,;\,w_j]}\int_{\mathbb{R}^d} e^{\imath\omega^t x}\,dP_X(x)\prod_{j=1}^{d}\frac{e^{-\imath a_j\omega_j}-e^{-\imath b_j\omega_j}}{\imath\omega_j}\,d\omega$$

Ahora, notando que $\left|\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,e^{\imath\,\omega^t x}\right| \le b_j-a_j$ es uniformamente acotado, se puede evocar el teorema de Fubini 1-6 para intercambiar las integrales, así que, con $e^{\imath\,\omega^t x}=\prod_{j=1}^{d}e^{\imath\,\omega_j x_j}$, tenemos

$$\int_{\times_{j=1}^{d}[-w_j\,;\,w_j]} \Phi_X(\omega)\prod_{j=1}^{d}\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\,d\omega = \int_{\mathbb{R}^d}\left(\prod_{j=1}^{d}\int_{-w_j}^{w_j}\frac{e^{-\imath\,\omega_j(a_j-x_j)}-e^{-\imath\,\omega_j(b_j-x_j)}}{\imath\,\omega_j}\,d\omega_j\right)dP_X(x)$$

Se nota que

$$\begin{aligned}\int_{-w_j}^{w_j}\frac{e^{-\imath\,\omega_j(a_j-x_j)}-e^{-\imath\,\omega_j(b_j-x_j)}}{\imath\,\omega_j}\,d\omega_j &= -\int_{-w_j}^{w_j}\frac{e^{+\imath\,\omega_j(a_j-x_j)}-e^{+\imath\,\omega_j(b_j-x_j)}}{\imath\,\omega_j}\,d\omega_j\\ &= \int_{-w_j}^{w_j}\frac{\sin(\omega_j(b_j-x_j))-\sin(\omega_j(a_j-x_j))}{\omega_j}\,d\omega_j\end{aligned}$$

por cambio de variables $\omega_j\to-\omega_j$ en la primera linea, tomando entonce la media suma de los terminos derecho/izquierdo dando la secunda linea. Seguimos notando que

$$\int_{-w}^{w}\frac{\sin(\omega(c-x))}{\omega}\,d\omega = \operatorname{sign}(c-x)\int_{-w|c-x|}^{w|c-x|}\frac{\sin(\omega)}{\omega}\,d\omega$$

es decir, de $\lim_{T\to+\infty}\int_{-T}^{T}\frac{\sin\omega}{\omega}d\omega=\pi$ (Gradshteyn & Ryzhik, 2015, Ec. 3.721), se obtiene

$$\lim_{w\to+\infty}\int_{-w}^{w}\frac{\sin(\omega(c-x))}{\omega}\,d\omega = \pi\operatorname{sign}(c-x)$$

[26] Se prolonga la función $\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}$ en $\omega_j=0$ por su límite $\lim_{\omega_j\to0}\frac{e^{-\imath\, a_j\omega_j}-e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}=b_j-a_j$.

Se acaba la prueba de $\left|\frac{\sin(\omega_j(b_j-x_j)-\sin(\omega_j(a_j-x_j))}{\omega_j}\right| < 2$ conjuntamente al teorema de convergencia dominada 1-5 permitiendo permutar integral y límite, y de $\operatorname{sign}(b_j - x_i) - \operatorname{sign}(a_j - x_i) = 2\,\mathbb{1}_{(a_j\,;\,b_j)}(x_j) + \mathbb{1}_{\{a_j\,;\,b_j\}}(x_j)$. □

Dos teoremas de inversión en los casos particular continuo y discreto permiten respectivamente volver a la densidad de probabilidad o a la masa de probabilidad.

Teorema 1-33 (Inversión, caso continuo)**.** *Si Φ_X es integrable, entonces P_X admite una densidad tal que*

$$p_X(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \Phi(\omega)\, e^{-\imath\,\omega^t x}\, d\omega$$

Demostración. Varias pruebas existen (ej. (Sasvári, 2013, p. 21)). Una bastante directa puede ser de salir de la fórmula general del teorema 1-32, de fijar a y poner $b \equiv x \in \mathbb{R}^d$. Se toma la derivada $\frac{\partial^d}{\partial x_1 \cdots \partial x_d}$ de la integral $\frac{1}{(2\pi)^d}\int_{\times_{j=1}^d [-w_j\,;\,w_j]} \Phi_X(\omega) \prod_{j=1}^d \frac{e^{-\imath\, a_j\omega_j} - e^{-\imath\, b_j\omega_j}}{\imath\,\omega_j}\, d\omega$ y se evoca el teorema de convergencia dominada para intercambiar derivación e integración, y luego tomar el límite, para obtener el resultado. □

Teorema 1-34 (Inversión, caso discreto)**.** *Para cualquier $x \in \mathbb{R}^d$,*

$$\lim_{w_1\to\infty,\ldots,w_d\to\infty} \frac{1}{2^d w_1 \ldots w_d}\int_{\times_{j=1}^d [-w_j\,;\,w_j]} \Phi(\omega)\, e^{-\imath\,\omega^t x}\, d\omega = P_X(x)$$

Demostración. Por definición de la función característica, y aplicando el teorema de Fubini como en el caso general (mismo enfoque),

$$\begin{aligned}
\frac{1}{2^d w_1 \ldots w_d}\int_{\times_{j=1}^d [-w_j\,;\,w_j]} \Phi_X(\omega) e^{-\imath\omega^t x}\, d\omega &= \frac{1}{2^d w_1 \ldots w_d}\int_{\times_{j=1}^d [-w_j\,;\,w_j]}\int_{\mathbb{R}^d} e^{\imath\omega^t y}\, dP_X(y)\, e^{-\imath\omega^t x} d\omega \\
&= \int_{\mathbb{R}^d}\left(\prod_{j=1}^d \frac{1}{2w_j}\int_{-w_j}^{w_j} e^{\imath\,(y_j-x_j)w_j} dw_j\right) dP_X(y) \\
&= \int_{\mathbb{R}^d}\prod_{j=1}^d \frac{\sin(w_j(y_j-x_j))}{(y_j-x_j)w_j} dP_X(y)
\end{aligned}$$

con el límite $\lim_{y_j\to x_j} \frac{\sin(w_j(y_j-x_j))}{w_j(y_j-x_j)} = 1$. Además, con el mismo enfoque que en el cas general acotando el integrande, por el teorema de convergencia dominada, se puede intercambiar límite e integral, así que, por $\lim_{x_j\to\infty} \frac{\sin(w_j(y_j-x_j))}{w_j(y_j-x_j)} = \mathbb{1}_{x_j}(y_j)$, lo cierra la prueba □

Como las funciónes G_X y M_X, la función característica tiene entre otros propiedades similares a las a las de los teoremas 1-28 y 1-29:

Teorema 1-35. *Sean X e Y dos vectores aleatorios d-dimensionales independientes, A una matriz de $\mathbb{R}^{d'\times d}$ y $b = \begin{bmatrix} b_1 & \cdots & b_{d'}\end{bmatrix}^t \in \mathbb{R}^{d'}$. Entonces para cualquier $\omega = \begin{bmatrix}\omega_1 & \cdots & \omega_{d'}\end{bmatrix}^t \in \mathbb{R}^{d'}$:*

$$\Phi_{AX+b}(\omega) = e^{\imath\omega^t b}\Phi_X\left(A^t\omega\right),$$

y para cualquier $\omega = \begin{bmatrix}\omega_1 & \cdots & \omega_d\end{bmatrix}^t \in \mathbb{R}^d$*:*

$$\Phi_{X+Y}(\omega) = \Phi_X(\omega)\,\Phi_Y(\omega)$$

Además, para $X_n, \quad n \in \mathbb{N}$ *una sucesión de vectores aleatorios indepedientes de misma distribución (resp. generadora de momentos)* P_X *(resp.* M_X*)* e N *una variable aleatoria definida sobre* $\mathbb{N}$*, independiente de los* X_n*, y* $S_N = \sum_{n=0}^{N} X_n$*,*

$$\Phi_{S_N}(\omega) = G_N\big(\Phi_X(\omega)\big),$$

Un otro resultado interesante se víncula a la noción de mezcla de escala y toma la forma siguiente:

Teorema 1-36. *Sea* X *vector aleatorio de función característica* Φ_X *y* R *variable aleatoria independiente de* X *y de medida de probabilidad* P_R*. Entonces, la función característica de* RX *es dada por*

$$\Phi_{RX}(\omega) = \int_{\mathbb{R}} \Phi_X(r\omega)\, dP_R(r)$$

Demostración. Por definición de la función característica, la fórmula de esperanza total 1-23 y el teorema 1-24, se tiene

$$\begin{aligned}
\Phi_{RX}(\omega) &= \mathrm{E}\left[e^{\imath\,\omega^t RX}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[e^{\imath\,\omega^t RX}\,\middle|\, R\right]\right] \\
&= \int_{\mathbb{R}} \mathrm{E}\left[e^{\imath\,\omega^t rX}\,\middle|\, R = r\right] dP_R(r) \\
&= \int_{\mathbb{R}} \mathrm{E}\left[e^{\imath\,\omega^t rX}\right] dP_R(r) \\
&= \int_{\mathbb{R}} \Phi_X(r\omega)\, dP_R(r)
\end{aligned}$$

□

Cumulant generating function ....

Se define los cumulantes. coinciden con los momentos centrados solo para $k = 1, 2, 3$**.**

teorema de Polya?

hablar de la cota de Chernoff con la mgf o pgf?

1.9 Vectores aleatorios complejos y matrices aleatorias en algunas palabras.

En varios campos, como en mecanica cuantica, procesamiento de señales, estadistica, estamos frente a datos complejas o puestas en matrices. Aún que se puede poner en biyección $\mathbb{C}$ y $\mathbb{R}^2$

o, similarmente, $\mathcal{M}_{d,d'}(\mathbb{R})$ y $\mathbb{R}^{dd'}$, puede parecer más natural trabajar conservando la estructura de los datos. Además, permite frecuentemente usar escrituras más compactas que pasar por vectores reales. Conservando la estructura implica ciertas matices en las escrituras de las distribuciones de probabilidades, momentos o funciones generadoras. Los vamos a ver brevemente, dejando el lector a libros especializados como (Gupta & Nagar, 1999) para tener más detalles.

1.9.1 Vectores aleatorios complejos

Formalmente, un vector aleatorio complejo se define de la misma manear que en el caso real, de la manera siguiente:

Definición 1-41 (Vector aleatorio complejo)**.** *Un vector aleatorio complejo es una función medible*

$$Z : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d), P_Z).$$

donde $\mathcal{B}(\mathbb{C}^d)$ son los borelianos de $\mathbb{C}^d$, σ-álgebra generada por los productos cartesianos $(-\infty\,;\, b_1] + \imath(-\infty\,;\, c_1] \times \cdots \times (-\infty\,;\, b_d] + \imath(-\infty\,;\, c_d]$ y donde la medida P_Z sobre $\mathcal{B}(\mathbb{C}^d)$ es la medida imágen de P. Como en el caso real,

$$(Z \in B) \equiv Z^{-1}(B) = \{\omega \in \Omega \;\big|\; Z(\omega) \in B\} \qquad \textit{y} \qquad P_Z(B) = P(Z \in B).$$

Sin embargo, se puede poner en biyección $\mathbb{C}^d$ y $\mathbb{R}^{2d}$ de tal manera de que se puede definir naturalmente un vector complejo aleatorio a partir de un vector aleatorio real de la manera alternativa equivalente (Lapidoth, 2017, Cap. 17):

Definición 1-42 (Vector aleatorio complejo – definición equivalente)**.** *Un vector aleatorio complejo se define como*

$$Z = X + \imath Y$$

donde $\widetilde{Z} \equiv \begin{bmatrix} X \\ Y \end{bmatrix}$ es un vector aleatorio de $\mathbb{R}^{2\,d}$. La medida de probabilidad imágen es entonces

$$P_Z \equiv P_{\widetilde{Z}} = P_{X,Y}$$

Resuelte de esta definición equivalente los hechos siguientes:

- La función de repartición de Z se escribe como la función de repartición conjunta de X e Y,

  $$F_Z \equiv F_{\widetilde{Z}} = F_{X,Y}$$

  Notando de que es una función de x e y, F_Z hace aparecer explicitamente ambos z y z^* complejos conjugados.

- Si la medida $P_{\widetilde{Z}}$ admite una derivada de Radon-Nykodým con respeto a la medida de Lebesgue sobre $\mathbb{R}^{2\,d}$, se define la densidad de probabilidad de Z como $f_Z \equiv f_{\widetilde{Z}} = f_{X,Y}$. A partir de la función de repartición, se escribe entonces o a través de la derivada $2\,d$-ésima de $F_{X,Y}$ con respeto a las compontentes x_i e y_i o, de manera equivalente,

$$f_Z(z) = \frac{\partial^{2\,d}}{\partial z_1 \cdots \partial z_d\, \partial z_1^* \cdots \partial z_d^*}$$

- Los momentos de orden K siendos definido a partir de las componentes de X y de Y, se definen también bajo la forma

$$m_{k_1,\ldots,k_d\,;\,k'_1,\ldots,k'_d} = \mathrm{E}\left[\prod_{i=1}^{d} Z_i^{k_i} \prod_{i=1}^{d} Z_i^{*\,k'_i}\right] \quad \text{con} \quad \sum_i (k_i + k'_i) = K$$

y similarmente para los momentos centrales $\zeta_{k_1,\ldots,k_d\,;\,k'_1,\ldots,k'_d}$. En particular,

  - La media de $Z = X + \imath Y$ es definida por

$$m_Z = \mathrm{E}[Z] = \mathrm{E}[X] + \imath\,\mathrm{E}[Y]$$

    La media de Z^* no lleva información más de orden 1.

  - La matriz de covarianza es definida por

$$\Sigma_Z \equiv \mathrm{Cov}[Z] \equiv \mathrm{E}\left[(Z - m_Z)(Z - m_Z)^\dagger\right]$$

    donde $Z^\dagger = (Z^*)^t$ dicho *transconjugado* (transpuesta conjugada). Fijense de que, volviendo al vector $\widetilde{Z}^t = \begin{bmatrix} X^t & Y^t \end{bmatrix}$ tenemos por un lado

$$\Sigma_{\widetilde{Z}} = \begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_{X,Y}^t & \Sigma_Y \end{bmatrix}$$

    conteniendo todas las convarianzas, y por el otro lado,

$$\Sigma_Z = (\Sigma_X + \Sigma_Y) - \imath\left(\Sigma_{X,Y} - \Sigma_{X,Y}^t\right)$$

    Se puede ver que la covarianza de Z no contiene todos los terminos de orden 2. Por eso, se define también la *pseudo-covarianza*, sin terminos conjugados,

$$\check{\Sigma}_Z \equiv \mathrm{pCov}[Z] \equiv \mathrm{E}\left[(Z - m_Z)(Z - m_Z)^t\right]$$

    Ahora, se puede ver que

$$\check{\Sigma}_Z = (\Sigma_X - \Sigma_Y) + \imath\left(\Sigma_{X,Y} + \Sigma_{X,Y}^t\right)$$

    Entonces, se recupera inmediatamente Σ_X, Σ_Y y $\Sigma_{X,Y}$ a partir de Σ_Z y $\check{\Sigma}_Z$; Claramente, los momentos centrales de orden 2 son dados por <u>ambas</u> Σ_Z y $\check{\Sigma}_Z$.

Los momentos así definidos heriden naturalmente de las propiedades de las del caso real.

- Se puede ver que $\Sigma_Z \in P_d(\mathbb{C})$ (hermitica semi-definida positiva), es decir que $\Sigma_Z = \Sigma_Z^\dagger$ y $\forall\, \mu \in \mathbb{C}, \quad \mu^\dagger \Sigma_Z \mu \geq 0$. Al revés, $\check{\Sigma}_Z \notin P_d(\mathbb{C})$; esta matriz es solamente simétrica $\check{\Sigma}_Z = \check{\Sigma}_Z^t \in S_d(\mathbb{C})$.

- Las generadoras son respectivamente equivalentes a las de $\widetilde{Z}$, o usando a la vez Z y Z^*. Por ejemplo, para la función característica, se la puede definir de argumento complejo como

$$\Phi_Z(\omega) = \mathrm{E}\left[e^{\imath \Re e\{\omega^\dagger Z\}}\right] \quad \text{con} \quad \omega \in \mathbb{C}^d$$

  (ver por ejemplo (Lapidoth, 2017, Cap. 17)). Las funciones generadoras así definidas heriden naturalmente de las propiedades de las del caso real. Entre otros, interpretando la función característica como función de ambos ω y ω^*, y derivando como si serían variables "independientes":

  - $\mathrm{E}[Z] = -2\,\imath\, \nabla_{\omega^*} \Phi_Z|_{\omega=0}$,
  - $\mathrm{Cov}[Z] = -4\, \mathcal{H}_{\omega^*,\omega} \Phi_Z|_{\omega=0} + 4\, \nabla_{\omega^*} \Phi_Z \nabla_\omega^t \Phi_Z|_{\omega=0}$,
  - $\mathrm{pCov}[Z] = -4\, \mathcal{H}_{\omega^*} \Phi_Z|_{\omega=0} + 4\, \nabla_{\omega^*} \Phi_Z \nabla_{\omega^*}^t \Phi_Z|_{\omega=0}$

En el marco de vectores complejos, aparece una subclase particular invariante por rotación, lo qu es conocido como vectores circulares:

Definición 1-43 (Vector aleatorio complejo circular)**.** *Un vector aleatorio complejo Z es dicho* circular *en torno*[27] *a un vector $\mu \in \mathbb{C}^d$ si para cualquier $\theta \in [0\,;\, 2\pi)$,*

$$e^{\imath\theta}\,(Z - \mu) \overset{\mathrm{d}}{=} Z - \mu$$

donde $\overset{\mathrm{d}}{=}$ significa "igualdad en distribución" (ver notaciones).

Los vectores circular tienen propiedades particulares importantes

- Si Z es circular al torno de un vector μ y admite una media, entonces

$$m_Z = \mathrm{E}[Z] = \mu$$

  Eso viene del hecho de que $e^{\imath\theta}\,\mathrm{E}\,[Z - \mu] = \mathrm{E}\left[e^{\imath\theta}(Z - \mu)\right] = \mathrm{E}\,[Z - \mu]$. Entonces, para cualquier $\theta \in [0\,;\, 2\pi)$, $\left(1 - e^{\imath\theta}\right)\mathrm{E}\,[Z - \mu] = 0$, lo que proba que $\mathrm{E}\,[Z - \mu] = 0$.

- Si Z es circular al torno de un vector μ y admite momentos de orden 2, entonces la pseudo-covarianza es nula,

$$\check{\Sigma}_Z = \mathrm{pCov}[Z] = 0$$

  Recordandose que $m_Z = \mu$, eso viene del hecho de que $e^{2\imath\theta}\,\mathrm{E}\,[(Z - m_Z)(Z - m_Z)^t] = \mathrm{E}\left[\left(e^{\imath\theta}(Z - m_Z)\right)\left(e^{\imath\theta}(Z - m_Z)\right)^t\right] = \mathrm{E}\,[(Z - m_Z)(Z - m_Z)^t]$. Entonces, para cualquier $\theta \in$

[27] En la literatura, la noción de circular es dada para $\mu = 0$ (Lapidoth, 2017, Def. 24.3.2), pero se extiende sin costo adicional al caso de la definición dada en este libro.

$[0\,;\,2\pi),\ \left(1-e^{2\,\imath\theta}\right)\mathrm{E}\left[(Z-m_Z)(Z-m_Z)^t\right]=0$, lo que cierra la prueba. La consecuencia es que en el contexto circular,

$$\Sigma_X=\Sigma_Y \quad \text{y} \quad \Sigma_{X,Y}^t=-\Sigma_{X,Y}$$

Fijense de que si la pseudo-covarianza de un vector aleatorio complejo es nula, eso no implica de que el vector es circular. Por ejemplo, sea Z tomando valores sobre $\mathcal{Z}=\{1+\imath\,;\,1-\imath\,;\,-1+\imath\,;\,-1-\imath\}$ con probabilidades $p=\left[\frac{1}{3}\quad\frac{1}{4}\quad\frac{1}{5}\quad\frac{13}{60}\right]^t$. No puede ser circular porque, por ejemplo $e^{\imath\frac{\pi}{4}}Z$ toma sus valores en $\{\sqrt{2}\,;\,-\sqrt{2}\,;\,\imath\sqrt{2}\,;\,-\imath\sqrt{2}\}\neq\mathcal{Z}$ o, por ejemplo, $e^{\imath\frac{\pi}{2}}Z$ toma sus valores en $\mathcal{Z}$ pero con el vector de probabilidad $p'=\left[\frac{1}{5}\quad\frac{1}{3}\quad\frac{13}{60}\quad\frac{1}{4}\right]^t\neq p$.

Cuando la pseudo-covarianza es nula se dice a veces que el vector es circular al orden 2. Más precisamente, en la literatura, se usa la definición siguiente (Lapidoth, 2017, Def. 17.4.1):

Definición 1-44 (Vector aleatorio complejo propio)**.** *Un vector aleatorio complejo Z es dicho* propio *(proper en ingles) si admite momentos hasta el orden 2 y ambos,*

$$\mathrm{E}[Z]=0, \qquad \mathrm{pCov}[Z]=0$$

Se podría ampliar esta definición hablando de vector propio al torno de un vector μ, conservando solamente la nulidad de la pseudo-covarianza. Como li vímos tranatdo de vectores circulares, tenemos la implicación siguiente:

Teorema 1-37 (Circularidad)**.** *Sea Z vector alatorio complejo. Entonces,*

$$Z \text{ circular en torno de } m \quad \Longrightarrow \quad Z \text{ propio en torno de } m$$

Los vectores propios tienen propiedades particulares, entre otros las siguientes.

Teorema 1-38 (Conservación del caracter propio por transformación lineal)**.** *Sea Z vector aleatorio complejo propio de $\mathbb{C}^d$, entonces, para cualquier matriz $A\in\mathcal{M}_{d',d}(\mathbb{C})$, el vector aleatorio AZ es propio.*

Demostración. La prueba es obiva, notando que $\mathrm{E}[AZ]=A\,\mathrm{E}[Z]=0$ y que $\mathrm{pCov}[AZ]=A\,\mathrm{pCov}[Z]A^t=0$. □

Teorema 1-39 (Caracter propio y proyección)**.** *Un vector aleatorio Z complejo de $\mathbb{C}^d$ es propio si y solamente para cualquier $c\in\mathbb{C}^d$, la variable c^tZ es propia.*

Demostración. Claramente, de $\mathrm{E}\left[c^tZ\right]=c^t\,\mathrm{E}[Z]$ y $\mathrm{pCov}\left[c^tZ\right]=c^t\,\mathrm{pCov}[Z]c$ tenemos que si Z es propio, $\mathrm{E}[Z]=0\ \Rightarrow\mathrm{E}\left[c^tZ\right]=0$ y $\mathrm{pCov}[Z]=0\ \Rightarrow\mathrm{pCov}\left[c^tZ\right]=0$.
Reciprocamente, si para cualquier c la variable c^tZ es propia, se puede elegir d vectores c_i^t puestas en una matriz C invertible. Entonces $\mathrm{E}[CZ]=0$ por hypotesis, lo que da $C\,\mathrm{E}[Z]=0\ \Rightarrow\ \mathrm{E}[Z]=0$ de la invertibilidad. De la misma manera, tenemos por hypothesis $\mathrm{pCov}[CZ]=0$ lo que significa que $C\,\mathrm{pCov}[Z]C^t=0\ \Rightarrow\ \mathrm{pCov}[Z]=0$ de la invertibilidad de C. □

1.9.2 Matrices aleatorias

De la misma manera que se puede querer trabajar con vectores complejos, a veces los datos son naturalmente puestas en matrices aleatorias. Por ejemplo, se puede querer estimar una matriz de covarianza a partir de una secuencia de vectores aleatorios $X_i,\ i = 1, \ldots, n$ de media nula y de misma ley. Un estimador natural es de reemplazar el promedio estadistico por un promedio empirico usando las observaciones/los vectores aleatorios $\widehat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^t$ (en practica, se reemplaza los X_i por observaciones/sampleos x_i y se evalua entonces un sampleo del estimador). Claramente $\widehat{\Sigma}_X$ es aleatoria por construcción, y tiene naturalmente la estructura de una matriz.

En lo que sigue, nos enfocaremos en las matrices reales. Para el caso complejo, se prodrá referirse a la subsección anterior. Notando que se puede poner en biyección $\mathcal{M}_{d,d'}(\mathbb{R})$ y $\mathbb{R}^{dd'}$, une manera de tratar de matrices aleatorias X puede ser de trabajar con su vectorización $\mathrm{vec}\,(X) = \begin{bmatrix} X_1^t & \cdots & X_{d'}^t \end{bmatrix}^t$ donde X_i es la i-ésima columna de X, las vectorizaciones de las operaciones matriciales (Magnus & Neudecker, 1979, Cap. 2) (ver también (Neudecker & Wansbeek, 1983; Harville, 2008)), y referirse a la sección tratando de vectores aleatorios. Pero resulte a veces más directo conservar la estructura matricial y trabajar con esa.

Formalmente, una matriz aleatorio real se define de la misma manear que en el caso de vectores reales, de la manera siguiente:

Definición 1-45 (Matriz aleatoria real)**.** *Una matriz aleatoria real es una función medible*

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathcal{M}_{d,d'}(\mathbb{R}), \mathcal{B}\left(\mathcal{M}_{d,d'}(\mathbb{R})\right), P_X)\,.$$

donde $\mathcal{B}\left(\mathcal{M}_{d,d'}(\mathbb{R})\right)$ son los borelianos de $\mathcal{M}_{d,d'}(\mathbb{R})$, σ-álgebra generada por los productos cartesianos $\bigtimes_{i=1,j=1}^{i=d,j=d'} (-\infty\,;\, b_{i,j}]$ y donde la medida P_X sobre $\mathcal{B}\left(\mathcal{M}_{d,d'}(\mathbb{R})\right)$ es la medida imágen de P. Nuevamente,

$$(X \in B) \equiv X^{-1}(B) = \{\omega \in \Omega \;\big|\; X(\omega) \in B\} \qquad \textit{y} \qquad P_X(B) = P(X \in B).$$

En lo que sigue, nos vamos a concentrar sobre dos situaciones particulares: (i) el caso general de matrices de $\mathcal{M}_{d',d}$ y (ii) el conjunto de matrices simétricas $S_d(\mathbb{R})$ (o un subconjunto) que tiene la particularidad de tener componentes iguales. El el último caso, a veces resuelte más comodo tener en cuenta esta simetria, es de decir que la matriz tiene a los más $\frac{d(d+1)}{2}$ componentes linealmente independientes.

1.9.2.1. Caso general

De manera general, trabajamos en el contexto de matrices no necesiaramente cuadradas, y con componentes potencialemnte linealmente independientes (sin simetria particular). En este marco general, X viviendo sobre $\mathcal{X} \subseteq \mathcal{M}_{d,d'}(\mathbb{R})$:

- La función de repartición F_X es la distribución conjunta de las componentes $X_{i,j}$. Si admite una densidad, se define como $p_X = \dfrac{\partial^{dd'} F_X}{\prod_{i=1}^{d}\prod_{j=1}^{d'} \partial x_{i,j}}$.

- Los momentos se definen como en el caso de vectores; Por ejemplo, todos los momentos de orden K son dados por

$$m_K = \mathrm{E}\left[X^{\otimes K}\right]$$

donde $\cdot^{\otimes K}$ es K veces el producto de Kronecker, y similarmente para los momentos centrales ζ_K. En particular,

  - La media es definida por

$$m_X = \mathrm{E}\left[X\right]$$

  - La covarianza es definida por

$$\Sigma_X \equiv \mathrm{Cov}[X] = \mathrm{E}\left[(X - m_X) \otimes (X - m_X)\right] = \mathrm{E}\left[X \otimes X\right] - m_X \otimes m_X$$

    Tipicamente, si la vemos en bloques de tamaño $d \times d$, la componente (k,l)-ésima del bloc (i,j)-ésima corresponde a $\mathrm{Cov}\left[X_{i,j}, X_{k,l}\right]$.
    Más generalmente, tratando de covarianza etre dos matrices aleatorias X e Y, se define la matriz de covarianza conjunta como

$$\Sigma_{X,Y} \equiv \mathrm{Cov}\left[X, Y\right] : \mathrm{E}\left[(X - m_X) \otimes (Y - m_Y)\right] = \mathrm{E}\left[X \otimes Y\right] - m_X \otimes m_Y$$

- Se puede escribir también las funciones generadoras con una forma matricial; por ejemplo, tratando de la función caractéristica, va a ser una función de dd' variables que se puede poner en una matriz de $\mathcal{M}_{d,d'}(\mathbb{R})$ de tal manera que

$$\Phi_X(\omega) = \mathrm{E}\left[e^{\imath\, \mathrm{Tr}\left(\omega^t X\right)}\right] \quad \text{con} \quad \omega \in \mathcal{M}_{d,d'}(\mathbb{R})$$

- Ahora es sencillo ver de que, si existent, se puede recuperar los momentos por diferenciación como en el caso de vectores,

$$-\imath \left.\frac{\partial \Phi_X}{\partial \omega_{i,j}}\right|_{\omega=0} = \mathrm{E}[X_{i,j}]$$

o

$$-\left.\frac{\partial^2 \Phi_X}{\partial \omega_{i,j} \partial \omega_{k,l}}\right|_{\omega=0} = \mathrm{E}\left[X_{i,j} X_{k,l}\right]$$

Se prodría referirse a (Magnus & Neudecker, 1999, Cap. 8) para usar las reglas de derivación matricial para hacer los calculos en la mayoría de los casos que se encuentran en la literatura (vamos a ver ejemplos en la sección 1.10).

1.9.2.2. Caso simétrico

En el contexto simétrico, *i. e.,* el espacio de llegada es $\mathcal{X} \subseteq S_d(\mathbb{R})$ (ej. el cono $P_d^+(\mathbb{R})$), aparece que por lo más la matriz tiene $\frac{d(d+1)}{2}$ componentes linealmente independientes. A veces resuelte más comodo definir la función característica en este caso con $\omega \in S_d(\mathbb{R})$ para respectar las simetrias del problema y no tener ninguna degenrencia de esa misma (ver por ejemplo (Peddada & Richards, 1991; Anderson, 2003)). A veces, es aún dificil o imposible calcular en
$M_{d,d}(\mathbb{R})$ entero. Eso tiene consecuencias:

- Si existen, se puede recuperar los momentos por diferenciación como en el caso de vectores, pero hay que tener en cuenta el hecho de que si $i \neq j$, $\omega_{i,j} = \omega_{j,i}$ aparece dos veces en ω. Entonces, por ejemplo, se puede ver inmediatamente que

$$-\imath \left.\frac{\partial \Phi_X}{\partial \omega_{i,j}}\right|_{\omega=0} = \left(2 - \mathbb{1}_{\{i\}}(j)\right) \mathrm{E}[X_{i,j}]$$

  o que

$$-\left.\frac{\partial^2 \Phi_X}{\partial \omega_{i,j} \partial \omega_{k,l}}\right|_{\omega=0} = \left(2 - \mathbb{1}_{\{i\}}(j)\right)\left(2 - \mathbb{1}_{\{l\}}(k)\right) \mathrm{E}\left[X_{i,j} X_{k,l}\right]$$

Descomposición de Bartlett, versión de la descomposición de Cholesky al caso de Wishart. Viene de que si $M \in \mathcal{P}_d(\mathbb{K})$, existe una matriz triangular inferior L tal que $M = LL^\dagger$(?, ?, Teo. 14.5.11)

Hablar de convergencia? DIBUJAR en el plano complejo, escalar, Φ_X? Unas curvas son "bellas" ver Sasvari 2013.

1.10 Algunos ejemplos de distribuciones de probabilidad

En esta sección, vamos a ver unos ejemplos de distribuciones que se encuentran frecuentemente en problema prácticos de varias areas cientificas (estadística, física, ingenería,...). Daremos las caracteristicas de cada ley presentada, así que sus propiedades remarcables. El numéro de leyes de probababilidad es tan importante que es dificil, para no decir imposible, ser exahustivo. Además, existen muchas relaciones entre leyes. En esta sección, vamos a ver algunas leyes que aparecen frecuentemente, y nos enfocaremos solamente sobre algunos vínculos entre leyes (los principales). Para tener más detalles, se puede referirse a los libros especializados en este marco, como por ejemplo (Spiegel, 1976; Johnson, Kotz & Kemp, 1992; Johnson et al., 1997; Johnson, Kotz & Balakrishnan, 1995a, 1995b; Kotz, Balakrishnan & Johnson, 2000; Gupta & Nagar, 1999; Fang et al., 1990; Samorodnitsky & Taqqu, 1994).

1.10.1 Distribuciones de variable discreta

1.10.1.1. Variable con certeza

El caso $X = a \in \mathbb{R}^d$ deterministico ($\forall\, \omega,\ X(\omega) = a$) puede ser ver visto como un caso degenerado de vector aleatorio. Visto así, sus características principales vistas en las secciones anteriores son resimudas en la tabla siguiente:

Dominio de definición	$\mathcal{X} = \{a\}, \quad a \in \mathbb{R}^d$
Distribución de probabilidad	$p_X(x) = \mathbb{1}_{\{a\}}(x)$
Promedio	$m_X = a$
Covarianza	$\Sigma_X = 0$
Sesgo (caso escalad)	$\gamma_X = 0$
Curtosis por exceso (caso escalar)	$\bar{\kappa}_X = -3$
Generadora de probabilidad	$G_X(z) = \prod_{i=1}^{d} z_i^{a_i}$ para $z_i \in \mathbb{C}$ si $a_i \geq 0$ y $\mathbb{C}^*$ si no
Generadora de momentos	$M_X(u) = e^{a^t u}$ para $u \in \mathbb{C}^d$
Función característica	$\Phi_X(\omega) = e^{\imath\, a^t \omega}$

La función de masa y función de repartición son representadas en la figura Fig. 1-14 en el caso escalar.

(a) (b)

Figura 1-14: Ilustración de una distribución cierta (a), y la función de repartición asociada (b).

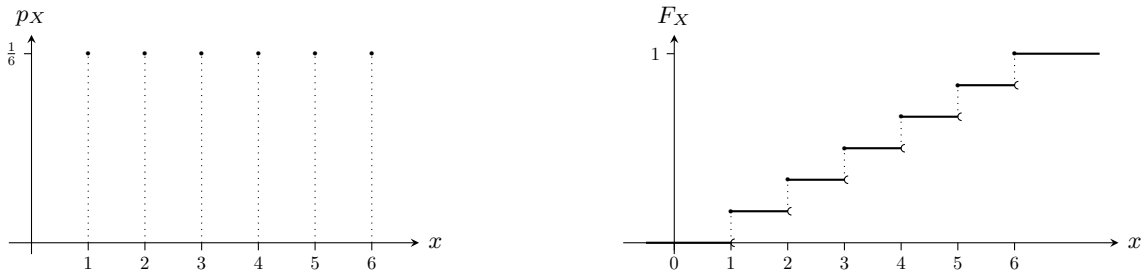
Poner acá la ley de los gran números? Más notas historicas.

1.10.1.2. Ley Uniforme sobre un "intervalo" de $\mathbb{Z}$

Se denota $X \sim \mathcal{U}\{a\,;\, b\}$ con $(a, b) \in \mathbb{Z}^2,\ b \geq a$. Las características de X son las siguientes:

Parámetros	$(a, b) \in \mathbb{Z}^2,\ b \geq a$
Dominio de definición	$\mathcal{X} = \{a\,;\, a+1\,;\, \ldots\,;\, b\}$
Distribución de probabilidad	$p_X(x) = \frac{1}{b-a+1}$
Promedio	$m_X = \dfrac{a+b}{2}$
Varianza	$\sigma_X^2 = \dfrac{(b-a)(b-a+2)}{12}$
Sesgo	$\gamma_X = 0$
Curtosis por exceso	$\bar{\kappa}_X = -\dfrac{6}{5}\dfrac{(b-a)(b-a+2)+2}{(b-a)(b-a+2)}$
Generadora de probabilidad	$G_X(z) = \dfrac{z^a - z^{b+1}}{1-z}$ para [28] $z \in \mathbb{C}$ si $a \geq 0$ y $\mathbb{C}^*$ sino
Generadora de momentos	$M_X(u) = \dfrac{e^{au} - e^{(b+1)u}}{1-e^u}$ para [29] $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = \dfrac{e^{\imath a\omega} - e^{\imath(b+1)\omega}}{1-e^{\imath\omega}}$

La distribución de masa de probabilidad y función de repartición de una variable uniforme $\mathcal{U}\{a\,;\, b\}$ son representadas en la figura Fig. 1-15.



(a) (b)

Figura 1-15: Ilustración de una densidad de probabilidad uniforme (a), y la función de repartición asociada (b). $a = 1,\ b = 6$ (ej. dado equilibriado).

Cuando $b = a$, la variable tiende a una variable cierta $X = a$.

[28] En el caso límite $z \to 1,\ \lim_{z\to 1} \frac{z^a - z^{b+1}}{1-z} = b + 1 - a$

[29] En el caso límite $u \to 0,\ \lim_{u\to 0} \frac{e^{au} - e^{(b+1)u}}{1-e^u} = b + 1 - a$, y similarmente para la función característica.

La distribución uniforme aparece por ejemplo en el tiro de un dado equilibriado con $a = 1,\ b = 6$.

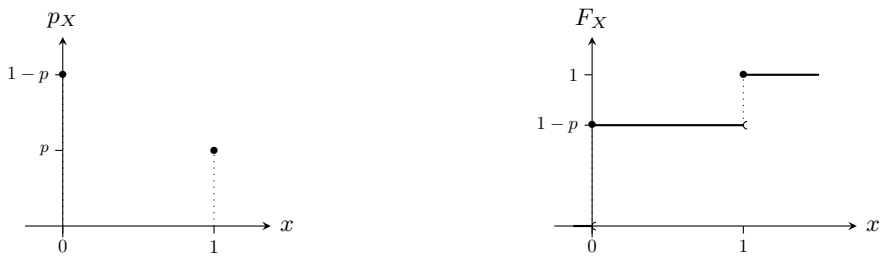
1.10.1.3. Ley de Bernoulli

Esta ley aparece cuando se hace una experiencia con 2 estados posible, tipo un tiro de moneda. Apareció en trabajos muy antiguos, entre otros el de J. Bernoulli tratando de la ley de gran números (Bernoulli, 1713; Hald, 1990).

Se denota $X \sim \mathcal{B}(p)$ con $p \in [0\,;\,1]$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{0\,;\,1\}$
Parámetro	$p \in [0\,;\,1]$
Distribución de probabilidad	$p_X(1) = 1 - p_X(0) = p$
Promedio	$m_X = p$
Varianza	$\sigma_X^2 = p\,(1-p)$
Sesgo	$\gamma_X = \dfrac{1-2\,p}{\sqrt{p\,(1-p)}}$
Curtosis por exceso	$\overline{\kappa}_X = \dfrac{1-6\,p+6\,p^2}{p\,(1-p)}$
Generadora de probabilidad	$G_X(z) = 1-p+pz$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = 1-p+p\,e^u$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = 1-p+p\,e^{\imath\omega}$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-16.



(a) (b)

Figura 1-16: Ilustración de una distribución de probabilidad de Bernoulli (a), y la función de repartición asociada (b), con $p = \frac{1}{3}$.

Nota que cuando $p = 0$ (resp. $p = 1$) la variable es cierta $X = 0$ (resp. $X = 1$).

La ley de Bernoulli tiene una propiedad de reflexividad trivial:

Lema 1-12 (Reflexividad)**.** *Sea* $X \sim \mathcal{B}(p)$. *Entonces*

$$1 - X \sim \mathcal{B}(1-p)$$

Demostración. El resultado es inmediato de $P(1-X=1) = P(X=0) = 1-p$. □

1.10.1.4. Ley Binomial

Esta ley apareció en trabajos muy antiguos, entre otros el de J. Bernoulli en 1713 (Bernoulli, 1713; Hald, 1990). Se la puede ver como una extension de la ley de Bernoulli a $n \geq 1$ tiros de una moneda, contando por ejemplo cuantas veces aparecen una cara.

Se denota $X \sim \mathcal{B}(n,p)$ con $n \in \mathbb{N} \setminus \{0\,;\,1\}, \quad p \in [0\,;\,1]$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{0\,;\,\ldots\,;\,n\}$
Parámetros	$n \in \mathbb{N} \setminus \{0\,;\,1\}, \quad p \in [0\,;\,1]$
Distribución de probabilidad	$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$
Promedio	$m_X = n\,p$
Varianza	$\sigma_X^2 = n\,p\,(1-p)$
Sesgo	$\gamma_X = \dfrac{1-2\,p}{\sqrt{n\,p\,(1-p)}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1-6\,p+6\,p^2}{n\,p\,(1-p)}$
Generadora de probabilidad	$G_X(z) = (1-p+pz)^n$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = (1-p+p\,e^u)^n$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = (1-p+p\,e^{\imath\omega})^n$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-17.

Otros ilustraciones para otros p?

Cuando $n = 1$, se recupera la lei de Bernoulli $\mathcal{B}(p) \equiv \mathcal{B}(1,p)$. Adémas, se muestra sencillamente usando la generadora de probabilidad que

Lema 1-13. *Sean* $X_i \sim \mathcal{B}(p), \quad i = 1,\ldots,n$ *independientes, entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{B}(n,p)$$

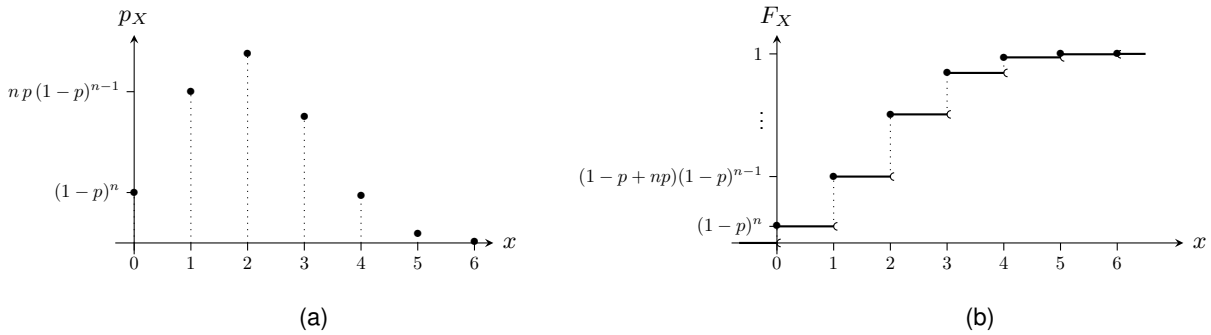


Figura 1-17: Ilustración de una distribución de probabilidad Binomial (a), y la función de repartición asociada (b), con $n = 6, \quad p = \frac{1}{3}$.

De este resultado, se puede notar que, por ejemplo, le distribución binomial aparece en el conteo de eventos independientes de misma probabilidad entre n.

También, la ley binomial tiene una propiedad de reflexividad, consecuencia directa de la de Bernoulli:

Lema 1-14 (Reflexividad)**.** *Sea* $X \sim \mathcal{B}(n,p)$. *Entonces*

$$n - X \sim \mathcal{B}(n, 1-p)$$

Si tomamos el ejemplo de una moneda que se tira n veces de maneras independientes, con probabilidad p que aparezca una cara, X representa el número de caras tiradas. Entonces, $n - X$ es el número de secas: en $n - X$ se intercambian los roles de la cara y seca. Más formalmente:

Demostración. El resultado es inmediato de la propiedad de reflexividad de la ley de Bernoulli, conjuntamente al lema **??**. Alternativamente, se nota que $P(n - X = x) = P(X = n - x) = \binom{n}{n-x} p^{n-x}(1-p)^x = \binom{n}{x}(1-p)^x p^{n-x}$ notando que $\binom{n}{n-x} = \binom{n}{x}$. □

Nota que cuando $p = 0$ (resp. $p = 1$) la variable es cierta $X = 0$ (resp. $X = n$).

1.10.1.5. Ley Binomial negativa

Formas particulares de esta ley aparecieron en los años 1679 en trabajos de Blaise Pascal (Pascal, 1679; Hald, 1990) [30] y un poco más tarde de P. R. de Montmort (de Montmort, 1713, p. 233-248). Esta ley aparece cuando se repite una experencia binaria exito/facasco con probabilidad p de exito, manera independiente hasta el r-ésimo facasco (r fijo), contendo el número de excitos obtenidos cuando se para la experiencia.

Se denota $X \sim \mathcal{B}_-(r,p)$ con $r \in \mathbb{N}^*, \quad p \in [0\,;\,1)$ y sus características son las siguientes:

[30] De hecho aparece en una carta de 1654 que mando B. PAscal a P. de Fermat, publicada mucho tiempo después.

Dominio de definición	$\mathcal{X} = \mathbb{N}$
Parámetros	$r \in \mathbb{N}^*, \quad p \in [0\,;\,1)$
Distribución de probabilidad	$p_X(x) = \binom{x+r-1}{x} p^x (1-p)^r$
Promedio	$m_X = \dfrac{r\,p}{1-p}$
Varianza	$\sigma_X^2 = \dfrac{r\,p}{(1-p)^2}$
Sesgo	$\gamma_X = \dfrac{1+p}{\sqrt{r\,p}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1+4\,p+p^2}{r\,p}$
Generadora de probabilidad	$G_X(z) = \left(\dfrac{1-p}{1-p\,z}\right)^r$ para $\lvert z\rvert < p^{-1}$
Generadora de momentos	$M_X(u) = \left(\dfrac{1-p}{1-p\,e^u}\right)^r$ para $\Re e\,\{u\} < -\ln p$
Función característica	$\Phi_X(\omega) = \left(\dfrac{1-p}{1-p\,e^{i\omega}}\right)^r$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-18. **Otros**

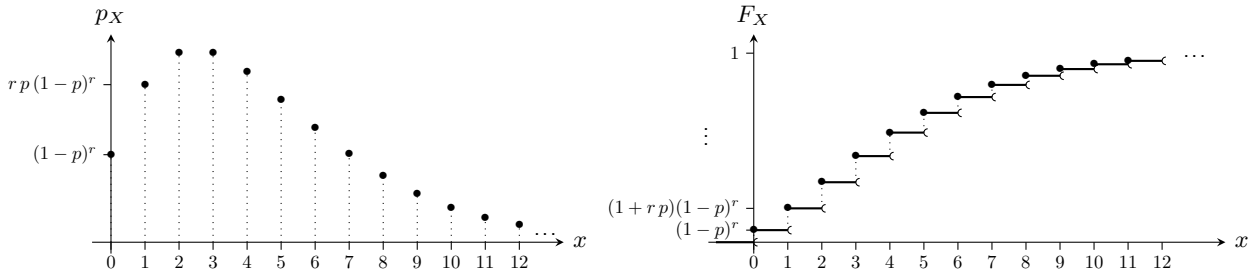


Figura 1-18: Ilustración de una distribución de probabilidad binomial negativa (a), y la función de repartición asociada (b), con $r = 3, \quad p = \frac{3}{5}$.

ilustraciones para otros r, p?

Recordar que esta ley aparece cuando se repite una experencia binaria $X_i \in \{0\,;\,1\}, i = 1, \ldots$ con $P(X_i = 1) = p$ de manera independiente (X_i independientes) hasta que r variables valen 0, con r fijo. El número de excito X sigue una ley $\mathcal{B}_-(r, p)$ (el calculo es directo). Dicho de otra manera, $X = \sum_{i=1}^{N} X_i$ con N variable aleatoria tal que $X_N = 0$ y $r = \sum_{i=1}^{N} (1 - X_i)$: condicionalmente a

N, la variable X es binomial de parámetro p, *i. e.*, $P(X=x|N=n)=\binom{n}{x}p^x(1-p)^{n-x}$. Se puede ver que $P(N=n)=\binom{n}{r-1}(1-p)^r p^{n-r}$ y la ley de la binomial negativa se recupera a través del teorema de probabilidad total 1-1 o también, a través del teorema 1-27.

Esta distribución se generaliza para $r\in\mathbb{R}_+^*$ pero se pierde la interpretación que vímos en el párafo anterior.

Nota: cuando $p=0$ la variable es cierta $X=r$.

1.10.1.6. Ley Multinomial

Esta ley es una generalización de la ley binomial y aparece por ejemplo cuando se repite una experiencia a k estados n veces de manera independiente y nos interesamos a la probabilidad que el primer evento aparece n_1 veces, el secundo n_2 veces, ...(ej. para $k=6$, contamos los números de 1, de 2, ...cuando tiramos n veces este dado). Aparece también esta ley por la primera vez en el trabajo de J. Bernoulli (Bernoulli, 1713; Hald, 1990) (ver también el ensayo de Montmort de 1708 con otras notaciones (de Montmort, 1713)).

Se denota $X\sim\mathcal{M}(n,p)$ con $n\in\mathbb{N}^*$ y $p=\begin{bmatrix}p_1 & \cdots & p_k\end{bmatrix}^t\in\Delta_{k-1}$ the $(k-1)$-simplex estandar (ver figure 1-30-(a) y notaciones). Entonces, a pesar de que se escribe X de manera k-dimensional, el vector partenece a un espacio claramente $d=k-1$ dimensional y en el caso $k=2$ se recupera la ley binomial. Las características de $X\sim\mathcal{M}(n,p)$ son las siguientes:

Dominio de definición [31]	$\mathcal{X}=\left\{x\in\{0\,;\,\dots\,;\,n\}^k \;\middle|\; \sum_{i=1}^k x_i=n\right\}$
Parámetros [32]	$n\in\mathbb{N}^*,\quad p\in\Delta_{k-1}$
Distribución de probabilidad [33]	$p_X(x)=\dfrac{n!}{\prod_{i=1}^k x_i!}\prod_{i=1}^k p_i^{x_i}$
Promedio	$m_X=n\,p$
Covarianza [34]	$\Sigma_X=n\left(\operatorname{diag}p-p\,p^t\right)$
Generadora de probabilidad [35]	$G_X(z)=\left(p^t z\right)^n$ para $z\in\mathbb{C}^k$
Generadora de momentos [36]	$M_X(u)=\left(p^t e^u\right)^n,\ e^u=\begin{bmatrix}e^{u_1} & \cdots & e^{u_k}\end{bmatrix}^t$ para $u\in\mathbb{C}^k$
Función característica [37]	$\Phi_X(\omega)=\left(p^t e^{\imath\omega}\right)^n$

[31] De hecho, se puede considerar que el vector aleatorio es $(k-1)$-dimensional $\widetilde{X}=\begin{bmatrix}\widetilde{X}_1 & \cdots & \widetilde{X}_{k-1}\end{bmatrix}^t$ definido sobre el dominio $\widetilde{\mathcal{X}}=\left\{x\in\{0\,;\,\dots\,;\,n\}^{k-1},\ \sum_{i=1}^{k-1}x_i\leq n\right\}$.

Su masa de probabilidad es representadas en la figura Fig. 1-19.

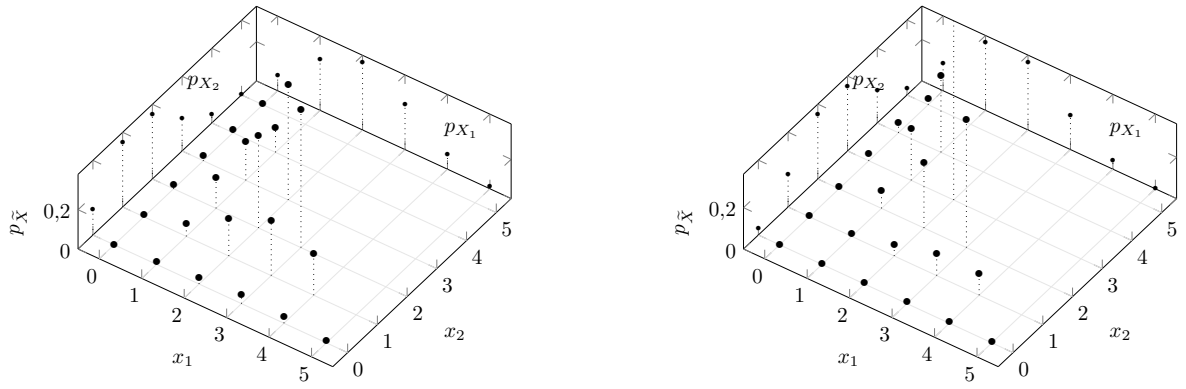


Figura 1-19: Ilustración de una distribución de probabilidad multinomial para $k = 3$ del vector $(k-1)$-dimensional $\widetilde{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^t$ ($X_3 = 1 - X_1 - X_2$) con las marginales p_{X_1}, p_{X_2} (ver notas de pie 31 y 33). Es dibujada solamente la distribución sobre $\widetilde{\mathcal{X}}$, siendo esta nula afuera de $\widetilde{\mathcal{X}}$. Los parámetros son $n = 5$ y $p = \begin{bmatrix} \frac{2}{5} & \frac{1}{3} & \frac{4}{15} \end{bmatrix}^t$ (a), $p = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}^t$ (b).

Notar: cuando $p = \mathbb{1}_i$, la variable es cierta $X = n\mathbb{1}_i$.

Otros ilustraciones para otros n, p?

Vectores de distribución multinomial tienen una propiedade notable con respecto a una permutación de variable, parecidas a la de la binomial:

Lema 1-15 (Efecto de una permutación)**.** *Sea* $X \sim \mathcal{M}(n, p)$, $p \in \Delta_{k-1}$ *y* $\Pi \in \mathfrak{S}_k(\mathbb{R})$ *matriz de permutación. Entonces*

$$\Pi X \sim \mathcal{M}(n, \Pi p)$$

Demostración. El resultado es inmediato saliendo de la función característica y aplicando el teorema 1-35 (recordar que $\Pi^{-1} = \Pi^t$). Más directamente, notando la permutation σ tal que $\Pi = \sum_{i=1}^{k} \mathbb{1}_i \mathbb{1}_{\sigma(i)}^t$,

---

[32] El parámetro de $\widetilde{X}$ es $\widetilde{p} = \begin{bmatrix} p_1 & \cdots & p_{k-1} \end{bmatrix}^t \in \left\{ q \in [0\,;\,1]^{k-1} \mid \sum_{i=1}^{k-1} q_i \leq 1 \right\}$.

[33] La masa de probabilidad de $\widetilde{X}$ es $p_{\widetilde{X}}(x) = \frac{n!}{\prod_{i=1}^{k-1} x_i!(n-\sum_{i=1}^{k-1} x_i)!} \prod_{i=1}^{k-1} p_i^{x_i} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{n-\sum_{i=1}^{k-1} x_i}$.

[34] $\Sigma_X \in P_k(\mathbb{R})$, pero de $\mathbb{1}^t \Sigma_X \mathbb{1} = 0$ viene $\Sigma_X \notin P_k^+(\mathbb{R})$. Eso es la consecuencia directa del hecho de que X d-dimensional, vive sobre Δ_{k-1}, $(d-1)$-dimensional.

[35] Notar: $G_{\widetilde{X}}(\widetilde{z}) = G_X\left(\begin{bmatrix} \widetilde{z} & 1 \end{bmatrix}^t\right)$ y al revés $G_X(z) = z_k^n\, G_{\widetilde{X}}\left(\begin{bmatrix} \frac{z_1}{z_k} & \cdots & \frac{z_{k-1}}{z_k} \end{bmatrix}^t\right)$.

[36] Notar: $M_{\widetilde{X}}(\widetilde{u}) = M_X\left(\begin{bmatrix} \widetilde{u} & 0 \end{bmatrix}^t\right)$ y $M_X(u) = e^{n\,u_k} M_{\widetilde{X}}\left(\begin{bmatrix} u_1 - u_k & \cdots & u_{k-1} - u_k \end{bmatrix}^t\right)$.

[37] Notar: $\Phi_{\widetilde{X}}(\widetilde{\omega}) = \Phi_X\left(\begin{bmatrix} \widetilde{\omega} & 0 \end{bmatrix}^t\right)$ o $\Phi_X(\omega) = e^{\imath\, n\, \omega_k} \Phi_{\widetilde{X}}\left(\begin{bmatrix} \omega_1 - \omega_k & \cdots & \omega_{k-1} - \omega_k \end{bmatrix}^t\right)$.

se puede ver que $P(\Pi X = x) = P(X = \Pi^{-1}x) = \dfrac{n!}{\prod_{i=1}^{k} x_{\sigma^{-1}(i)}!} \prod_{i=1}^{k} p_i^{x_{\sigma^{-1}(i)}} = \dfrac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_{\sigma(i)}^{x_i}$ por cambio de indices. □

Además ley multinomial exhibe una stabilidad remplazando dos componentes por su suma:

Lema 1-16 (Stabilidad por agregación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_k \end{bmatrix}^t \sim \mathcal{M}(n,p)$, $p \in \Delta_{k-1}$ *y* $G^{(i,j)}$ *matriz de agrupación de las* (i,j)*-ésima componentes (ver notaciones). Entonces,*

$$G^{(i,j)}X \sim \mathcal{M}\left(n, G^{(i,j)}p\right)$$

Este resultado es intuitivo en el hecho de que vuelve a agrupar los estados i e j en un estado, que tiene entonces la probabilidad $p_i + p_j$ de aparecer.

Demostración. Suponemos $i < j$ (el otro caso se recupera por simetría). A partir de la función característica y el teorema 1-35 se tiene,

$$\begin{aligned} \forall\, \omega \in \mathbb{R}^{k-1}, \quad \Phi_{G^{(i,j)}X}(\omega) &= \Phi_X\left(G^{(i,j)\,t}\omega\right) \\ &= \left(\sum_{l=1}^{k} p_l\, e^{\imath \left(G^{(i,j)\,t}\omega\right)_l}\right)^n \end{aligned}$$

Ahora, se nota que $G^{(i,j)\,t}\omega = \begin{bmatrix} \omega_1 & \cdots & \omega_{j-1} & \omega_i & \omega_{j+1} & \cdots & \omega_{k-1} \end{bmatrix}^t$, entonces

$$\begin{aligned} \forall\, \omega \in \mathbb{R}^{k-1}, \quad \Phi_{G^{(i,j)}X}(\omega) &= \left(\sum_{l=1, l\neq j}^{k} p_l\, e^{\imath\, \omega_l} + p_j\, e^{\imath\, \omega_i}\right)^n \\ &= \left(\sum_{l=1, l\neq i, l\neq j}^{k} p_l\, e^{\imath\, \omega_l} + (p_i + p_j)\, e^{\imath\, \omega_i}\right)^n \end{aligned}$$

lo que cierra la prueba. Se puede tener un enfoque más directo, con los mismos pasos que en la prueba del lema 1-21 tratando de la ley hipergeometrica multivaluada. □

De este lema, aplicado de manera recursiva, se obtiene los corolarios siguientes:

Corolario 1-8. *Sea* $X \sim \mathcal{M}(n,p)$*, entonces* $X_i \sim \mathcal{B}(n, p_i)$*.*

Al final, por una analisis combinatorial, se muestra sencillamente un resultado similar al de la binomial como suma de Bernoulli independientes:

Lema 1-17. *Sean* $U_i, \quad i = 1, \ldots, n,\ j = 1, \ldots, n$ *discretas sobre* $\mathcal{U} = \{1\,; \ldots ; k\}$ *de masa de probabilidad* $p_{U_i} = p \in \Delta_{k-1}$*, independientes, y* $X_i = \mathbb{1}_{U_i} \in \mathbb{R}^k$*. Entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{M}(n,p)$$

Nota: esta ley se generaliza de la misma manera que para la binomial negativa, dando una ley multinomial negativa o, de manera equivalente, generalizando la binomial negativa a más de dos clases se obtiene la ley multinomial negativa. **Anadirlo en una seccion?**

1.10.1.7. Ley hipergeometrica

Esta ley aparece por ejemplo cuando se hace una experiencia con una población de tamaño n (ej. n bolas en una urna), que pueden partenecer a dos clases, con k numéro de elementos de la primera clase (a veces dicho estados de excito; ej. k bolas negras), $n-k$ numéro de elementos de la secunda clase, y se hace m tiros si reemplazamiento. X es el número de tiros parteneciendo en la primera clase (número de excitos). Esta ley apareció en trabajos de de Moivre en 1710 (de Moivre, 1710; Hald, 1990).

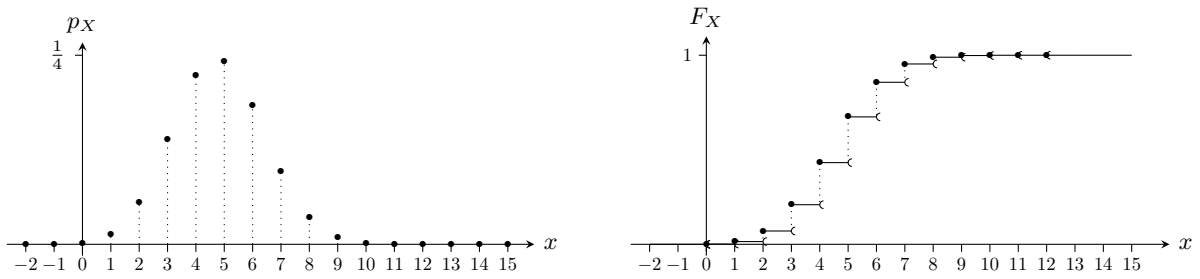
Se denota $X \sim \mathcal{H}(n,k,m)$ con $n \in \mathbb{N}^*, \quad k \in \{0\,;\,\ldots\,;\,n\}, \quad m \in \{0\,;\,\ldots\,;\,m\}$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \{\text{máx}(0, k+m-n)\,;\,\ldots\,;\,\text{mín}(k,m)\}$
Parámetros	$n \in \mathbb{N}^*$ (población) $k \in \{0\,;\,\ldots\,;\,n\}$ (número de estados exitosos) $m \in \{0\,;\,\ldots\,;\,n\}$ (número de tiros)
Distribución de probabilidad	$p_X(x) = \frac{\binom{k}{x}\binom{n-k}{m-x}}{\binom{n}{m}}$
Promedio	$m_X = \frac{m}{n}\,k$
Varianza [38]	$\sigma_X^2 = \begin{cases} \frac{m\,(n-m)}{n^2(n-1)}\,k\,(n-k) & \text{si} \quad n > 1 \\ 0 & \text{si} \quad n = 1 \end{cases}$
Generadora de probabilidad	$G_X(z) = \frac{\binom{n-k}{m}}{\binom{n}{m}}\ {}_2F_1(-m,-k;n-m-k+1;z)$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = \frac{\binom{n-k}{m}}{\binom{n}{k}}\ {}_2F_1\left(-m,-k;n-m-k+1;e^u\right)$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = \frac{\binom{n-k}{m}}{\binom{n}{m}}\ {}_2F_1\left(-m,-k;n-m-k+1;e^{\imath\,\omega}\right)$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-20.

Otros ilustraciones para otros n,k,m?

[38] En el caso degenerado $n=1$, o $m=0$, o $m=1=n$; en ambos casos, la variable es cierta (ver fin de la subsección).



(a) (b)

Figura 1-20: Ilustración de una distribución de probabilidad Hipergeometrica (a), y la función de repartición asociada (b), con $n = 100, \quad k = 12, \quad m = 40$.

Poner el Sesgo (ya lo tengo)? El Curtosis (lo tengo que simplificar)? muy pesadas... Momento factorial $f_q = \frac{(m)^{\underline{q}}(k)^{\underline{q}}}{(n)^{\underline{q}}}$ permitiendo calcular todo.

Notar: la variable resuelta cierta en los casos siguientes

- $m = 0 \Rightarrow X = 0$: no se sortean elementos, así que siempre se sortea 0 elementos de la primera clase;
- $m = n \Rightarrow X = k$: si se sortean todos los elementos de la población, se sortean todos los k de la primera clase;
- $k = 0 \Rightarrow X = 0$: si la primera clase no tiene elementos, no se puede tirar elementos de esta clase;
- $k = n \Rightarrow X = m$: al revés si la secunda clase no tiene elementos, todos los sorteados partenecen a la primera clase.

La ley tiene propiedades de reflexividad del mismo tipo que para la ley binomial:

Lema 1-18 (Reflexividad)**.** *Sea* $X \sim \mathcal{H}(n, k, m)$. *Entonces*

$$m - X \sim \mathcal{H}(n, n - k, m) \qquad \textit{y} \qquad k - X \sim \mathcal{H}(n, k, n - m)$$

Se puede ver que si en una urna con bolas negras y blancas, con k bolas negras, y X es el número de bolas negras sorteadas, $m - X$ representa las bolas blancas sorteadas. Es decir que en $m - X$ se intercambia los roles de las bolas negras y blancas. De la misma manera, $k - X$ representa las bolas negras que quedan en la urna, entre las $n - m$ que quedan, es decir que en $k - X$ se intercambia los roles de las bolas sorteadas y las que quedan en la urna. Más formalmente:

Demostración. El primer resultado es inmediato de $P(m - X = x) = P(X = m - x) = \frac{\binom{k}{m-x}\binom{n-k}{x}}{\binom{n}{m}}$. El secundo de $P(k - X = x) = P(X = k - x) = \frac{\binom{k}{k-x}\binom{n-k}{m-k+x}}{\binom{n}{m}} = \frac{\binom{k}{x}\binom{n-k}{n-m-x}}{\binom{n}{n-m}}$ notando que $\binom{a}{b} = \binom{a}{a-b}$. □

1.10.1.8. Ley hipergeometrica Negativa

Esta ley aparece por ejemplo cuando se hace una experiencia del mismo tipo que para la hipergeometrica, con una población de tamaño n (ej. n bolas en una urna), que pueden partenecer a dos clases, con k numéro de elementos de la primera clase estados de excito; ej. k bolas negras), $n-k$ numéro de elementos de la secunda clase. Pero en lugar de hacer m tiros fijos, se hace tiros hasta que r elementos de la seconda clase (fracascos) sean tiradas. X es el número de tiros parteneciendo en la primera clase (número de excitos). Es decir que cuando $X = x$, tenemos k elementos de la primera clase en los "primeros" $x + r - 1$ tiros, el últimos parteneciendo a la seconda clase. Parece que se encuentran las primeras huellas de esta ley en trabajos del marquesano de Condorcet en 1785 (Marquis de Condorcet, 1785).

Se denota $X \sim \mathcal{H}_-(n,k,r)$ con $n \in \mathbb{N}^*$, $k \in \{0\,;\ldots;n\}$, $m \in \{0\,;\ldots;n-k\}$ y sus características son las siguientes:

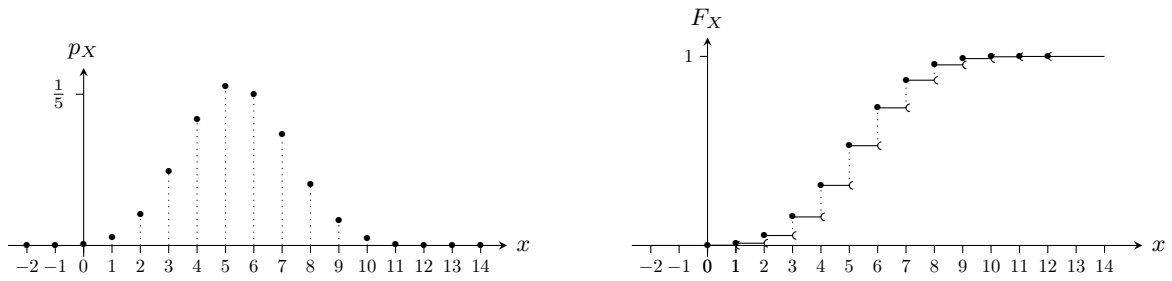
Dominio de definición	$\mathcal{X} = \{0\,;\ldots;k\}$
Parámetros	$n \in \mathbb{N}^*$ (población) $k \in \{0\,;\ldots;n\}$ (número de estados exitosos) $r \in \{0\,;\ldots;n-k\}$ (número de fracascos para parar)
Distribución de probabilidad [39]	$p_X(x) = \begin{cases} \frac{\binom{x+r-1}{x}\binom{n-r-x}{k-x}}{\binom{n}{k}} & \text{si} \quad r > 0 \\ \mathbb{1}_{\{0\}}(x) & \text{si} \quad r = 0 \end{cases}$
Promedio	$m_X = \frac{r\,k}{n-k+1}$
Varianza	$\sigma_X^2 = \frac{r\,k(n+1)(n-k-r+1)}{(n-k+1)^2(n-k+2)}$
Generadora de probabilidad	$G_X(z) = \frac{\binom{n-r}{k}}{\binom{n}{k}}\ {}_2F_1(r,-k;r-n;z)$ sobre $\mathbb{C}$
Generadora de momentos	$M_X(u) = \frac{\binom{n-r}{k}}{\binom{n}{k}}\ {}_2F_1\left(r,-k;r-n;e^u\right)$ sobre $\mathbb{C}$
Función característica	$\Phi_X(\omega) = \frac{\binom{n-r}{k}}{\binom{n}{k}}\ {}_2F_1\left(r,-k;r-n;e^{\imath\,\omega}\right)$

Poner sesgo y curtosis? Expresiones muy pesadas... Momento factorial $f_q = \frac{(r)^{\overline{q}}(k)_{\underline{q}}}{(n-k+1)^{\overline{q}}}$ permi-

[39] Para los $x + r - 1$ primeros tiros, de la primera clase hay $\binom{k}{x}$ combinaciones posibles, y $\binom{n-k}{r-1}$ de la seconda clase, sobre los $\binom{n}{x+r-1}$ combinaciones posibles en total. Para el último tiro, quedan $n-k-(r-1)$ posibilidades de la seconda clase sobre las $n-x-(r-1)$ elementos que quedan.

tiendo calcular todo.

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-21.



(a) (b)

Figura 1-21: Ilustración de una distribución de probabilidad Hipergeometrica (a), y la función de repartición asociada (b), con $n = 100, \quad k = 12, \quad r = 40$.

Otros ilustraciones para otros n, k, r?

Notar: cuando $k = 0$, la variable es cierta $X = r$ (se sortean solamente elementos de la seconda clase, así que para siempre cuando se han tirados r elementos); cuando $r = 0$, también la variable es cierta $X = 0$ (no se sortan bolas, así que no hay de la primera clase).

1.10.1.9. Ley hipergeometrica multivariada

Esta ley aparece por ejemplo cuando se se generaliza la ley hipergeometrica con $c > 2$ clases con k_i numéro de elementos de la clase i, $\sum_i k_i = n$. Se estudia esta ley, entre otros, por la primera vez, en el ensayo de Montmort en 1708 (de Montmort, 1713), o más tarde, en 1740, en trabajos de Simpson (Simpson, 1740; Hald, 1990).

Se denota $X \sim \mathcal{HM}(n, k, m)$ con $n \in \mathbb{N}, \quad k = \begin{bmatrix} k_1 & \cdots & k_c \end{bmatrix}^t \in \left\{ \{0\,; \ldots ; n\}^c \;\middle|\; \sum_{i=1}^c k_i = n \right\}$, $m \in \{0\,; \ldots ; n\}$.

Entonces, como en el caso de la ley multinomial, a pesar de que se escribe X de manera c-dimensional, el vector partenece a un espacio claramente $(c-1)$-dimensional. Notar que en el caso $c = 2$ se recupera la ley hipergeometrica.

Sus características son las siguientes:

Dominio de definición [40]	$\mathcal{X} = \left\{ x \in \bigtimes_{i=1}^{c} \{0\,;\,\ldots\,;\,k_i\} \;\middle|\; \sum_{i=1}^{c} x_i = m \right\}$
Parámetros [41]	$n \in \mathbb{N}^*$ (población) $c \in \mathbb{N}^*$ (número de clases) $k \in \left\{ q \in \{0\,;\,\ldots\,;\,n\}^c \;\middle|\; \sum_{i=1}^{c} q_i = n \right\}$ (números de elementos de cada clase) $m \in \{0\,;\,\ldots\,;\,n\}$ (número de tiros)
Distribución de probabilidad [42]	$p_X(x) = \dfrac{\prod_{i=1}^{c} \binom{k_i}{x_i}}{\binom{n}{m}}$
Promedio	$m_X = \dfrac{m}{n}\, k$
Covarianza [43]	$\Sigma_X = \begin{cases} \frac{m(n-m)}{n^2(n-1)} \left(n \operatorname{diag} k - k k^t\right) & \text{si} \quad n > 1 \\ 0 & \text{si} \quad n = 1 \end{cases}$

Ver si se calcula Phi

La masa de probabilidad es representada en la figura Fig. 1-22.

Notar: cuando $c = 2$ se recupera la ley hipergeometrica; además X resuelta cierta en los casos siguientes (ver subesección anterior para las explicaciones/ilustraciones):

- $m = 0 \;\Rightarrow\; X = 0$;
- $m = n \;\Rightarrow\; X = k$;
- $k = n\mathbb{1}_i \;\Rightarrow\; X = m\mathbb{1}_i$.

Vectores de distribución hipergeometricas multivaluada tienen propiedades notables similares a las de la hipergeometricas y de la multinomial, a saber de tipo reflexividad, con respecto a una permutación de variable y con respecto a una agregación.

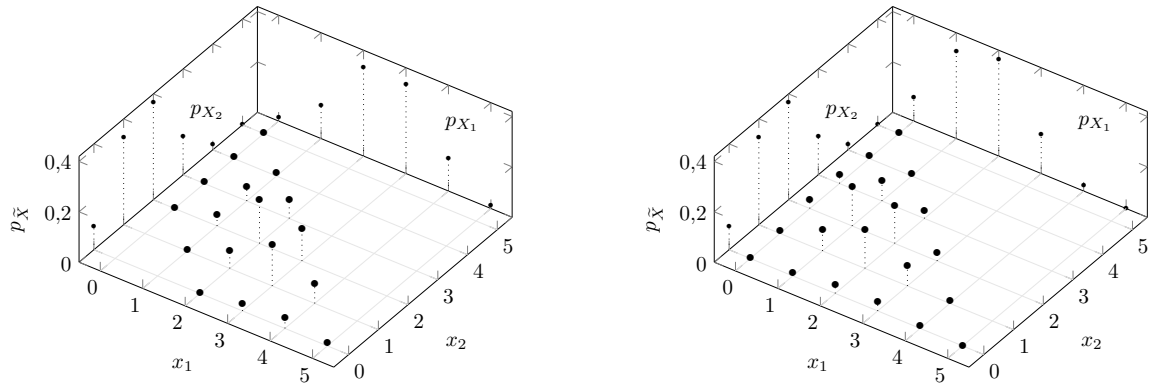
---

[40] De hecho, se puede considerar que el vector aleatorio es $(c-1)$-dimensional $\widetilde{X} = \begin{bmatrix} \widetilde{X}_1 & \cdots & \widetilde{X}_{c-1} \end{bmatrix}^t$ definido sobre el dominio $\widetilde{\mathcal{X}} = \left\{ x \in \bigtimes_{i=1}^{c-1} \{0\,;\,\ldots\,;\,k_i\} \;\middle|\; \text{máx}\left(0\,,\, \sum_{i=1}^{c-1} k_i + m - n\right) \leq \sum_{i=1}^{c-1} x_i \leq m \right\}$.

[41] Los parámetros de $\widetilde{X}$ son $n \in \mathbb{N}^*$, $m \in \{0\,;\,\ldots\,;\,n\}$ y $\widetilde{k} = \begin{bmatrix} k_1 & \cdots & k_{c-1} \end{bmatrix}^t \in \left\{ q \in \{0\,;\,\ldots\,;\,n\}^{c-1} \;\middle|\; \sum_{i=1}^{c-1} q_i \leq n \right\}$.

[42] La masa de probabilidad de $\widetilde{X}$ es $p_{\widetilde{X}}(x) = \dfrac{\prod_{i=1}^{c-1} \binom{k_i}{x_i} \binom{n - \sum_{i=1}^{c-1} k_i}{m - \sum_{i=1}^{c-1} x_i}}{\binom{n}{m}}$..

[43] Ver notas de pie 38 y 34.



(a) (b)

Figura 1-22: Ilustración de una distribución de probabilidad hipergeometrica multivariada para $c = 3$ del vector $(c-1)$-dimensional $\widetilde{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^t$ $(X_3 = m - X_1 - X_2)$ con las marginales p_{X_1}, p_{X_2} (ver notas de pie 40 y 42). Es dibujada solamente la distribución sobre $\widetilde{\mathcal{X}}$, siendo esta nula afuera de $\widetilde{\mathcal{X}}$. Los parámetros son $n = 18,\ m = 5,\ k = \begin{bmatrix} 9 & 6 & 3 \end{bmatrix}^t$ (a), $k = \begin{bmatrix} 6 & 6 & 6 \end{bmatrix}^t$ (b).

Lema 1-19 (Reflexividad)**.** *Sea* $X \sim \mathcal{HM}(n, k, m)$*. Entonces*

$$k - X \sim \mathcal{HM}(n, k, n - m)$$

Como el en contexto escalar, si en una urna tenemos bolas de c colores diferentes, con un número k_i para el i-ésimo color, X_i es el numéro de este color que se sorteó y $k_i - X_i$ representan las de este color que quedan en la urna, entre las $n - m = \sum_{i=1}^{c}(k_i - X_i)$ que quedan, es decir que en $k - X$ se intercambia los roles de las bolas sorteadas y las que quedan en la urna. Más formalmente:

Demostración. Sea $Y = k - X$. De $P(Y = y) = P(k - X = y) = P(X = k - y) = \frac{\prod_{i=1}^{c}\binom{k_i}{k_i - y_i}}{\binom{n}{m}} = \frac{\prod_{i=1}^{c}\binom{k_i}{y_i}}{\binom{n}{n-m}}$ notando que $\binom{a}{b} = \binom{a}{a-b}$. Se cierra la prueba recordandose que $\sum_{i=1}^{c} k_i = n$ y $\sum_{i=1}^{c} x_i = m$, dando $\sum_{i=1}^{c} k_i = n$ y $\sum_{i=1}^{c} y_i = n - m$. □

Lema 1-20 (Efecto de una permutación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_c \end{bmatrix}^t \sim \mathcal{HM}(n, k, m)$ *y* $\Pi \in \mathfrak{S}_c(\mathbb{R})$ *matriz de permutación. Entonces*

$$\Pi X \sim \mathcal{HM}\left(n, \Pi k, m\right)$$

Demostración. La prueba sigue paso paso la de la multinomial. Notando la permutation σ tal que $\Pi = \sum_{i=1}^{c} \mathbb{1}_i \mathbb{1}_{\sigma(i)}^t$, se puede ver que $P(\Pi X = x) = P(X = \Pi^{-1}x) = \dfrac{\prod_{i=1}^{c}\binom{k_i}{x_{\sigma^{-1}(i)}}}{\binom{n}{m}} = \dfrac{\prod_{i=1}^{c}\binom{k_{\sigma(i)}}{x_i}}{\binom{n}{m}}$

por cambio de indices. □

Lema 1-21 (Stabilidad por agregación). *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_c \end{bmatrix}^t \sim \mathcal{HM}(n,k,m)$ *y* $G^{(i,j)}$ *matriz de agrupación de las* (i,j)*-ésima componentes (ver notaciones). Entonces,*

$$G^{(i,j)}X \sim \mathcal{HM}\left(n, G^{(i,j)}k, m\right)$$

Este resultado es intuitivo en el hecho de que vuelve a agrupar las clases i e j en una clase, que tiene entonces $k_i + k_j$ elementos.

Demostración. Del lema precediente, notando que existen matrices de permutación [44] $\Pi_k \in \mathfrak{S}_k(\mathbb{R})$ y $\Pi_{k-1} \in \mathfrak{S}_{k-1}(\mathbb{R})$ tal que $G^{(i,j)} = \Pi_{k-1}\, G^{(c-1,c)}\, \Pi_k$, se puede concentrarse en el caso $(i,j) = (c-1,c)$. Ahora, claramente,

$$\begin{aligned} P(G^{(c-1,c)}X = x) &= P\left(\bigcap_{i=1}^{c-2}\left(X_i = x_i\right) \cap \left(X_{c-1} + X_c = x_{c-1}\right)\right) \\ &= \sum_{t=0}^{x_{c-1}} P\left(\bigcap_{i=1}^{c-2}\left(X_i = x_i\right) \cap \left(X_{c-1} = t\right) \cap \left(X_c = t - x_{c-1}\right)\right) \\ &= \frac{\prod_{i=1}^{c-2}\binom{k_i}{x_i}}{\binom{n}{m}} \sum_{t=0}^{x_{c-1}} \binom{k_{c-1}}{t}\binom{k_c}{x_c - t} \end{aligned}$$

Se cierra la prueba de la identidad de Chu-Vandermonde [45] $\sum_{t=0}^{l}\binom{r}{t}\binom{s}{l-t} = \binom{r+s}{l}$ (Knuth, 1997, Ec. (21), p. 59) o (Gradshteyn & Ryzhik, 2015, Ec. 0.156). □

De este lema, aplicado de manera recursiva, se obtiene el corolario siguiente:

Corolario 1-9. *Sea* $X \sim \mathcal{HM}(n,k,m)$*, entonces* $X_i \sim \mathcal{H}(n,k_i,m)$.

Nota: esta ley se generaliza de la misma manera que para la hipergeometrica negativa, dando una ley hipergeometrica negativa multivariada o, de manera equivalente, generalizando la hipergeometrica negativa a más de dos clases se obtiene la ley hipergeometrica negativa. **Anadirlo en una seccion?**

1.10.1.10. Ley Geométrica

[44] Π_k pone las componentes i e j el las posiciones $c-1$ y c, sin cambiar el orden de las precedientes; Π_{k-1} trazlada la última componente en la posición $\min(i,j)$.

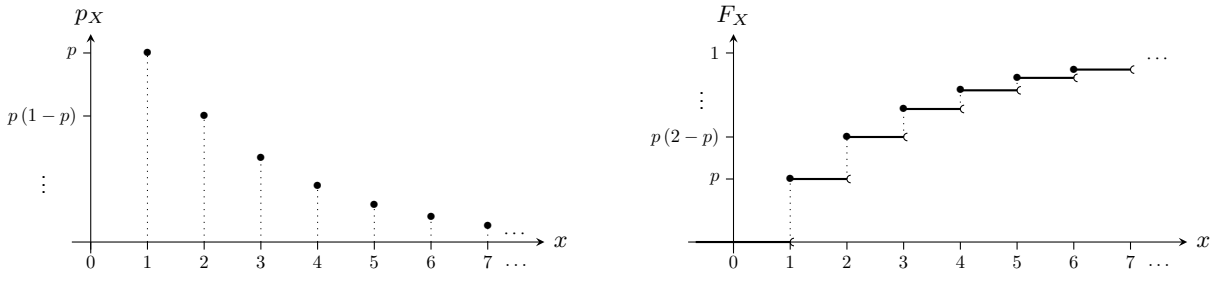
[45] Este identidad es debido a A.-T. Vandermonde en 1772, pero esta conocida desde 1303 por el matematico chino Chu Shi-Chieh, explicando la denominación de este identidad (Andersen & Larsen, 1994) o (Askey, 1975, p. 59-60). Se prueba escribiendo $(1+x)^{r+s} = (1+x)^r(1+x)^s$ y desorrando con la formula del binomio cada potencia.

La ley geométrica es un caso particular de la ley binomial negativa para $r = 1$, como ya lo hemos evocado. Dicho de otra manera, esta distribución aparece en el conteo de conteo de une repetición de una experiencia de maneja independiente hasta que occure un evento de probabilidad p; por ejemplo el número de tiro de un dado equilibriado hasta que occurre un "6" sigue una ley geométrica de parámetro $p = \frac{1}{6}$.

Se denota $X \sim \mathcal{G}(p)$ con $p \in (0\,;\,1]$ y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \mathbb{N}^*$
Parámetro	$p \in (0\,;\,1]$
Distribución de probabilidad	$p_X(x) = (1-p)^{x-1}p$ (convención $0^0 = 1$)
Promedio	$m_X = \frac{1}{p}$
Varianza	$\sigma_X^2 = \dfrac{1-p}{p^2}$
Sesgo	$\gamma_X = \dfrac{2-p}{\sqrt{1-p}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6-6\,p+p^2}{1-p}$
Generadora de probabilidad	$G_X(z) = \dfrac{pz}{1-(1-p)z}$ para $\lvert z\rvert < \frac{1}{1-p}$
Generadora de momentos	$M_X(u) = \dfrac{p\,e^u}{1-(1-p)\,e^u}$ para $\Re e\,\{u\} < -\ln(1-p)$
Función característica	$\Phi_X(\omega) = \dfrac{p\,e^{\imath\omega}}{1-(1-p)\,e^{\imath\omega}}$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-23. **Otros**



(a) (a)

Figura 1-23: Ilustración de una distribución de probabilidad Geométrica (a), y la función de repartición asociada (b), con $p = \frac{1}{3}$.

ilustraciones para otros p?

Como ya lo hemos evocado, esta ley esta vínculada con la binomial negativa, siendo un caso particular:

Lema 1-22 (Vínculo con la ley Binomial negativa)**.** *Sea* $X \sim \mathcal{B}_-(1,p)$*, entonces tenemos también*

$$X \sim \mathcal{G}(1-p).$$

Si volvemos a la representation de $X \sim \mathcal{B}_-(1,p)$ como $X = \sum_{i=1}^{N} X_i$ con $X_i \sim \mathcal{B}(p)$ independientes, N tal que $X_N = 0$ y $\sum_{i=1}^{N}(1-X_i) = 1$, aparece que, también, $N \sim \mathcal{G}(1-p)$.

Nota que cuando $p = 1$ la variable es cierta $X = 1$.

1.10.1.11. Ley de Poisson

Esta ley fue introducida por Poisson en 1837 como caso límite de la ley binomial para n grande, con el producto np fijo (Poisson, 1837, Cap. 3), (Hald, 1990). Se intereso en su estudio del comportamentio probabilistico del conteo de experiencia de Bernoulli bajo la hipotesis de independencia (dando lugar a la ley binomial) en ciencia humana, para una población importante (n grande), pero con un valor promedio dado. De hecho, se conocia esta ley, también como caso límite de la binomial, por lo menos desde un trabajo de de Moivre unas decadas antes (de Moivre, 1710). Aparecio también más tarde en muchos procesos físicos, como el conteo de desintegración atomica por secundo en un material radioactivo, o, (aproximadamente) a través del conteo de particulas que caen en una pequeña superficia cuanto se tiras particulas uniformamente en una grande superficia en trabajos de W. S. Gosset [46] (Student, 1907).

Se denota $X \sim \mathcal{P}(\lambda)$ con $\lambda \in \mathbb{R}_+^*$ llamada *taza*, y sus características son las siguientes:

[46] Fue connocido bajo en nombre “Student”; ver nota de pie 64.

Dominio de definición	$\mathcal{X} = \mathbb{N}$
Parámetro	$\lambda \in \mathbb{R}_+^*$
Distribución de probabilidad	$p_X(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$
Promedio	$m_X = \lambda$
Varianza	$\sigma_X^2 = \lambda$
Sesgo	$\gamma_X = \dfrac{1}{\sqrt{\lambda}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{1}{\lambda}$
Generadora de probabilidad	$G_X(z) = e^{\lambda(z-1)}$ para $z \in \mathbb{C}$
Generadora de momentos	$M_X(u) = e^{\lambda(e^u-1)}$ para $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = e^{\lambda\,(e^{\imath\omega}-1)}$

Su masa de probabilidad y función de repartición son representadas en la figura Fig. 1-24. **Otros**

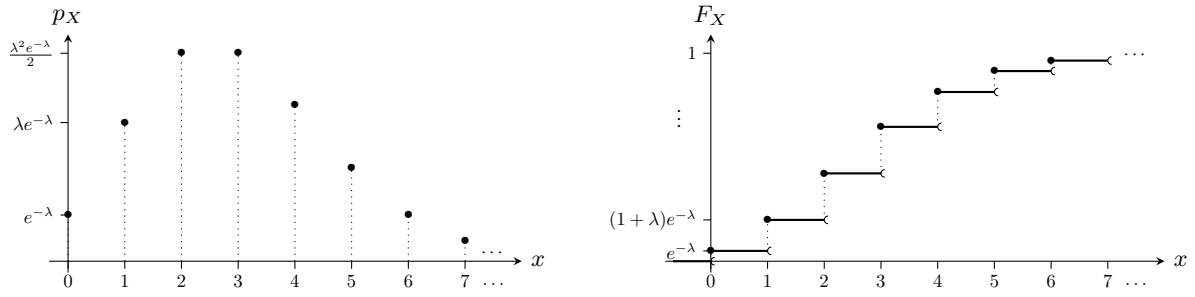


Figura 1-24: Ilustración de una distribución de probabilidad de Poisson (a), y la función de repartición asociada (b), con $\lambda = 3$.

ilustraciones para otros λ?

Adémas, se muestra sencillamente usando la generadora de probabilidad que

Lema 1-23 (Stabilidad)**.** *Sean* $X_i \sim \mathcal{P}(\lambda_i), \quad i = 1, \ldots, n$ *independientes, entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{P}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Como lo hemos introducido, la ley de Poisson esta vínculada a la ley binomial, como caso límite:

Lema 1-24 (Vínculo con la ley binomial)**.** *Sean* $X_n \sim \mathcal{B}\left(n\,,\,\frac{\lambda}{n}\right)$ *con* $\lambda > 0$ *fijo, entonces*

$$X_n \xrightarrow[n\to\infty]{\mathrm{d}} X \sim \mathcal{P}(\lambda)$$

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución (ver notaciones).*

Demostración. Se sale de la forma de la distribución binomial y de la formula de Stirling [47] para $\log\Gamma(z) = \left(z - \frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + o(1)$ en $z \to +\infty$ (Stirling, 1730; Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015). □

Aparece que la ley de Poisson esta vínculada también a la ley binomial negativa, también como caso límite:

[47] De hecho, esta formula es probablemente debida previamente a A. De Moivre (de Moivre, 1733, 1756; Pearson, 1924; Le Cam, 1986; Dutka, 1991; Deming, 1933), y fue mejorada por Stirling más tarde. Fue mejorada aún más por el famoso matemático S. Ramanujan recientemente (Andrew & Berndt, 2013, § 4.1).

Lema 1-25 (Vínculo con la binomial negativa)**.** *Sean* $X_r \sim \mathcal{B}_- \left(\frac{\lambda}{r+\lambda}, r\right)$ *con* $\lambda > 0$ *fijo, entonces*

$$X_R \xrightarrow[r\to\infty]{\mathrm{d}} X \sim \mathcal{P}(\lambda)$$

Demostración. Se sale de nuevo la forma de la distribución binomial negativa y de la formula de Stirling para probarlo. □

Más allá del contexto discreto, esta ley esta también vínculada a ley exponencial, por el processo dicho de Poisson. Si eventos pueden aparecer de manera aleatoria en el tiempo tal que, entre dos eventos, el tiempo sigue una ley exponencial de parámetro λ, y que estos tiempos son independientes, entonces dado un intervalo T de tiempo , el número de estos eventos sigue una ley de Poisson de parámetro λT. Lo vamos a ver en el ejemplo de la ley exponencial más adelante.

Al final, nota que cuando $\lambda = 0$ la variable es cierta $X = 0$ (usando la convención $0^0 = 1$).

1.10.1.12. Distribución seria de potencia (power series distributions)?

Estadística de los números de ocupación de niveles energéticos: distribuciones de Maxwell–Boltzmann, de Fermi–Dirac, y de Bose–Einstein (Rényi, 2007, p. 37-38)

Leyes de los grandes números; DeMoivre-Laplace

1.10.2 Distribuciones de variable continua

Antes de ir más adelante, notamos que, tratando de un vector aleatorio X continuo, de densidad de probabilidad $p_X(x)$, para cualquier $a \in \mathbb{R}^*$ el vector $Y_a = aX$ va a ser obviamente continuo, de densidad de probabilidad $p_{Y_a}(y) = \frac{1}{|a|} p_X\left(\frac{y}{a}\right)$. Ahora, cuando $a \to 0$, queda claro que el vector Y_a tiende al vector 0, deterministico. O, de un punto de vista de variable aleatoria, Y_a tiende al vector cierto, discreto. Un punto que puede parecer sopredente es que tal vector no admite una densidad de probabilidad más. De hecho la densidad p_{Y_a} tiende a una función generalizada, o distribución de Schwarz, más precisamente la "función Dirac", como lo hemos visto en el fin de la sección 1.3.4. Como lo hemos enfatizado, en tal caso preferimos seguir trabajando con la medida de probabilidad, que tiende a la medida de Dirac.

1.10.2.1. Distribución uniforme sobre un intervalo

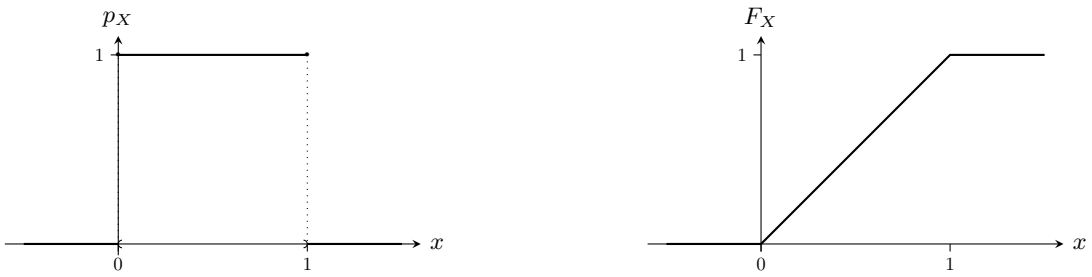
Esta distribución es la más natural que se usa cuando queremos modelar una falta de información sobre una variable, sabiendo que vive en un espacio de volumen finito: sin a priori más, una tendencia natural/intuitiva es de asignar la "misma probabilidad" a cada punto del conjunto. En particular, aparece así naturalmente en la inferencia Bayesiana que consiste a modelar como aleatorio un parámetro que se quierre inferir (Robert, 2007) (la ley es dicha ley *a priori*; ver también (Bayes, 1763) o (de Laplace, 1820); tal a priori es conocido como a a priori de Laplace).

Se denota $X \sim \mathcal{U}([a\,;\,b])$. Las características de X son las siguientes:

Dominio de definición	$\mathcal{X} = [a\,;\,b]$
Parámetros	$(a, b) \in \mathbb{R},\ b > a$
Densidad de probabilidad	$p_X(x) = \frac{1}{b-a}$
Promedio	$m_X = \dfrac{a+b}{2}$
Varianza	$\sigma_X^2 = \dfrac{(b-a)^2}{12}$
Sesgo	$\gamma_X = 0$
Curtosis por exceso	$\bar{\kappa}_X = -\dfrac{6}{5}$
Generadora de momentos	$M_X(u) = \dfrac{e^{bu} - e^{au}}{u}$ para [48] $u \in \mathbb{C}^d$
Función característica	$\Phi_X(\omega) = \dfrac{e^{\imath a\omega} - e^{\imath b\omega}}{\imath\,\omega}$

Obviamente, se puede escribir $X \overset{\mathrm{d}}{=} a + (b-a)U$ donde $\overset{\mathrm{d}}{=}$ significa que la equalidad es en distribución (las variables tienen la misma distribución de probabilidad), con $U \sim \mathcal{U}([0\,;\,1])$ llamada *uniforme estandar*.

La densidad de probabilidad y función de repartición de la variable estandar son representadas en la figura Fig. 1-25.



(a) (b)

Figura 1-25: Ilustración de una densidad de probabilidad uniforme (a), y la función de repartición asociada (b).

Una nota importante es que cada ley continua es vínculada a la ley uniforme sobre $(0\,;\,1)$ de la

[48] En el caso límite $u \to 0,\ \lim_{u\to 0} \frac{e^{bu}-e^{au}}{u} = b - a$, y similarmente para la función característica

manera siguiente:

Lema 1-26 (Inversión)**.** *Sea X, continua sobre $\mathcal{X} \subset \mathbb{R}$, de función de repartición F_X. Entonces*

$$U \equiv F_X(X) \sim \mathcal{U}(0\,;\,1)$$

Reciprocamente, definiendo la función de repartición inversa (o quantile)

$$F_X^{-1}(u) = \text{ínf}\{x \;\big|\; F(x) \geq u\}$$

si $V \sim \mathcal{U}(0\,;\,1)$,

$$Y = F_X^{-1}(V) \quad \Rightarrow \quad F_Y(y) = F_X(y)$$

Cuando F_X se inversa sencillamente eso da una manera sencille de tirar sampleos de función de repartición F_X a partir de sampleos tirados según una ley uniforme.

Demostración. Inmediatamente, F_X siendo creciente,

$$\begin{aligned} P(U \leq u) &= P(F_X(X) \leq u) \\ &= P(X \leq F_X^{-1}(u)) \\ &= F_X\left(F_X^{-1}(u)\right) \end{aligned}$$

Similarmente

$$\begin{aligned} P(Y \leq y) &= P(F_X^{-1}(V) \leq y) \\ &= P(V \leq F_X(y)) \\ &= F_X(y) \end{aligned}$$

□

De manera general, para cualquier ensemble $\mathcal{D} \subset \mathbb{R}^d$ de volumen $|\mathcal{D}|$ la variable uniforma sobre $\mathcal{D}$ tiene la densidad con respecto a la medida "natural" sobre $\mathcal{D}$ (Lebesque, discreta,...) constante sobre $\mathcal{D}$,

$$p_X(x) = \frac{1}{|\mathcal{D}|}\mathbb{1}_{\mathcal{D}}(x)$$

La media va a ser el centro de gravedad de $\mathcal{D}$.

Vamos a ver el el capítulo siguiente que esta distribución es la distribución definida sobre un conjunto de volumen finito que maximiza la entropía, *i. e.,* que es la "menos informativa". Por ejemplo, si se busca un parámetro modelizado como aleatorio (enfoque Bayesiano), definido sobre un conjunto de volumen finito, sin a priori más, una tendencia natural/intuitiva es de asignar la "misma probabilidad" a cada punto del conjunto. Aparece así naturalmente en la inferencia Bayesiana (Robert, 2007).

Notar que cuando $b \to a$, la variable tiende a una variable cierta $X = a$ (ver principio de esta sección).

1.10.2.2. Distribución normal o Gaussiana multivariada real

En el caso escalar, esta ley parece aparecer por unas de las primeras veces en trabajos de de Moivre como approximación de la ley binomial para n grande, usando la formula de Stirling (de Moivre, 1730, 1733, 1756; Pearson, 1924; Pearson, de Moivre & Archibald, 1926; Deming, 1933; Hald, 1984, 1990; Johnson et al., 1995a; Hald, 2006). Se puede ver también el trabajo de F. Galton, quien contruyó un experimento, la caja dicha de Galton, que ilustra por una parte como se puede obtener la ley binomial como suma de Bernoulli, y la convergencia a la Gausiana (Galton, 1889, Figs. 7-9, p. 63) o (Pearson, 1920, p. 38). Aparte de Moivre, la ley gausiana fue desarollado mucho por los matemáticos como Gauss en el estudio del movimiento de planetas con perturbaciones (predicción de la trayectoria de Céres) (Gauss, 1809; Pearson, 1924; Hald, 2006), basado en trabajos de A. M. Legendre (Legendre, 1805; Hald, 2006), o Laplace en mismo tipos de problema (Laplace, 1809a, 1809b; de Laplace, 1820; Pearson, 1924; Hald, 2006). De hecho, apoyandose en trabajos de de Moivre, la formalizó antes y más claramente Laplace, quien revandicó entonces su partenidad (ver por ejemplo (Pearson, 1920)). Por eso, esta ley es también conocida como ley de Laplace-Gauss.

En el contexto multivariado, la extensión natural de la ley binomial siendo al ley multinomial, es sin sorpresa que se introdujo la gausiana multivaluada como approximación de la multinomial. Este trabajo es debido entre otros a J. L. Lagrange en los años 1770, con correcciones debido unas decadas después a A. de Morgan (de Morgan, 1838). Pero apareció antes en el caso bidimensional, en particular a través del estudio del coeficiente de correlación entre variables aleatorias (ver por ejemplo trabajos de Galton (Galton, 1877a, 1877b; Pearson, 1920)).

A pesar de que parece menos natural en la modelisación de fenomenos aleatorio que leyes uniformes, la ley gausiana es seguramente unas de las más importante en probabilidad, sino que la más importante y la más expendida en la naturaleza. Eso viene sin duda del teorema del límite centrale. En dos palabras, cuando se suman un numero importante de variables aleatorias (independientes, de misma ley, admitiendo una varianza, o con menos restricciones (Athreya & Lahiri, 2006, Cap. 11)), corectamente normalizado, esta suma tiende a una gausiana [49]. En la naturaleza, se puede ver el ruido (señales) como suma de un número importante de fuentes de ruido independientes, justificando el modelo gausiano (Feller, 1971; Le Cam, 1986; Ash & Doléans-Dade, 1999; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Rényi, 2007; Billingsley, 2012). Adémas, como lo vamos a ver en el capitulo **??**, esta ley es la de incerteza máxima (maximizando la entropía) teniendo una dada varianza. Aparece naturalmente en termodínamica (gaz perfecto, con un número muy alto de particulas) (Maxwell, 1867; Boltzmann, 1896, 1898; Gibbs, 1902; Jaynes, 1965). En estimación, bajo la hipotesis gausiana, los estimadores de parámetros minimizando el error cuadratico promedio son generalmente lineal (Kay,

[49] De hecho, la approximación de la ley binomial por una gausiana cuando n es grande es una caso particular del teorema, siendo la binomial una suma de Bernoulli independientes.

1993; Robert, 2007). Todas estas consideraciones dan a la ley gausiana un rol central en la teoría de las probabilidades.

Se denota $X \sim \mathcal{N}(m, \Sigma)$ con $m \in \mathbb{R}^d$ y $\Sigma \in P_d^+(\mathbb{R})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{R})$ símetricas definidas positivas. Las características de la Gaussiana son las siguientes:

Dominio de definición	$\mathcal{X} = \mathbb{R}^d$
Parámetros	$m \in \mathbb{R}^d,\ \Sigma \in P_d^+(\mathbb{R})$
Densidad de probabilidad	$p_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \lvert\Sigma\rvert^{\frac{1}{2}}} e^{-\frac{1}{2}(x-m)^t \Sigma^{-1}(x-m)}$
Promedio	$m_X = m$
Covarianza	$\Sigma_X = \Sigma$
Sesgo (caso escalar)	$\gamma_X = 0$
Curtosis por exceso (caso escalar)	$\bar{\kappa}_X = 0$
Generadora de momentos	$M_X(u) = e^{u^t \Sigma u + u^t m}$ para $u \in \mathbb{C}^d$
Función característica	$\Phi_X(\omega) = e^{-\frac{1}{2}\omega^t \Sigma \omega + \imath \omega^t m}$

Nota: trivialmente, se puede escribir $X \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} N + m$ con $N \sim \mathcal{N}(0, I)$ donde N es dicha *Gausiana estandar* o *centrada-normalizada*. Las características de X son vínculadas a las de N (y vice-versa) por transformación afine (ver secciones anteriores).

La densidad de probabilidad gausiana y la función de repartición en el caso escalar son representadas en la figura Fig. 1-26-(a) y (b) y una densidad en un contexto bi-dimensional figura Fig. 1-26(c).

La gaussiana tiene un par de propiedades particulares:

Teorema 1-40 (Stabilidad)**.** *Sean $A_i, i = 1, \ldots, n$ matrices de $\mathbb{R}^{d' \times d}, d' \leq d$ de rango lleno, $b_i \in \mathbb{R}^{d'}$ y $X_i \sim \mathcal{N}(m_i, \Sigma_i)$ independientes, entonces*

$$\sum_{i=1}^{n} (A_i X_i + b_i) \sim \mathcal{N}\left(\sum_{i=1}^{n} (m_i + b_i)\ ,\ \sum_{i=1}^{n} A_i \Sigma_i A_i^t \right)$$

En particular, cualquier combinación lineal de los componentes de un vector Gaussiano da una gaussiana. Reciprocamente, si cualquier combinación lineal de los componentes de un vector aleatorio sigue una ley gaussiana, entonces el vector es gaussiano.

Demostración. Este resultato se proba usando función característica de la gaussiana, conjuntamente al teorema 1-35. □

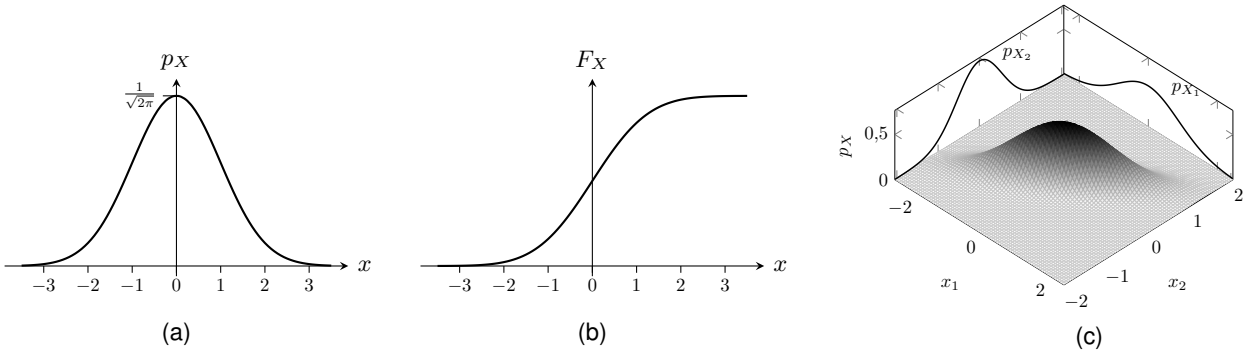


Figura 1-26: Ilustración de una densidad de probabilidad gaussiana escalar estandar (a), y la función de repartición asociada (b), así que una densidad de probabilidad gaussiana bi-dimensional centrada y de matriz de covarianza $\Sigma_X = R(\theta)\Delta^2 R(\theta)^t$ con $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ matriz de rotación y $\Delta = \operatorname{diag}\left(\begin{bmatrix} 1 & a \end{bmatrix}\right)$ matriz de cambio de escala, y sus marginales $X_1 \sim \mathcal{N}\left(0, \cos^2\theta + a^2\sin^2\theta\right)$ y $X_2 \sim \mathcal{N}\left(0, \sin^2\theta + a^2\cos^2\theta\right)$ (ver más adelante). En la figura, $a = \frac{1}{4}$ y $\theta = \frac{\pi}{6}$.

Teorema 1-41 (Independencia)**.** *Sea* $X \sim \mathcal{N}(m, \Delta)$ *con* $\Delta = \operatorname{diag}\left(\begin{bmatrix} \sigma_1^2 & \cdots & \sigma_d^2 \end{bmatrix}^t\right)$ *diagonal. Entonces las componentes* $X_i \sim \mathcal{N}(m_i, \sigma_i^2)$ *son independientes.*

Demostración. Este resultato se proba trivialmente escribiendo la densidad de probabilidad, notando que se factorisa. □

Hemos visto que cuando un vector tiene componentes independientes, la matriz de covarianza es diagonal (lema 1-6), pero que la reciproca es falsa en general. El último teorema muestra que la reciproca vale en el caso gausiano.

Volvemos ahora al rol central de la gausiana como modelo probabilistico muy frecuente de fenomeos aleatorios. Este rol particular viene del teorema del límite central que ya introdujimos. A veces es conocido como teorema de Lindenberg-Feller (por lo menos la forma con condiciones más debiles que en la formulación original). Para unas de las formulaciones originales, se puede referirse al trabajo de Laplace de 18P9 o de 1912 (Laplace, 1809a, 1809b; de Laplace, 1820). El nombre “central” viene de un documento de G. Pólya de 1920, titulado “Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem” (“Sobre el teorema del límite central del cálculo probabilístico y el problema de los momentos; el teorema es central... (Polya, 1920; Le Cam, 1986)). Se enuncia de manera siguiente (Spiegel, 1976; Brockwell & Davis, 1987; Ash & Doléans-Dade, 1999; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Billingsley, 2012):

Teorema 1-42 (Teorema del límite central)**.** *Sea* $\{X_i\}_{i\in\mathbb{N}^*}$ *una serie de vectores aleatorios independientes, de misma ley, y que admiten un promedio* m *y una matriz de covarianza* Σ*. Entonces*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{m}(X_i - m) \xrightarrow[n\to+\infty]{\mathrm{d}} Y \sim \mathcal{N}(0, \Sigma)$$

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución (ver notaciones).*

Demostración. Hay varias pruebas de este resultado. Quizás la más simple uas la función cáracteristica. Sin perdida de generalidad, supongamos que $m = 0$. Sea $Y_n = \frac{1}{n}\sum_{i=1}^{m} X_i$. Sea ω fijo. Por independencia y relaciones del teorema 1-35:

$$\begin{aligned}
\Phi_{Y_n}(\omega) &= \left(\Phi_{X_i}\left(\frac{\omega}{\sqrt{n}}\right)\right)^n \\
&= \left(\Phi_{X_i}(0) + \frac{1}{\sqrt{n}}\omega^t\nabla_\omega\Phi_{X_i}(0) + \frac{1}{2n}\omega^t\mathcal{H}_\omega\Phi_{X_i}(0)\omega + o\left(n^{-1}\right)\right)^n \\
&= \left(1 - \frac{1}{2n}\omega^t\Sigma\omega + o\left(n^{-1}\right)\right)^n \\
&\xrightarrow[n\to+\infty]{} \exp\left(-\frac{1}{2}\omega^t\Sigma\omega\right)
\end{aligned}$$

porque $\Phi_{X_i}(0) = 1$, X_i siendo de media nula el gradiente de la función característica se cancela en $\omega = 0$, y $\mathcal{H}_\omega\Phi_{X_i}(0) = -\Sigma$. Se reconoce ahora la función característica de la gausiana, lo que prueba que la función característica de Y_n converge simplemente hacia la función característica de la gausiana. Se cierra la pruba usando el teorema de convergencia de Lévy, diciendo que la convergencia simple de la función característica implica la convergencia en distribución (Ash & Doléans-Dade, 1999; Billingsley, 2012; Athreya & Lahiri, 2006). □

Existen varias variantes de este teorema que enunciamos, sin dar la prueba. Dejamos el lector a libros más especializados como (Ash & Doléans-Dade, 1999; Billingsley, 2012; Athreya & Lahiri, 2006; Lindeberg, 1922).

Teorema 1-43 (Teorema de Lindenberg-Feller)**.** *Sean* $\{X_i\}_{i\in\mathbb{N}^*}$ *vectores aleatorios independientes, no necesariamente de misma distribución de probabilidad, con* X_i *de media* $m_i = \mathrm{E}[X_i]$ *y de matriz de covarianza* $\Sigma_i \in P_d^+(\mathbb{K})$*. Sean* $C_n = \sum_{i=1}^n \Sigma_i$*,* c_n^2 *al autovalor más pequeña de* C_n*, y* $Y_n = C_n^{-\frac{1}{2}}\sum_{i=1}^n (X_i - \mathrm{E}[X_i])$*.*

$$\textit{Si}\quad \lambda_n > 0 \quad \textit{y} \quad \forall\,\varepsilon > 0, \quad \lim_{n\to+\infty}\sum_{i=1}^{n}\mathrm{E}\left[\left\|\frac{X_i - m_i}{c_n}\right\|^2 \mathbb{1}_{[\varepsilon\,;+\infty)}\left(\left\|\frac{X_i - m_i}{c_n}\right\|\right)\right] = 0$$

entonces

$$Y_n \xrightarrow[n\to+\infty]{\mathrm{d}} Y \sim \mathcal{N}(0, I)$$

En numerosos libros, este teorema es dado en el caso escalar. Se extiende sencillamente al caso multivariado gracia a lo que es conocido como *teorema de Cramér-Wold*, diciendo que una secuencia de vecotores aleatorios $Y_n \xrightarrow{\mathrm{d}} Y$ si y solamente para cualquier $u \in \mathbb{R}^d$ $u^tY_n \xrightarrow{\mathrm{d}} u^tY$ (Ash & Doléans-Dade, 1999; Athreya & Lahiri, 2006; Billingsley, 2012).

Sin dar la prueba, la condición de Lindenberg dice que si la suma de las “dispersiónes” de los vectores normalizado por los que es basicamente la varianza más de los componentes de la suma

(una vez diagonalizada) se concentra asintoticamente, la suma renormalizada de lo vectores centrados tiende a la gausiana (en distribución).

Se puede ver que se satisface la condición de Lindeberg en el caso de variables independientes de misma ley, del hecho que $C_n = n\Sigma$, lo que da $c_n^2 = nc^2$ con c^2 autovalor más pequeña de Σ. A continuación da la condición $\lim_{n\to\infty} \mathrm{E}\left[\|X_i - m_i\|^2 \mathbb{1}_{[\varepsilon\,;+\infty)}\left(\left\|\frac{X_i - m_i}{\sqrt{n}c}\right\|\right)\right] = 0$, satisfecha porque el argumento de la función indicadora tiende a 0 (casi siempre).

Un otra caso "trivial" aparece cuando la secuencia es uniformamente acotada, *i. e.,* $\forall i, \quad \|X_i\| \leq M$. Se puede retomar los argumentos anteriores, remplazando las variables por la cota.

Nota: si se satisface la condición dicha *de Lyapunov*, si existe $\delta > 0$ tal que

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}\left[\frac{\|X_i - m_i\|^2}{c_n^{2+\delta}}\right] = 0,$$

se satisface la de Lindeberg (Ash & Doléans-Dade, 1999). Frecuentemente, es más sencillo verificar la condición más fuerte de Lyapunov para probar la convergencia de una suma a la gausiana.

Al final, se puede aún debilitar la condición de independencia sin perder la convergencia a la gausiana (Brockwell & Davis, 1987, Sec. 6.4).

1.10.2.3. Distribución normal o Gaussiana multivariada complejas

Por definición, un vector aleatorio complejo d-dimensional $Z = X + \imath Y$ es gausiano significa que el vector $2d$-dimensional $\widetilde{Z} = \begin{bmatrix} X^t & Y^t \end{bmatrix}^t$ es gausiano. Se puede entonces referirse en el caso de vectores gausianos, pero como lo presentamos en la sección 1.9.1, es frecuentemente más comodo trabajar con Z en lugar de $\widetilde{Z}$. En particular, en el marco de las comunicaciones en ingeneria, se trabaja con modulaciones dichas en fase y en cuadratura (señal multiplicado respectivamente por un seno y un coseno) y en lugar de trabajar con dos componente se considera una modulación con una exponencial compleja y la señal/variable compleja. Se puede por ejemplo referirse a (Lapidoth, 2017) (ver en particular el capitulo 24) o (Schreier & Scharf, 2003; Eriksson & Koivunen, 2006).

En el caso general, la gausiana real siendo completamente descritar por su media y su matriz de covarianza, la gausiana compleja va a ser completamente definida por la media, la matriz de covarianza y la pseudo-covarianza (ver Sec. 1.9 por las relaciones entre la covarianza de $\widetilde{Z}$ y estas matrices). $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$ con $m \in \mathbb{C}^d$, $\Sigma \in P_d^+(\mathbb{C})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{C})$ hermiticas definida positivas, y $\check{\Sigma} \in S_d(\mathbb{C})$ conjunto de las matrices de $\mathcal{M}_{d,d}(\mathbb{C})$ symmetricas (ver notaciones). Un caso particular aparece cuando Z es propio en torno de m, lo que es equivalente en el caso gausiano a tener Z circular (ver más adelante) en torno de 0, dado cuando $\check{\Sigma} = 0$: en este caso usaremos la misma notación, $Z \sim \mathcal{CN}(m, \Sigma)$. Las características de la gaussiana compleja son las siguientes (Lapidoth, 2017; Picinbono, 1996; Goodman, 1963; van den Bos, 1995; Schreier & Scharf, 2003; Eriksson & Koivunen, 2006):

Dominio de definición	$\mathcal{Z} = \mathbb{C}^d$
Parámetros	$m \in \mathbb{C}^d,\ \Sigma \in P_d^+(\mathbb{C}),\ \check{\Sigma} \in S_d(\mathbb{C})$
Densidad de probabilidad Caso general: Caso circular:	 $p_Z(z) = \frac{1}{\pi^d \lvert\Sigma\rvert^{\frac{1}{2}} \lvert P\rvert^{\frac{1}{2}}} e^{-(z-m)^\dagger P^{-1}(z-m) + \Re e\{(z-m)^t R^t P^{-1}(z-m)\}}$ con [50] $P = \Sigma - \check{\Sigma}\Sigma^{-*}\check{\Sigma}^\dagger, \quad R = \check{\Sigma}^\dagger \Sigma^{-1}$. $p_Z(z) = \frac{1}{\pi^d \lvert\Sigma\rvert}\, e^{-(z-m)^\dagger \Sigma^{-1}(z-m)}$
Promedio	$m_Z = m$
Covarianza	$\Sigma_Z = \Sigma$
Pseudo-covarianza	$\check{\Sigma}_Z = \check{\Sigma}$
Función característica Caso general: Caso circular:	 $\Phi_Z(\omega) = e^{-\frac{1}{4}\omega^\dagger \Sigma \omega - \frac{1}{4}\Re e\{\omega^\dagger \check{\Sigma} \omega^*\} + \imath \Re e\{\omega^\dagger m\}}, \quad \omega \in \mathbb{C}^d$ $\Phi_Z(\omega) = e^{-\frac{1}{4}\omega^\dagger \Sigma \omega + \imath \Re e\{\omega^\dagger m\}}, \quad \omega \in \mathbb{C}^d$

Notar que en el caso escalar propio (circular), la varianza de Z es $\sigma_Z^2 = 2\sigma^2$. El coefficiente 2 viene del hecho de que Z contiene dos componentes independientes de varianza σ^2.

Los vectores aleatorios complejos van a compartir las propiedades del caso real, siendo equivalente a un vector $2d$-dimensional gausiano real.

Primero, como en el caso real, la gausiana es estable por combinación lineal de vectores independientes:

Teorema 1-44 (Stabilidad)**.** *Sean $A_i, i = 1, \ldots, n$ matrices de $\mathbb{C}^{d' \times d}, d' \leq d$ de rango lleno, $b_i \in \mathbb{C}^{d'}$ y $Z_i \sim \mathcal{CN}(m_i, \Sigma_i, \check{\Sigma}_i)$ d-dimensionales, independientes, entonces*

$$\sum_{i=1}^n (A_i Z_i + b_i) \sim \mathcal{CN}\left(\sum_{i=1}^n (m_i + b_i)\ ,\ \sum_{i=1}^n A_i \Sigma_i A_i^\dagger\ ,\ \sum_{i=1}^n A_i \check{\Sigma}_i A_i^t\right)$$

vector gaussiano complejo da una gaussiana compleja. Reciprocamente, si cualquier combinación lineal de los componentes de un vector aleatorio sigue una ley gaussiana compleja, entonces el vector es gaussiano complejo.

---

[50] En (Picinbono, 1996) la expresión es ligieramente diferente, pero se recupera usando la simetría $\check{\Sigma}^* = \check{\Sigma}^\dagger$. Recordar que $\cdot^{-*} = (\cdot^*)^{-1}$ (ver notaciones).

Además, en el caso complejo se tiene una estabilidad combinando Z y Z^*:

Teorema 1-45. *Sean* $A \in \mathcal{M}_{d',d}(\mathbb{C})$, $B \in \mathcal{M}_{d',d}(\mathbb{C})$ *tales que ambas* $A+B$ *y* $A-B$ **sean de rango lleno**, $c \in \mathbb{C}^{d'}$ *y* $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$ d*-dimensonal, entonces*

$$AZ + BZ^* + c \sim \mathcal{CN}\left(\mu, C, \check{C}\right)$$

con

$$\begin{aligned} \mu &= Am + Bm^* + c \\ C &= A\Sigma A^\dagger + B\Sigma^* B^\dagger + A\check{\Sigma}B^\dagger + B\check{\Sigma}^* A^\dagger \\ \check{C} &= A\check{\Sigma}A^t + B\check{\Sigma}B^t + A\Sigma B^t + B\Sigma^* A^t \end{aligned}$$

Demostración. Tomando la forma real $2d$-dimensional Z es en biyección con $\widetilde{Z} = \begin{bmatrix} X \\ Y \end{bmatrix}$ y entonces Z^* en biyección con $\widetilde{Z^*} = \begin{bmatrix} X \\ -Y \end{bmatrix}$. Eso da $AZ + BZ^* + c$ en biyección con $\begin{bmatrix} A+B & 0 \\ 0 & A-B \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} \Re e\{c\} \\ \Im m\{c\} \end{bmatrix}$. Notando que $\begin{bmatrix} A+B & 0 \\ 0 & A-B \end{bmatrix}$ es de rango lleno, por el teorema 1-40 este vector es gausiano, lo que proba que $AZ + BZ^* + c$ es gausiano complejo. Las formas de la media, covarianza y pseudo-covarianza siguen de calculos directos. □

Evidentemente, se puede combinar los dos teoremas anteriores.

El teorema del límite central y sus variantes se recuperan del caso real.

Teorema 1-46 (Teorema del límite central (caso complejo))**.** *Sea* $\{Z_i\}_{i \in \mathbb{N}^*}$ *una serie de vectores aleatorios independientes, de misma ley, y que admiten un promedio* m*, una matriz de covarianza* Σ *y una matriz de pseudo-covarianza* $\check{\Sigma}$*. Entonces*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{m} (Z_i - m) \xrightarrow[n \to +\infty]{\mathrm{d}} W \sim \mathcal{CN}\left(0, \Sigma, \check{\Sigma}\right)$$

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución (ver notaciones).*

No lo presentamos, pero se transpone sencillamente el teorema de Lindenberg-Feller 1-43 al caso complejo.

Notamos también que, en el caso circular, se puede escribir naturalmente $Z \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}} N + m$ con $N \sim \mathcal{CN}(0, I)$ donde N es dicha *Gausiana estandar* o *centrada-normalizada*. Eso se generaliza en dos direcciones. La primera pone también en juega una gausiana estandar (Lapidoth, 2017):

Teorema 1-47. *Sea* $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$*. Entonces, existen matrices (no únicas)* $A \in \mathcal{M}_{d,d}(\mathbb{C})$, $B \in \mathcal{M}_{d,d}(\mathbb{C})$ *tales que*

$$Z \overset{\mathrm{d}}{=} AW + BW^* + m$$

con $W \sim \mathcal{CN}(0, I)$ *gausiana estandar.*

Demostración. Inmediatamente

$$Z = \begin{bmatrix} I & \imath I \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \overset{\mathrm{d}}{=} \begin{bmatrix} I & \imath I \end{bmatrix} M \begin{bmatrix} U \\ V \end{bmatrix}$$

con $U \sim \mathcal{N}(0, I)$ y $V \sim \mathcal{N}(0, I)$ independientes, y M tal que $MM^t = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^t & \Sigma_Y \end{bmatrix}$ (ej. raiz cuadrade de esta matriz de $P_{2d}^+(\mathbb{R})$, o descomposición de Cholesky (Horn & Johnson, 2013; Bhatia, 2007)). Ahora, volviendo a la forma compleja tenemos

$$Z \overset{\mathrm{d}}{=} \begin{bmatrix} I & \imath I \end{bmatrix} M \begin{bmatrix} I & I \\ -\imath I & \imath I \end{bmatrix} \begin{bmatrix} \frac{1}{2}(U + \imath V) \\ \frac{1}{2}(U - \imath V) \end{bmatrix}$$

Se cierra la prueba denotando

$$\begin{bmatrix} A & B \end{bmatrix} = \begin{bmatrix} I & \imath I \end{bmatrix} M \begin{bmatrix} I & I \\ -\imath I & \imath I \end{bmatrix}$$

y notando que $\frac{1}{2}(U + \imath V) \sim \mathcal{CN}(0, I)$. □

Nota, usando la descomposición de Cholesky, tenemos M triangular superior [51], y entonces bloc-triangular lo que conduce, por identificación de los bloques de $MM^t = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^t & \Sigma_Y \end{bmatrix}$, a una solución posible con

$$\begin{cases} A &= \left(\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^t\right)^{\frac{1}{2}} + \Sigma_Y^{\frac{1}{2}} - \imath\Sigma_{XY}\Sigma_Y^{-\frac{1}{2}} \\ B &= \left(\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^t\right)^{\frac{1}{2}} - \Sigma_Y^{\frac{1}{2}} + \imath\Sigma_{XY}\Sigma_Y^{-\frac{1}{2}} \end{cases}$$

La secunda extensión pone en juega una sola gausiana compleja sin su conjugada (Eriksson & Koivunen, 2006; Schreier & Scharf, 2003):

Teorema 1-48. *Sea $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$. Entonces, existe una matriz $C \in \mathcal{M}_{d,d}(\mathbb{C})$ tal que*

$$Z \overset{\mathrm{d}}{=} CW + m$$

con $W \sim \mathcal{CN}(0, I, \Delta)$ con $\Delta \in P_d(\mathbb{R})$ (real) diagonal.

Demostración. Eso viene de teoremas de diagonalización conjunta. Más precisamente, siendo $\Sigma \in P_d^+(\mathbb{C})$ y $\check{\Sigma} \in S_d(\mathbb{C})$, se aplica el teorema (Horn & Johnson, 2013, Teo. 7.6.5) diciendo que existe una matriz no singular (invertible) C tal que $\Sigma = CC^\dagger$ y $\check{\Sigma} = C\Delta C^t$ con Δ real diagonal con elementos positivos ($\Delta \in P_d(\mathbb{R})$ diagonal). Inmediatamente, por el teorema 1-44, tenemos

$$C^{-1}(Z - m) \overset{\mathrm{d}}{=} W \sim \mathcal{CN}(0, I, \Delta)$$

lo que cierra la prueba. □

[51] Se puede hacer el mismo razonamiento con la forma triangular inferior; se cambia los roles de X e Y en las matrices de covarianza.

Al final, vímos en la sección 1.9.1 que si un vector es circular, entonces su pseudo-covarianza es nula, pero la reciproca no vale en general. Aparece que en el contexto gausiano tenemos la reciproca:

Teorema 1-49 (Circularidad)**.** *Sea* $Z \sim \mathcal{CN}(m, \Sigma, \check{\Sigma})$*. Entonces,*

$$Z \text{ circular en torno de } m \iff Z \text{ propio en torno de } m$$

Demostración. Vímos la directa en la sección 1.9.1, teorema 1-37. Reciprocamente, si Z es propio en torno de m, por definición $\check{\Sigma} = 0$ y el resultado viene de la forma de la función característica por ejemplo: $\Phi_{Z-m}(\omega) = e^{-\frac{1}{4}\omega^\dagger \Sigma \omega} = \Phi_{Z-m}\left(e^{\imath\theta}\omega\right) = \Phi_{e^{\imath\theta}(Z-m)}(\omega)$. □

1.10.2.4. Distribución exponencial

A pesar de que sea un caso particular de la distribución Gamma que vamos a ver más adelante, estudiada por Pearson desde el año 1895 (Pearson, 1895), o apareció quizas un poco antes en trabajos de L. Boltzmann o de Whitworth (Balakrishnan & Basu, 1995) (como caso limíte de la ley de Poisson), apareció esta ley de manera "propia" mucho más tarde, entre otros en 1930 en (Kondo, 1930, Ec. (46)).

Se denota $X \sim \mathcal{E}(\lambda)$ con $\lambda \in \mathbb{R}_+^*$ llamada *taza* (inversa de *escala*), y sus características son las siguientes:

Dominio de definición	$\mathcal{X} = \mathbb{R}_+$
Parámetro	$\lambda \in \mathbb{R}_+^*$
Densidad de probabilidad	$p_X(x) = \lambda e^{-\lambda x}$
Promedio	$m_X = \dfrac{1}{\lambda}$
Varianza	$\sigma_X^2 = \dfrac{1}{\lambda^2}$
Sesgo	$\gamma_X = 2$
Curtosis por exceso	$\bar{\kappa}_X = 6$
Generadora de momentos	$M_X(u) = \dfrac{\lambda}{\lambda - u}$ para $\Re e\{u\} < \lambda$
Función característica	$\Phi_X(\omega) = \dfrac{\lambda}{\lambda - \imath\omega}$

Su densidad de probabilidad y función de repartición son representadas en la figura Fig. 1-27. **Poner escalas; Otros ilustraciones para otros λ?**

La ley exponencial es conocida como siendo *sin memoria*, es decir, si buscamos X visto como un

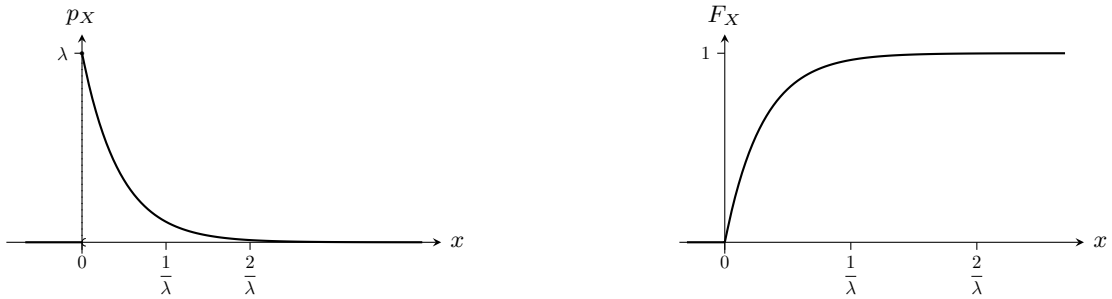


Figura 1-27: Ilustración de una densidad de probabilidad exponencial (a), y la función de repartición asociada (b), con $\lambda = 1{,}5$.

tiempo (ej. tiempo de desintegración de un atomo radioactivo) tal que

$$\forall\, x_0 \geq 0,\, x \geq 0, \quad P(X > x + x_0 | X > x_0) = P(X > x)$$

i. e., la probabilidad que $X > x + x_0$ (extra tiempo después de x_0) condicionalmente a $X > x_0$ es exactamente la de $X > x + x_0$ (se olvidó x_0), tenemos, por la definición de la probabilidad condicional

$$\forall\, x_0 \geq 0,\, x \geq 0, \quad \frac{1 - F_X(x + x_0)}{1 - F_X(x_0)} = 1 - F_X(x)$$

Por diferenciación con respeto a x eso da, en $x \to 0$,

$$\forall\, x_0 \geq 0, \quad F'_X(x_0) + \lambda F_X(x_0) = \lambda \qquad \text{con} \qquad \lambda = F'_X(0)$$

Teneiendo en cuenta de que F_X es una función de repartición, aparece que $F_X(x) = \left(1 - e^{-\lambda x}\right) \mathbb{1}_{\mathbb{R}_+}(x)$, ley exponencial.

Como lo hemos evocado tratando de la ley de Poisson, esta es vínculada intimamente a la ley exponencial a través del processo dicho de poisson:

Lema 1-27 (Vínculo con la ley de Poisson)**.** *Sea* $T_0 = 0$ *y* $\forall\, n \in \mathbb{N}^*$ *las variables aleatorias positivas* T_n *tales que* $T_{n+1} - T_n \geq 0$ *son independientes y de distribución* $\mathcal{E}(\lambda)$*. Fijamos* $T > 0$ *y sea* X *el número de variables* T_n *que partenecen a* $(0\,;\, T)$*, i. e.,* $T_X < T < T_{X+1}$*. Entonces*

$$X \sim \mathcal{P}(\lambda T)$$

Dicho de otra manera, si tenemos eventos que aparecen en tiempos aleatorios tales que los incrementos de tiempos entre eventos son independientes y de distribución exponencial de taza λ, el número de eventos en un intervalo de tiempo T dado sigue una ley de Poisson, de taza λT proporcional al intervalo, y proporcional a la taza de la ley exponencial. El parámetro λ representa la taza de evento por unidad de tiempo.

Demostración. Por definición,

$$\begin{aligned} P(X=n) &= P(X \leq n) - P(X \leq n-1) \\ &= P(T_{n+1} > T) - P(T_n > T) \\ &= F_{T_n}(T) - F_{T_{n+1}}(T) \end{aligned}$$

Ahora, notando que

$$T_n = \sum_{i=0}^{n-1} (T_{i+1} - T_i)$$

de la independencia de los incrementos de tiempo, y de las propiedades de la función característica, tenemos

$$\Phi_{T_n}(\omega) = \frac{\lambda^n}{(1 - \imath\,\omega)^n}$$

De la fórmula de inversion del teorema 1-33 se prueba que [52]

$$p_{T_n}(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} \mathbb{1}_{\mathbb{R}_+}(x)$$

Con integraciones por partes, se obtiene sencillamente

$$F_{T_n}(T) = 1 - \sum_{i=0}^{n-1} \frac{\lambda^i T^i e^{-\lambda T}}{i!}$$

lo que cierra la prueba. □

En física, se modela la ley de tiempo de desintegración como siendo exponencial, y se supone que los desintegraciuones son independientes, explicando el modelo de Poisson para el número de desintegración durante un tiempo dado.

Una otra característica de esta ley es su stabilidad con respecto al operador no lineal mínimo:

Lema 1-28 (Stabilidad por el mín)**.** *Sean* $X_i \sim \mathcal{E}(\lambda),\ i = 1, \ldots, n$ *independientes. Entonces,*

$$\min_{i=1,\ldots,n} X_i \equiv X \sim \mathcal{E}(n\lambda)$$

Demostración. Inmediatamente, para cualquief $x \geq 0$

$$\begin{aligned} 1 - F_X(x) &= P(X > x) \\ &= P\left(\bigcap_{i=1}^{n} (X_i > x)\right) \\ &= \prod_{i=1}^{n} P(X_i > x) \\ &= e^{-n\lambda x} \end{aligned}$$

[52] Una manera es de hacer una integración en el plano complejo y usar los lemas de Jordan y teorema de residuos (Carrier, Krook & Pearson, 2005) o (Ablowitz & Fokas, 2003, Cap. 4). Nota: de hecho se reconoce en Φ_{T_n} la función característica de una ley gamma $\mathcal{G}(n, \lambda)$, ley que vamos a ver en la sección 1.10.2.5.

La secunda linea viene de la equivalencia entre los eventos $\min_{i=1,\ldots,n} X_i > x$ y $\bigcap_{i=1}^{n} (X_i > x)$ y la tercera de la independencia de los X_i. □

1.10.2.5. Distribución Gamma

Pearson 1895 Se denota $X \sim \mathcal{G}(a,b)$ con $a \in \mathbb{R}_+^*$ llamado *parámetro de forma* y $b \in \mathbb{R}_+^*$ llamada *taza* (inversa de *escala*). **Cuando $a \in \mathbb{N}^*$ es entero, la ley es a veces conocida como ley de Erlang () A. K. Erlang, a Danish telephone engineer who is considered the founder of queueing theory Cf. Cox** Las características son:

Dominio de definición	$\mathcal{X} = \mathbb{R}_+$
Parámetros	$a \in \mathbb{R}_+^*$ (forma), $b \in \mathbb{R}_+^*$ (taza)
Densidad de probabilidad	$p_X(x) = \dfrac{b^a\, x^{a-1}\, e^{-bx}}{\Gamma(a)}$
Promedio	$m_X = \dfrac{a}{b}$
Varianza	$\sigma_X^2 = \dfrac{a}{b^2}$
Sesgo	$\gamma_X = \dfrac{2}{\sqrt{a}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6}{a}$
Generadora de momentos	$M_X(u) = \left(1 - \dfrac{u}{b}\right)^{-a}$ para $\Re e\{u\} < b$
Función característica	$\Phi_X(\omega) = \left(1 - \dfrac{\imath\omega}{b}\right)^{-a}$

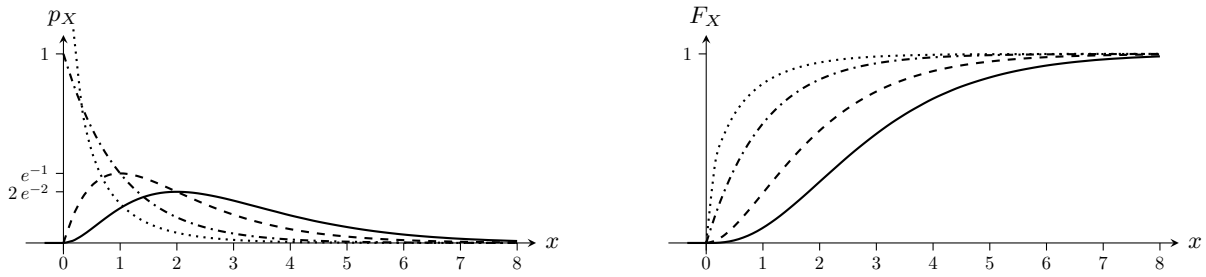
Nota: trivialmente, se puede escribir $X \overset{\mathrm{d}}{=} \frac{1}{b}G$ con $G \sim \mathcal{G}(a,1)$ donde G es estandardizada o normalizada. De nuevo, las características de X son vínculadas a las de G (y vice-versa) por transformación afine (ver secciones anteriores).

Una densidad de probabilidad gamma y la función de repartición asociada son representadas en la figura Fig. 1-28 para varios a y $b = 1$.

Nota que para $X \sim \mathcal{G}(1,b)$ es una variable exponencial de parámetro b, *i. e.,* $X \sim \mathcal{E}(b)$. Cuando $a < 1$, la densidad p_X diverge para $x \to 0$ (divergencia integrable). Además, se muestra también sencillamente con las funciones características que:

Lema 1-29 (Stabilidad)**.** *Sean* $X_i \sim \mathcal{G}(a_i, b)$, $i = 1, \ldots, n$ *independientes. Entonces*

$$\sum_{i=1}^{n} X_i \sim \mathcal{G}\left(\sum_{i=1}^{n} a_i\,,\, b\right)$$



(a) (b)

Figura 1-28: Ilustración de una densidad de probabilidad gamma (a), y la función de repartición asociada (b). $b = 1$ y $a = 0{,}5$ (linea punteada), 1 (linea mixta), 2 (linea guionada) y 3 (linea llena).

En particular, la suma de variable independientes de ley exponencial sigue una distribución de Erlang de parámetro de forma n.

Además, se muestra sencillamente por cambio de variables y la función característica un vínculo con variables gausianas:

Lema 1-30 (Vínculo con la gaussiana)**.** *Sean* $X_i \sim \mathcal{N}\left(0, \sigma^2\right),\ i = 1, \ldots, n$ *independientes. Entonces*

$$\sum_{i=1}^{n} X_i^2 \sim \mathcal{G}\left(\frac{n}{2}, \frac{1}{2\,\sigma^2}\right)$$

La distribución Gamma aparece entre otros en problema de inferencia Bayesiana como distribución a priori conjugado [53] del parámetro λ de la ley de Poisson (Robert, 2007).

Esta distribución aparece...

1.10.2.6. Distribución matriz-variada de Wishart

Este ejemplo es una generalización matriz-variada de la distribución gamma. Se puede ver una matriz como un vector, guardando por ejemplo sus columnas una bajo la precediente. Sin embargo, tal distribución apareciendo naturalmente en un contexto de estimación de matriz de covarianza (ver más adelante), es más natural verla matriz-variate. Tal distrubución es debido a J. Wishart (Wishart, 1928;

[53] En la inferencia Bayesiana, nos interesamos al paremetro (posiblemente multivariado) θ subyancente a una distribución. Por ejemplo, sabemos tener observaciones sorteados de una distribución de Poisson, pero con el parámetro λ desconocido y nos interesamos a $\theta \equiv \lambda$. El enfoque Bayesiano consiste a considerar el paremetro Θ aleatorio, tal que la distribución de las observaciones sea vista como distribución condicional $p_{X|\Theta=\theta}(x)$, llamada distribución de sampleo. Dados las observaciones $X = x$, la meta es de determinar la distribución dicha a posteriori $p_{\Theta|X=x}$ a partir de la cual se puede hacer estimación de θ dados las observaciones, calcular intervalos de confianza, etc. Por eso, el metodo se apoya sobre la regla de Bayes $p_{\Theta|X=x}(\theta) \propto p_{X|\Theta=\theta}(x) p_\Theta(\theta)$ así que se necesita elegir una distribución p_Θ dicha a priori. Una elección posible es tomarla en una familia parametrizada tal que la distribución a posterior partenece también a esta familia: es lo que se llama prior conjugado. La idea es que si vienen observaciones, en lugar de re-calcular el posterior, se puede actualizar solamente los parámetros (llamados hiperparámetros). **Ver nota de pie en el cap 2 a modificar**.

Gupta & Nagar, 1999; Anderson, 2003), y se denota $X \sim W_d(V, \nu)$ donde el dominio de definición es $P_d^+(\mathbb{R})$, conjunto matrices simetricas definida positivas, $V \in P_d^+(\mathbb{R})$ parámetro de escala y $\nu > d-1$ grados de libertad. Las características de la distribución son las siguientes:

Dominio de definición [54]	$\mathcal{X} = P_d^+(\mathbb{R}),\ d \in \mathbb{N}^*$
Parámetros	$V \in P_d^+(\mathbb{R})$ (escala) y $\nu > d-1$ (grados de libertad)
Densidad de probabilidad [55]	$p_X(x) = \dfrac{\lvert x\rvert^{\frac{\nu-d-1}{2}}\, e^{-\frac{1}{2}\operatorname{Tr}\left(V^{-1}x\right)}}{2^{\frac{d\nu}{2}}\, \lvert V\rvert^{\frac{\nu}{2}}\, \Gamma_d\left(\frac{\nu}{2}\right)}$
Promedio	$m_X = \nu V$
Covarianza	$\Sigma_X = \nu\big(J(V\otimes V) + (V\otimes I)K(V\otimes I)\big)$
Función característica [56]	$\Phi_X(\omega) = \lvert I - 2i\omega V\rvert^{-\frac{\nu}{2}},\quad \omega \in S_d(\mathbb{R})$

(ver (Peddada & Richards, 1991; Sultan & Tracy, 1996; Anderson, 2003)).

Fijense que p_X no es la distribución conjuntos de los componentes de X: el hecho de que X sea uan matriz aleatoria de $P_d^+(\mathbb{R})$ impone vínculos sobre sus compnentes; entre otros, $X_{i,j} = X_{j,i}$.

Inmediatamente, si $d = 1$, la distribución de Whishart $W_1(V, \nu)$ se reduce a la distribución Gamma $\mathcal{G}\left(\frac{\nu}{2}, \frac{1}{2V}\right)$. De este hecho, se la podría ver como extensión matriz-variada de la distribución gamma. La distribución de Wishart tiene varias otras propiedades como las siguientes.

Lema 1-31 (Stabilidad por transformación lineal)**.** *Sea $X \sim W_d(V, \nu)$ y $A \in \mathbb{R}^{d\times d'}$ con $d' \leq d$ y de rango lleno. Entonces*

$$A^t X A \sim W_{d'}\left(A^t V A, \nu\right)$$

En particular, si $d' = 1$, $A^t X A \sim G\left(\frac{\nu}{2}, \frac{1}{2\, A^t V A}\right)$. Más allá, tomando $A = \mathbb{1}_j$, aparece de que las componentes diagonales de X son de distribución gamma, $X_{j,j} \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{1}{2\, V_{j,j}}\right)$.

---

[54] De hecho, se puede considerar que la matriz aleatoria es equivalent a tener un vector $\frac{d(d+1)}{2}$-dimensional; por la simetria, claramente X tiene solamente $\frac{d(d+1)}{2}$ componentes diferentes; además, se puede probar que cualquier matriz $A \in P_d^+(\mathbb{R})$ se descompone bajo la forma $A = LL^t$ con L triangular inferior con elementos no nulos sobre su diagonal, llamado descomposición de Cholesky (Gupta & Nagar, 1999; Bhatia, 2007; Harville, 2008; Horn & Johnson, 2013) y reciprocamente. Eso muestra que A se define a partir de $\frac{d(d+1)}{2}$ "grados de libertad".

[55] La densidad de probabilidad corresponde a la densidad conjunta de los $\frac{d(d+1)}{2}$ elementos $X_{i,j}$, $1 \leq i \leq j \leq d$ (Wishart, 1928; Peddada & Richards, 1991; Sultan & Tracy, 1996; Gupta & Nagar, 1999; Anderson, 2003).

[56] **Se proba que la función generadora de momentos no existe en general**.

Demostración. El resultado es inmediato saliendo de la función caractéristica [56] y notando de que

$$\begin{aligned}\Phi_{A^tXA}(\omega) &= \mathrm{E}\left[e^{\imath\,\mathrm{Tr}\left(\omega^tA^tXA\right)}\right]\\ &= \Phi_X\left(A\omega^tA^t\right)\\ &= \left|I-2\imath A\omega A^tV\right|^{-\frac{\nu}{2}}\\ &= \left|I-2\imath\omega A^tVA\right|^{-\frac{\nu}{2}}\end{aligned}$$

de $\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$ (Harville, 2008) y de la identidad de Sylvester (Sylvester, 1851; Akritas, Akritas & Malaschonok, 1996) o (Harville, 2008, § 18.1) $|I+AB|=|I+BA|$. . □

De hecho, si los elementos diagonales son de distribución gamma, no es el caso de los elementos no-diagonales (Seber, 2004; Anderson, 2003) o (Gupta & Nagar, 1999, Teo. 3.3.4). De eso resuelte delicado llamar la distribución como gamma matriz-variada.

Lema 1-32 (Stabilidad por suma)**.** *Sea* $X_i\sim W_d(V,\nu_i)$, $i=1,\ldots,n$ *independientes. Entonces*

$$\sum_{i=1}^n X_i\sim W_d\left(V\,,\sum_{i=1}^n\nu_i\right)$$

Demostración. El resultado es inmediato saliendo de la función caractéristica [56] y notando que como el el context vectorial $\Phi_{\sum_i X_i}=\prod_i\Phi_{X_i}$. □

La distribución de Wishart aparece naturalmente en problemas de estimación de matriz de covarianza en el contexto gausiano:

Lema 1-33 (Vínculo con vectores gausianos (Seber, 2004))**.** *Sean* $X_i\sim\mathcal{N}(0,V)$, $i=1,\ldots,n>d-1$ *independientes y la matriz* $S=\sum_{i=1}^n X_iX_i^t$ *llamada* matriz de dispersión *(scatter matrix en inglés). Entonces,* $S\in P_d^+(\mathbb{R})$ *(c. s.) (*S *es simétrica definida positiva casi siempre, o con probabilidad uno) y* $S\sim W_d(V,n)$.

Este resultado permite también probar el lema **??** para $\nu=n$ entero escribiendo $X\overset{\mathrm{d}}{=}\sum_{i=1}^n X_iX_i^t$ tal que $A^tXA\overset{\mathrm{d}}{=}\sum_{i=1}^n A^tX_iX_i^tA=\frac{1}{n}\sum_{i=1}^n\left(A^tX_i\right)\left(A^tX_i\right)^t$ y notando que los $A^tX_i\sim\mathcal{N}(0,A^tVA)$ son independientes (Seber, 2004). Además, permite re-obtener las expreciones del promedio y de las covarianzas [57]. Notar que cuando los X_i tienen un promemedio, el lema conduce a lo que es conocido como Wishart no central (Anderson, 2003; Seber, 2004).

Que propedad mas? Ver Gupta Nagar 1999

La distribución Wishart aparece así naturalmente en problema de inferencia Bayesiana como distribución a priori conjugado [53] del parámetro p de la ley gaussiana multivariada (Robert, 2007).

---

[57] Para la covarianza, su usa la formula $\mathrm{E}[Y_1Y_2Y_3Y_4]=\mathrm{E}[Y_1Y_2]\,\mathrm{E}[Y_3Y_4]+\mathrm{E}[Y_1Y_3]\,\mathrm{E}[Y_2Y_4]+\mathrm{E}[Y_1Y_4]\,\mathrm{E}[Y_2Y_3]$ para $Y=\begin{bmatrix}Y_1 & Y_2 & Y_3 & Y_4\end{bmatrix}^t$ vector gausiano, formula que se obtiene por ejemplo a partir de la función característica de un vecor gausiano.

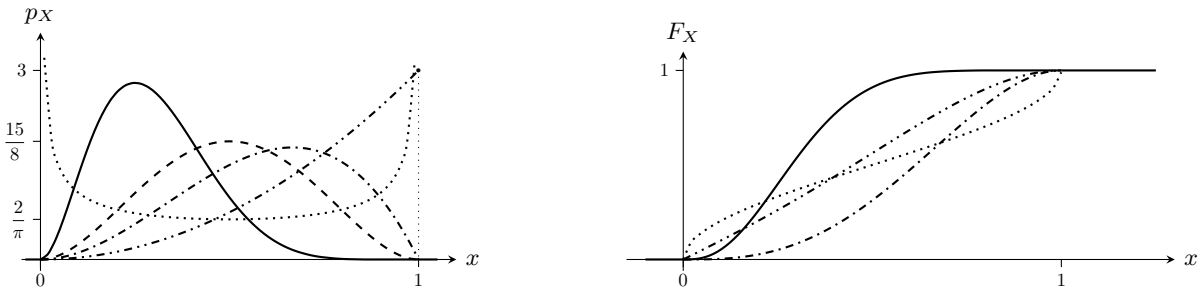
Y donde aparece mas?

1.10.2.7. Distribución Beta

Se denota $X \sim \beta(a,b)$ con $(a,b) \in \mathbb{R}_+^{*\,2}$ llamados *parámetros de forma*. Las características son:

Dominio de definición	$\mathcal{X} = [0\,;\,1]$
Parámetros	$(a,b) \in \mathbb{R}_+^{*\,2}$ (forma)
Densidad de probabilidad	$p_X(x) = \dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$
Promedio	$m_X = \dfrac{a}{a+b}$
Varianza	$\sigma_X^2 = \dfrac{ab}{(a+b)^2(a+b+1)}$
Sesgo	$\gamma_X = \dfrac{2\,(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6\,\big((a-b)^2(a+b+1) - ab(a+b+2)\big)}{a\,b\,(a+b+2)\,(a+b+3)}$
Generadora de momentos	$M_X(u) = {}_1F_1\,(a, a+b\,;\,u)$ para $u \in \mathbb{C}$
Función característica	$\Phi_X(\omega) = {}_1F_1\,(a, a+b\,;\,\imath\omega)$

Su densidad de probabilidad y función de repartición son representadas en la figura Fig. 1-29 para varios a y b.



(a) (b)

Figura 1-29: Ilustración de una densidad de probabilidad beta (a), y la función de repartición asociada (b). $(a,b) = (0{,}5\,,\,0{,}5)$ (linea punteada), $(3\,,\,1)$ (linea mixta doble punteada), $(3\,,\,2)$ (linea mixta), $(3\,,\,3)$ (linea guionada), $(3\,,\,7)$ (linea llena).

Nota que se recupera la ley uniforme sobre $[0\,;\,1]$ para $a = b = 1$. Se conoce la ley de $2\,\beta\left(\frac{1}{2}, \frac{1}{2}\right) - 1$ como *ley arcseno*.

Variables beta tienen también unas propiedades notables. Primero, por cambio de variables, se demuestra el lema siguiente:

Lema 1-34 (Reflexividad)**.** *Sea* $X \sim \beta(a,b)$*. Entonces*

$$1 - X \sim \beta(b,a)$$

Lema 1-35 (Vínculo con la ley gamma)**.** *Sea* $X \sim \mathcal{G}(a,c)$ *e* $Y \sim \mathcal{G}(b,c)$ *independientes. Entonces*

$$\frac{X}{X+Y} \sim \beta(a,b)$$

(independientemente de c*). Además,* $\frac{X}{X+Y}$ *y* $X+Y$ *son independientes.*

Demostración. La independencia de c es obvia del hecho de que para cualquier $\theta > 0$, $\theta^{-1}X \sim \mathcal{G}(a,\theta c)$ e $\theta^{-1}Y \sim \mathcal{G}(b,\theta c)$, la independencia con respeto a c viniendo de $\frac{\theta^{-1}X}{\theta^{-1}X+\theta^{-1}Y} = \frac{X}{X+Y}$. Entonces, se puede considerar $c = 1$ sin perdida de generalidad. Ahora, sea la transformación

$$\begin{array}{rccl} g: & \mathbb{R}_+^2 & \mapsto & [0\,;\,1]\times\mathbb{R}_+ \\ & (x,y) & \to & (u,v) = \left(\frac{x}{x+y}\,,\,x+y\right) \end{array}$$

Entonces, la transformación inversa se escribe

$$g^{-1}(u,v) = (uv\,,\,(1-u)v)$$

de matriz Jacobiana

$$\mathrm{J}_{g^{-1}} = \begin{bmatrix} v & u \\ -v & 1-u \end{bmatrix}$$

Del teorema de cambio de variables teorema **??**, notando que $\left|\mathrm{J}_{g^{-1}}\right| = v$ y de la independencia de X e Y, se obtiene para el vector aleatorio $W = \begin{bmatrix} U & V \end{bmatrix}^t$ la densidad de probabilidad

$$\begin{aligned} p_W(u,v) &= p_X(uv)\,p_Y((1-u)v)\,v \\ &= \frac{(uv)^{a-1}\,e^{-uv}}{\Gamma(a)} \times \frac{((1-u)v)^{b-1}\,e^{-(1-u)v}}{\Gamma(b)} \times v \\ &= \frac{u^{a-1}(1-u)^{b-1}}{B(a,b)} \times \frac{v^{a+b-1}e^{-v}}{\Gamma(a+b)} \end{aligned}$$

La densidad se obtiene por marginalización, *i. e.,* integrando sobre v la densidad conjunta, lo que cierra la prueba de la primera parte. Además, aparece claramente que U y V son independientes (se factorisa la densidad de probabilidad). Pasando, se recupera el hecho que $X+Y \sim \mathcal{G}(a+b,1)$. □

Lema 1-36 (Stabilidad por producto)**.** *Sea* $X \sim \beta(a,b)$ *e* $Y \sim \beta(a+b,c)$ *independientes. Entonces*

$$XY \sim \beta(a,b+c)$$

Demostración. Sean $U \sim \mathcal{G}(a,1)$, $V \sim \mathcal{G}(b,1)$ y $W \sim \mathcal{G}(c,1)$ independientes y sean $X = \frac{U}{U+V}$, $Y = \frac{U+V}{U+V+V}$ y $Z = U+V+W$. Del lema anterior $X \sim \beta(a,b)$ y $Y \sim \beta(a+b,c)$. Sea la transformación

$$\begin{array}{rccl} g: & \mathbb{R}_+^3 & \mapsto & [0\,;\,1]^2 \times \mathbb{R}_+ \\ & (u,v,w) & \to & (x,y,z) = \left(\frac{u}{u+v}\,,\,\frac{u+v}{u+v+w}\,,\,u+v+w\right) \end{array}$$

Entonces, la transformación inversa se escribe

$$g^{-1}(x,y,z) = (xyz\,,\,(1-x)yz\,,\,z(1-y))$$

de matriz Jacobiana

$$\mathrm{J}_{g^{-1}} = \begin{bmatrix} yz & xz & xy \\ -yz & (1-x)z & (1-x)y \\ 0 & -z & 1-y \end{bmatrix}$$

De nuevo, del teorema de cambio de variables teorema **??**, notando que $\left|\mathrm{J}_{g^{-1}}\right| = yz^2$ y de la independencia de U,V,W, se obtiene para el vector aleatorio $T = \begin{bmatrix} X & Y & Z \end{bmatrix}^t$ la densidad de probabilidad probabilidad

$$\begin{aligned} p_T(x,y,z) &= p_u(xyz)\,p_V((1-x)yz)\,p_W(y(1-z))\,yz^2 \\ &= \frac{(xyz)^{a-1}\,e^{-xyz}}{\Gamma(a)} \times \frac{((1-x)yz)^{b-1}\,e^{-(1-x)yz}}{\Gamma(b)} \times \frac{(z(1-y))^{c-1}\,e^{-z(1-y)}}{\Gamma(c)} \times yz^2 \\ &= \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \times \frac{y^{a+b-1}(1-y)^{c-1}}{B(a+b,c)} \times \frac{z^{a+b+c-1}e^{-z}}{\Gamma(a+b+c)} \end{aligned}$$

Eso proba que X, Y y Z son independientes (las densidades se factorizan). Además,

$$XY = \frac{U}{U+V} \times \frac{U+V}{U+V+W} = \frac{U}{U+V+W} \sim \beta(a,b+c)$$

el último resultado como consecuencia de los lemas 1-35 y 1-29. Eso cierra la prueba. □

Exponencial como limite de $nbeta(1,n)$. $e^{-X} \sim beta(l,1)$

La distribución beta aparece entre otros en problema de inferencia Bayesiana como distribución a priori conjugado [53] del parámetro p de la ley Binomial (Robert, 2007).

Esta distribución aparece...

La distribución beta se generaliza al caso matriz-variada X definido sobre $\mathcal{X}$ tal que X y $I-X$ son en $P_d^+(\mathbb{R})$; se denota $X \sim \beta_d(a,b)$ donde $(a,b) \in \mathbb{R}_+^{*\,2}$ y la densidad est dada por $p_X(x) = \frac{|x|^{a-\frac{d+1}{2}}|I-x|^{b-\frac{d+1}{2}}|}{B_p([a \quad b]^t)}$, $(a,b) \in \left(\frac{d-1}{2}\,;\,+\infty\right)^2$. Se refiera a (Gupta & Nagar, 1999, Cap. 5) para tener más informaciones.

1.10.2.8. Distribución de Dirichlet

Esta distribución teniendo su nombre de integrales on a simplex estudiados por M. Lejeune-Dirichlet y J. Liouville en 1839 (Gupta & Richards, 2001; Lejeune-Dirichlet, 1839; Liouville, 1839) en es una

extensión multivariada de las variables beta a veces conocida como *Beta multivariada* (Olkin & Rubin, 1964). Se nota $X \sim \mathcal{D}ir(a)$ con $a \in \mathbb{R}_+^{*\,k}$ y X vive sobre el $(k-1)$-simplex estandar Δ_{k-1}. a es llamado parámetro de forma. Como en el caso de vectores de distribución multinomial, a pesar de que se escribe X de manera k-dimensional, el vector partenece a una variedad $d = k-1$ dimensional y en el caso $k = 2$ se recupera la ley beta. A veces se parametriza la ley con un parámetro escalar $\alpha > 0$ y un vector del simplex estandar $\bar{a} \in \Delta_{k-1}$ tal que

$$a = \alpha\bar{a}, \quad \textit{i. e.,} \quad \alpha = \sum_{i=1}^{k} \alpha_i, \quad \bar{a} = \frac{a}{\alpha}$$

α es conocido como parámetro de *concentración* y el vector $\bar{a}$ como *medida de base*.

Las características de un vector de Dirichlet son:

Dominio de definición [58]	$\mathcal{X} = \Delta_{k-1},\ k \in \mathbb{N} \setminus \{0\,;\,1\}$
Parámetros	$a = \alpha\,\bar{a} \in \mathbb{R}_+^{*\,k}$ (forma) con $\alpha \in \mathbb{R}_+^*$ (concentración) y $\bar{a} \in \Delta_{k-1}$ (medida de base)
Densidad de probabilidad [59]	$p_X(x) = \dfrac{\prod_{i=1}^{k} x_i^{a_i - 1}}{B(a)}$
Promedio	$m_X = \bar{a}$
Covarianza [60]	$\Sigma_X = \dfrac{\operatorname{diag}(\bar{a}) - \bar{a}\bar{a}^t}{1+\alpha}$
Generadora de momentos [61]	$M_X(u) = \Phi_2^{(k)}(a, \alpha\,;\,u)$ para $u \in \mathbb{C}$
Función característica [62]	$\Phi_X(\omega) = \Phi_2^{(k)}(a, \alpha\,;\,\imath\omega)$

La figura Fig. 1-30 representa el dominio de definición del vector (a) y su densidad de probabildad con las marginales (ver más adelante) para $k = 3$.

---

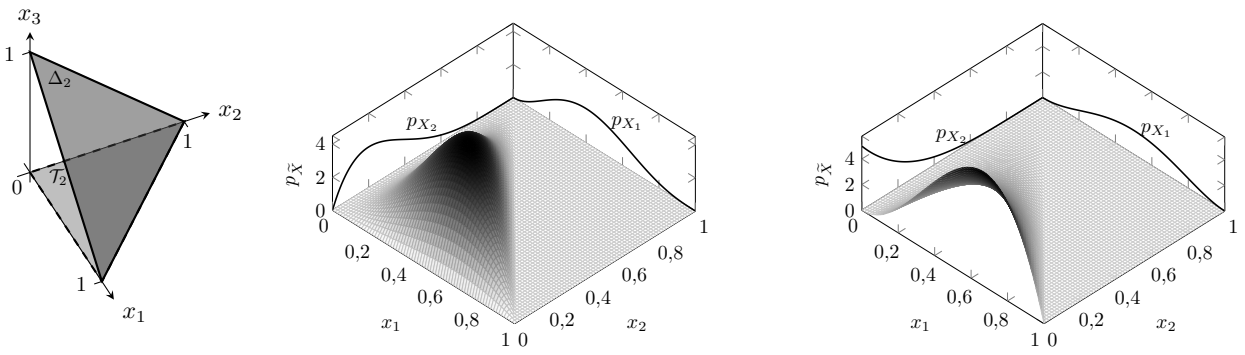
[58] De hecho, se puede considerar que el vector aleatorio es $(k-1)$-dimensional $\widetilde{X} = \begin{bmatrix} \widetilde{X}_1 & \cdots & \widetilde{X}_{k-1} \end{bmatrix}^t$ definido sobre el hipertriangulo $\widetilde{\mathcal{X}} = \mathcal{T}_{k-1} = \left\{ \widetilde{x} \in [0\,;\,1]^{k-1} \;\middle|\; \sum_{i=1}^{k-1} \widetilde{x}_i \leq 1 \right\}$, proyección del simplex sobre el hiperplano $x_k = 0$.

[59] La densidad de probabilidad es dada con respeto a la medida de Lebesgue restricta al simplex Δ_{k-1}. Tratando de $\widetilde{X}$, el vector tiene una densidad con respeto a la medida de Lebesgue usual, y es dada por $p_{\widetilde{X}}(\widetilde{x}) = \dfrac{\prod_{i=1}^{k-1} \widetilde{x}_i^{\,a_i - 1} \left(1 - \sum_{i=1}^{k-1} \widetilde{x}_i\right)^{a_k - 1}}{B(a)}$.

[60] Ver nota de pie **??**.

[61] Ver nota de pie 36, reemplazando n por 1.

[62] Ver nota de pie 36, reemplazando n por 1. Ademas, la forma de la función generadora de momento viene directamente de la escritura de las series de Taylor de $e^{u_i x_i}$ o de la forma integral de la función confluente hipergeométrica (Phillips, 1988).



(a) (b) (c)

Figura 1-30: Ilustración del dominio Δ_{k-1} de definición de la ley de Dirichlet para $k = 3$ (grise oscuro), con el dominio $(k-1)$-dimensional $\mathcal{T}_{k-1}$ del vector $\widetilde{X} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^t$ $(X_3 = 1 - X_1 - X_2)$ (grise claro) (a), y densidad de probabilidad de $\widetilde{X}$ con las marginales p_{X_1}, p_{X_2} (ver notas de pie **??** y 59). Los parámetros son $a = \begin{bmatrix} 3 & 2 & 2 \end{bmatrix}^t$ (b) y $a = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}^t$ (c).

Vectores de distribución de Dirichlet tienen también unas propiedades notables, parecidas a las de la beta:

Lema 1-37 (Reflexividad)**.** *Sea* $X \sim \mathcal{D}ir(a), a \in \mathbb{R}_+^{*k}$ *y* $\Pi \in \mathfrak{S}_k(\mathbb{R})$ *matriz de permutación. Entonces*

$$\Pi X \sim \mathcal{D}ir\left(\Pi a\right)$$

Demostración. El resultado es inmediato por cambio de variables $x \to \Pi x$, la Jacobiana siende Π, de valor absoluto determinente igual a 1 (ver sección 1.4). □

Además, se muestra una stabilidad remplazando dos componentes por su suma:

Lema 1-38 (Stabilidad por agregación)**.** *Sea* $X = \begin{bmatrix} X_1 & \cdots & X_k \end{bmatrix}^t \sim \mathcal{D}ir(a)$, $a = \begin{bmatrix} a_1 & \cdots & a_k \end{bmatrix}^t \in \mathbb{R}_+^{*k}$ *y* $G^{(i,j)}$ *matriz de agrupación de las* (i,j)*-ésima componentes (ver notaciones). Entonces,*

$$G^{(i,j)} X \sim \mathcal{D}ir\left(G^{(i,j)} a\right)$$

Demostración. Se puede probar este resultado a partir de la función característica, usando las propiedades de la función confluent hipergeométrica (Srivastava & Karlsson, 1985; Humbert, 1922; Appell, 1925; ?, ?, ?; Erdélyi, 1940). Pero se puede también tener un enfoque más directo. Del lema precediente, notando que existen matrices de permutación [63] $\Pi_k \in \mathfrak{S}_k(\mathbb{R})$ y $\Pi_{k-1} \in \mathfrak{S}_{k-1}(\mathbb{R})$ tal que $G^{(i,j)} = \Pi_{k-1} G^{(1,2)} \Pi_k$, se puede concentrarse en el caso $(i,j) = (1,2)$. Sea el cambio de variables $g : x = (x_1, \ldots, x_k) \mapsto u = (u_1, \ldots, u_k) = (x_1, x_1 + x_2, x_3, \ldots, x_k)$. Entonces

[63] Π_k pone las componentes i e j el las posiciones 1 y 2, sin cambiar el orden de las siguientes; Π_{k-1} trazlada la primera componente en la posición $\mín(i,j)$.

$g^{-1}(u) = (u_1, u_2 - u_1, u_3, \ldots, u_k)$ es de determinente de matriz Jacobiana igual a 1 dando para $U = g(X)$ la densidad

$$p_U(u) = \frac{u_1^{a_1-1}\,(u_2-u_1)^{a_2-1}\prod_{i=3}^{k} u_i^{a_i-1}}{B(a)}$$

sobre $g\,(\Delta_{k-1})$. Para $u_2 \in [0\,;\,1]$ tenemos $u_1 \in [0\,;\,u_2]$ así que, por marginalización en u_1 obtenemos la densidad

$$\begin{aligned} p_{G^{(1,2)}X}(u_2,\ldots,u_k) &= \frac{\prod_{i=3}^{k} u_i^{a_i-1}}{B(a)} \int_0^{u_2} u_1^{a_1-1}\,(u_2-u_1)^{a_2-1}\,du_1 \\ &= \frac{\prod_{i=3}^{k} u_i^{a_i-1}}{B(a)}\, u_2^{a_1+a_2-1} \int_0^{1} v_1^{a_1-1}\,(1-v_1)^{a_2-1}\,dv_1 \end{aligned}$$

con el cambio de variables $u_1 = u_2 v_1$. Se cierra la prueba notando que la integral vale $B(a_1, a_2)$ y que $\frac{B(a_1,a_2)}{B(a)} = \frac{1}{B\left(G^{(1,2)}a\right)}$. □

De este lema, aplicado de manera recursiva, se obtiene en corolario siguiente:

Corolario 1-10. *Sea $X \sim \mathcal{D}ir(a)$, entonces $X_i \sim \beta\,(a_i\,,\,\alpha - a_i)$.*

Naturalmente, la ley de Dirichelt siendo una extensión de la beta, existe también un vínculo entre esta ley y variables de distribución gamma:

Lema 1-39 (Vínculo con la ley gamma)**.** *Sea X vector k-dimensional de componente i-ésima $X_i \sim \mathcal{G}(a_i, c)$, $i = 1, \ldots, k$ independientes y a vector de componente i-ésima a_i. Entonces*

$$\frac{X}{\sum_{i=1}^{k} X_i} \sim \mathcal{D}ir(a)$$

(independientemente de c). Además, $\frac{X}{\sum_{i=1}^{k} X_i}$ y $\sum_{i=1}^{k} X_i$ son independientes.

Demostración. La prueba sigue exactamente los mismos pasos que la del lema 1-35 trabajando con $\widetilde{X}$. □

Naturalmente, la distribución de Dirichlet, extensión de la ley beta, aparece entre otros en problema de inferencia Bayesiana como distribución a priori conjugado [53] del parámetro p de la ley multinomial (Robert, 2007), extensión de la ley binomial.

Polya urn schemes (ver Ash entre otros) , Chinese restaurant

La distribución de Dirichlet se generaliza al caso matriz-variada X definido sobre $\mathcal{P}_{d,k}(\mathbb{R})$, conjuntos de k-uplet de matrices de $P_d^+(\mathbb{R})$ cumpliando la relación de completud (ver notaciones); se denota $X \sim \mathcal{D}ir_d(a)$ donde $a \in \left(\frac{d-1}{2}\,;\,+\infty\right)^k$ la densidad est dada por $p_X(x) = \frac{\prod_{i=1}^{k} |x_i|^{a_i - \frac{d+1}{2}}}{B_d(a)}$. Se refiera a (Gupta & Nagar, 1999, Cap. 6) para tener más informaciones.

1.10.2.9. Distribución Student-t multivariada

En el caso escalar, esta ley fue introducida inicialmente por F. R. Helmert (Helmert, 1875, 1876; Sheynin, 1995) y J. Lüroth (Lüroth, 1876; Pfanzagl, 1996). Pero es más conocida por su introducción

por William Sealy Gosset [64] en 1908, trabajando sobre variables centradas normalizadas por el promedio y varianza empiricos (Student, 1908). Fue estudiada entre otros intensivamente por el famoso matematico R. Fisher (Fisher, 1925). En la literatura, esta ley es conocida bajo los nombres *Student*, *Student-t* o simplemente *t-distribución* o aún bajo el nombre *Pearson tipo IV* debido a la familia de Pearson (Pearson, 1895).

Dominio de definición	$\mathcal{X} = \mathbb{R}^d$
Parámetro	$\nu \in \mathbb{R}_+^*$ (grados de libertad), $m \in \mathbb{R}^d$ (posición), $\Sigma \in P_d^+(\mathbb{R})$ (matriz caracerística)
Densidad de probabilidad	$p_X(x) = \dfrac{\Gamma\left(\frac{\nu+d}{2}\right)}{\pi^{\frac{d}{2}}\nu^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)\,\lvert\Sigma\rvert^{\frac{1}{2}}}\left(1+\dfrac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right)^{-\frac{\nu+d}{2}}$
Promedio	$m_X = m$ si $\nu > 1$; no existe si no [65].
Covarianza [66]	$\Sigma_X = \dfrac{\nu}{\nu-2}\,\Sigma$ si $\nu > 2$; no existe si no [65].
Sesgo (caso escala)	$\gamma_X = 0$ si $\nu > 3$; no existe si no [65].
Curtosis por exceso	$\bar{\kappa}_X = \dfrac{6}{\nu-4}$ si $\nu > 4$; no existe si no [65].
Función característica [67]	$\Phi_X(\omega) = \dfrac{\nu^{\frac{\nu}{4}}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)}\,e^{\imath\omega^t m}\,\left(\omega^t\Sigma\omega\right)^{\frac{\nu}{4}}\,K_{\frac{\nu}{2}}\left(\sqrt{\nu\,\omega^t\Sigma\omega}\right)$

Nota: nuevamente se puede escribir $X \overset{\mathrm{d}}{=} \Sigma^{\frac{1}{2}}S + m$ con $S \sim t_\nu(0, I)$ donde S es dicha *Student-t estandar* y las características de X son vínculadas a las de S (y vice-versa) por transformación lineal (ver secciones anteriores).

---

[64] De hecho, Gosset fue un estudiante trabajando en la fábrica de cerveza irlandesa Guiness sobre estadística relacionada a la quiímica de la cerveza. A pesar que hay varias explicaciones sobre el hecho de que se publicó este trabajo bajo el nombre "Student". Unas es que fue para que no se sabe que la fábrica estaba trabajando sobre estas estadisticas para estudiar la calidad de la cerveza (Wendl, 2016).

[65] De manera general, esta ley admite momentos de orden k si y solamente si $\nu > k$.

[66] Fijense de que Σ no es la covarianza, pero es proporcional a la covarianza... cuando existe. Se podría imaginar renormalizar la ley tal que Σ_X y Σ coinciden, pero no sería posible en el caso $\nu \leq 2$.

[67] Se muestra sencillamente que la función generatriz de momentos puede existir si y solamente si $\Re e\{u\} = 0$. La función genetratriz de momentos restricta al producto cartesiano de bandas $\Re e\{u\} = 0$ es nada más que la función característica. Además, esta función fue calculdada, especialmente en el caso multivariado, relativamente recientemente (Sutradhar, 1986; Hurst, 1995; Kibria & Joarder, 2006; Song, Park & Kim, 2014).

La densidad de probabilidad Student-t estandar y la función de repartición en el caso escalar son representadas en la figura Fig. 1-31-(a) y (b) y una densidad en un contexto bi-dimensional figura Fig. 1-31(c).

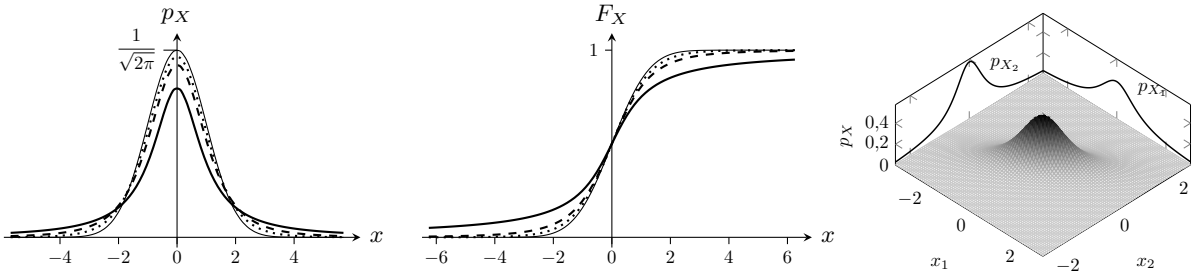


Figura 1-31: Ilustración de una densidad de probabilidad Student-t escalar estandar (a), y la función de repartición asociada (b) con $\nu = 1$ (linea llena), $\nu = 3$ (linea guionada), $\nu = 7$ (linea punteada) y $\nu \to +\infty$ (linea llena fina; ver más adelante) grados de libertad, así que una densidad de probabilidad Student-t bi-dimensional con $\nu = 1$ grado de libertad, centrada, y de matriz característica $\Sigma = R(\theta)\Delta^2 R(\theta)^t$ con $R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ matriz de rotación y $\Delta = \operatorname{diag}\left(\begin{bmatrix} 1 & a \end{bmatrix}\right)$ matriz de cambio de escala, y sus marginales $X_1 \sim t_\nu\left(0, \cos^2\theta + a^2\sin^2\theta\right)$ y $X_2 \sim t_\nu\left(0, \sin^2\theta + a^2\cos^2\theta\right)$ (ver más adelante). En la figura, $a = \frac{1}{3}$ y $\theta = \frac{\pi}{6}$.

Nota: el caso $\nu = 1$ es conocido como distribución de *Cauchy* o *Cauchy Breit-Wigner* (Samorodnitsky & Taqqu, 1994; ?, ?, ?). Es un caso particular también de distrubución α-estables (Samorodnitsky & Taqqu, 1994).

Contrariamente al caso gaussiano, de la forma de la densidad de probabilidad, es claro que si la matriz Σ es diagona, la densidad no factoriza, así que las componentes del vector no son independientes. Este ejemplo muestra claramente que la reciproca del lema 1-6 es falsa en general.

Sin embargo, las distribuciones Student-t tienen varias propiedades notables.

Lema 1-40 (Stabilidad por transformación lineal)**.** *Sea $X \sim t_\nu(m, \Sigma)$, A matriz de $\mathbb{R}^{d' \times d}$ con $d' \leq d$, y de rango lleno y $b \in \mathbb{R}^{d'}$. Entonces*

$$AX + b \sim t_\nu(Am + b, A\Sigma A^t)$$

En particular los componentes de X son student-t,

$$X_i \sim t_\nu(m_i, \Sigma_{i,i})$$

Demostración. La prueba es inmediata usando la función característica y sus propiedades por transformación lineal. La condición sobre A es necesaria y suficiente para que $A\Sigma A^t \in P_{d'}^+(\mathbb{R})$. □

Lema 1-41 (Vínculo con las distribuciones Gamma y Gausiana (mezcla Gaussiana de escala))**.** *Sea* $V \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ *y* $G \sim \mathcal{N}(0, I)$ *independientes. Entonces*

$$\frac{G}{\sqrt{V}} \sim t_\nu(0, I)$$

Dicho de otra manera, se puede escribir $X \sim t_\nu(m, \Sigma)$ *esticasticamente bajo la forme* $X \overset{\mathrm{d}}{=} \sqrt{\frac{\nu}{V}}\Sigma^{\frac{1}{2}}G + m$ *donde* $\overset{\mathrm{d}}{=}$ *significa que la igualdad es en distribución.*

Demostración. Lo más simple es de salir de la formula de probabilidad total vista pagina **??**, notando que condicionalmente a $V = v$ la variable es gausiana de covarianza $\frac{1}{\sqrt{v}}I$,

$$\begin{aligned} p_X(x) &= \int_{\mathbb{R}} p_{X|V=v}(x)\, p_V(v)\, dv \\ &\propto \int_0^{+\infty} v^{\frac{d}{2}} e^{-\frac{v}{2}x^t x} v^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}v}\, dv \\ &\propto \left(1 + \frac{x^t x}{\nu}\right)^{-\frac{d+\nu}{2}} \int_0^{+\infty} u^{\frac{d+\nu}{2}-1} e^{-u}\, du \\ &\propto \left(1 + \frac{x^t x}{\nu}\right)^{-\frac{d+\nu}{2}} \end{aligned}$$

con $\propto$ significando "proporcional a" (el coeficiente es lo de normalización) y el cambio de variables $v = \frac{2\,u}{\nu + x^t x} = \frac{\frac{2}{\nu}}{1+\frac{x^t x}{\nu}}\, u$. □

Nota: este lema permite también probar el lema 1-40 escribiendo $AX + b \overset{\mathrm{d}}{=} \sqrt{\frac{\nu}{V}} A\Sigma^{\frac{1}{2}}G + Am + b$.

Lema 1-42 (Límite Gausiana)**.** *Sea* $X_\nu \sim t_\nu(m, \Sigma)$ *vector Student-*t *parametizado por* ν *sus grados de libertad. Entonces*

$$X_\nu \xrightarrow[\nu\to\infty]{\mathrm{d}} = X \sim \mathcal{N}(m, \Sigma)$$

donde $\xrightarrow{\mathrm{d}}$ *significa que el límite es en distribución.*

Demostración. La prueba es inmediata tomando el logaritmo de la densidad de probabilidad, usando la formula de Stirling [68] para $\log\Gamma(z) = \left(z - \frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + o(1)$ en $z \to +\infty$ (Stirling, 1730; Abramowitz & Stegun, 1970; Gradshteyn & Ryzhik, 2015) y $-\frac{d+\nu}{2}\log\left(1 + \frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu}\right) = -\frac{d+\nu}{2}\left(\frac{(x-m)^t\Sigma^{-1}(x-m)}{\nu} + o\left(\nu^{-1}\right)\right) = -\frac{(x-m)^t\Sigma^{-1}(x-m)}{2} + o(1)$. □

Sampling distribution, Gosset, Fisher 25.

Applications

La distribución Student-t se generaliza al caso matriz-variada X definido sobre $M_{d,d'}(\mathbb{R})$; se denota $X \sim t_\nu(M, \Sigma, \Omega)$ donde $M \in M_{d,d'}(\mathbb{R})$, $\Sigma \in P_d^+(\mathbb{R})$, $\Omega \in P_{d'}^+(\mathbb{R})$ y la densidad est dada por $p_X(x) =$

[68] Ver nota de pie [47]

$\frac{\Gamma_d\left(\frac{\nu+d+d'-1}{2}\right)}{\pi^{\frac{\nu d}{2}}\Gamma_d\left(\frac{\nu+d-1}{2}\right)\,|\Sigma|^{\frac{d'}{2}}\,|\Omega|^{\frac{d}{2}}}\left|I+\Sigma^{-1}(x-M)\Omega^{-1}(x-M)^t\right|^{-\frac{\nu+d+d'-1}{2}}$. Se refiera a (Gupta & Nagar, 1999, Cap. 4) para tener más informaciones.

von Mises en el circulo? Y multivariate (von Mises-Fisher)? Cantor = singular? A ver con la idea de mixta

1.10.2.10. Familia exponencial (Darmois, 1935; Koopman, 1936; Andersen, 1970; Kay, 1993; Lehmann & Casella, 1998; Robert, 2007).; Muchas de estas leyes entran en una familia que juega un rol particular en problema de maximizacion de entropie (ver cap 2): es la familia exponencial...

Caso discreto misma forma (en footnote)

1.10.2.11. Familia eliptica Invariante por rotacion... GSM

Teorema del límite central, y relajando la independencia, y versiones con leyes diferentes pero uniformamente acotadas.

hablar de simulación? Metoto inverso, mezcla, rejeccion, a traves de la condicional para el caso vectorial?

EPÍLOLOGO

Este libro surgio de la experiencia de los autores en el dictado del curso semestral “Métodos de geometría diferencial en teoría de la información”, que se imparte en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y en la Facultad de Matemática, Astronomía y Física de la Universidad Nacional de Córdoba. **. . . acabar** *Los autores*

Referencias

Ablowitz, M. J. & Fokas, A. S. (2003). *Complex Variables: Introduction and Applications* (2nd ed.). New-York: Cambridge University Press.

Abramowitz, M. & Stegun, I. A. (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th printing. New-York: Dover.

Akritas, A. A., Akritas, E. K., & Malaschonok, G. I. (1996). Various proofs of sylvester's (determinant) identity. *Mathematics and Computers in Simulation*, *42*(4-6), 585–593.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Andersen, E. S. & Larsen, M. E. (1994). Combinatorial summation identities. In *Analysis: Algebra and Computers in Mathematical Research: Proceedings of the Twenty-first Nordic Congress of Mathematicians*, Lecture Notes in Pure and Applied Mathematics, (pp. 1–23). CRC Press.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken, New-Jersey, USA: John Wiley & Sons.

Andrew, G. E. & Berndt, B. C. (2013). *Ramanujan's Lost Notebook-Part IV*. New-York: Springer.

Appell, P. (1925). *Sur les fonctions hypergéométriques de plusieurs variables, les pôlynomes d'Hermite et autres fonctions hypersphériques dans l'hyperespace*, volume fascicule 3. Paris: Mémorial des Sciences Mathématiques; Gauthier-Villars.

Ash, R. B. & Doléans-Dade, C. A. (1999). *Probability and Measure Theory* (2nd ed.). San Diego, CA, USA: Academic Press.

Askey, R. A. (1975). *Orthogonal Polynomials and Special Functions*. SIAM.

Athreya, K. B. & Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. New-York: Springer.

Balakrishnan, N. & Basu, A. P. (1995). *The Exponential Distribution: theory, Methods and Applications*. Amsterdam, The Netherlands: Gordon an Breach Publishers.

Barnard, G. A. (1958). Studies in the history of probability and statistics: IX. Tomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*(3-4), 293–295.

Barone, J. & Novikoff, A. (1978). A history of the axiomatic formulation of probability from Borel to Kolmogorov: Part I. *Archive for History of Exact Sciences*, *18*(2), 123–190.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Bellhouse, D. (2005). Decoding cardano's liber de ludo aleae. *Historia Mathematica*, *32*(2), 180–202.

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilæ reticularis*. Basel, Switzeland: Thurneysen Brothers.

Bhatia, R. (2007). *Positive Definite Matrices*. Princeton: Princeton University Press.

Bienaymé, I.-J. (1853). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrées. *Comptes Rendus de l'Académie des Sciences.*, *37*, 158–176.

Billingsley, P. (2012). *Probability and Measure* (3rd ed.). Hoboken, NJ, USA: John Willey & Sons.

Bochner, S. (1932). Ein konvergenzsatz für mehrvariablige fouriersche integrale. *Mathematische Zeitschrift*, *34*(1), 440–447.

Bochner, S. (1959). Monotonic functions, Stieltjes integrals and harmonic analysis. In *Lectures on Fourier Integrals* (pp. 292–331). Princeton University Press.

Bogachev, V. I. (2007a). *Measure Theory*, volume I. Berlin: Springer.

Bogachev, V. I. (2007b). *Measure Theory*, volume II. Berlin: Springer.

Boltzmann, L. (1896). *vorlesungen über Gastheorie - I*. Leipzig, Germany: Verlag von Johann Ambrosius Barth.

Boltzmann, L. (1898). *vorlesungen über Gastheorie - II*. Leipzig, Germany: Verlag von Johann Ambrosius Barth.

Borel, E. (1898). *Le0.98628cons sur la théorie des fonctions*. Paris: Gauthier-Villars et fils.

Borel, E. (1909). *Éléments de la théorie des probabilités*. Paris: A. Hermann & fils.

Bouniakowsky, V. (1859). Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies. *Mémoires de l'Académie Impériale des Sciences de Saint-Petersbourg*, *I*(9).

Brémaud, P. (1988). *An Introduction to Probabilitic Modeling*. New-York: Springer.

Brockwell, P. J. & Davis, R. A. (1987). *Time Series: Theory and Methods* (2nd ed.). New-York: Springer Verlag.

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, *11*(3), 368–385.

Cardano, J. (1663). *Liber de ludo aleae, en "Opera Omnia"*, volume 1, (pp. 262–276). Lyon: cura Caroli Sponii.

Carrier, G. F., Krook, M., & Pearson, C. E. (2005). *Function of a Complex Variable: Theory and Technique*. Philadelphia: SIAM.

Cauchy, A.-L. (1821). *Cours d'analyse de l'école royale polytechnique*, volume 1: analyse algébrique. Paris: Imprimerie royale (digital version, Cambrige, 2009).

Cohn, D. L. (2013). *Measure Theory* (2nd ed.). New-York: Springer.

Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *Comptes rendus de l'Acadéie des Sciences*, *200*, 1265–1966.

de Laplace, P. S. (1820). *Théorie analytique des Probabilités* (3ème ed.). Paris: Gauthier-Villars.

de Moivre, A. (1710). De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus. *Philosophical Transactions of the Royal Society of London*, *27*(329), 213–264.

de Moivre, A. (1730). *Miscellanea analytica de seriebus et quadraturis*. London: Londini: J. Tonson & J. Watts.

de Moivre, A. (1733). Approximatio ad summam terminorum binomii $(a + b)^n$ in seriem expansi. Very few copies published privately in London (see also The Doctrine of Chance).

de Moivre, A. (1756). *The Doctrine of Chances : or, a method for calculating the probabilities of events in play* (3rd ed.). London: AMS Chelsea Publishing.

de Montmort, P. R. (1713). *Essay d'analyse sur les jeux de hazard* (2nd ed.). Paris: Jacque Quillau, Imprimeur Juré Libraire de l'Université.

de Morgan, A. (1838). *An Essay on Probabilities and on their application to Life Contingencies and Insurance Offices*. London, UK: Longman, Orme, Brown, Green & Longmans.

Deming, W. E. (1933). De moivre's "miscellanea analytica", and the origin of the normal curve. *Nature*, *132*(3340), 713–713.

Dutka, J. (1991). The early history of the factorial function. *Archive for History of Exact Sciences*, *43*(3), 225–249.

(E. D. Sylla, Translator), J. B. (1713). *The Art of Conjecturing - Together with a "Letter to a Friend on Set in Court Tennis"*. Johns Hopkins University Press.

Eaton, M. L. (1981). On the projections of isotropic distributions. *Annals of Statistics*, *9*(2), 391–400.

Erdélyi, A. (1940). Integration of a certain system of linear partial differential equations of hypergeometric type. *Proceedings of the Royal Society of Edinburgh*, *59*, 224–241.

Eriksson, J. & Koivunen, V. (2006). Complex random vectors and ICA models: Identifiability, uniqueness, and separability. *IEEE Transactions on Information Theory*, *52*(3), 1017–1029.

Euler, L. (1741). Observationes analyticae varias de combinationibus. *Commentarii academiae scientiarum Petropolitanae*, *13*, 64–93.

Euler, L. (1750). De partitione numerorum. *Novi Commentarii academiae scientiarum Petropolitanae*, *3*, 125–169.

Euler, L. (1768). *Lettres à une princesse d'Allemagne sur divers sujets de physique & de philosophie*, volume 2. Saint Petersbourg, Rusia: Académie Impériale des Sciences de Saint Petersbourg.

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Monographs on statistics and probability 36. London: Chapman & Hall.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3 ed.)., volume 1. New-York: John Wiley & Sons, Inc.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. New-York: John Wiley & Sons, Inc.

Ferentinos, K. (1982). On Tchebycheff's type inequalities. *Trabajos de Estadística e Investigación*

Operativa, *33*(1), 125–132.

Fisher, R. A. (1925). Applications of “Student’s” distributions. *Metron*, *5*(3), 90–104.

Fourier, J. (1822). *Théorie Analitique de la Chaleur*. Paris: Firmin Didot, père et fils.

Galton (1877a). Typical laws of heredity. *Journal of the Royal Institution of Great Britain*, *8*, 282–301.

Galton, F. (1877b). Typical laws of heredity. *Nature*, *15*, 492–495.

Galton, F. (1889). *Natural Inherence*. London, UK: McMillan.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hambourg, Germany: Perthes et Besser.

Gauss, C. F. (1810). *Disquisitio de elementis ellipticis Palladis, ex oppositionibus annorum 1803, 1804, 1805, 1807, 1808, 1809*. Göttingen, Germany: Gottingae : Apud H. Dieterich.

Gel’fand, I. M. & Shilov, G. E. (1964). *Generalized Functions*, volume 1: Properties and Operations. New-York: Academic Press.

Gel’fand, I. M. & Shilov, G. E. (1968). *Generalized Functions*, volume 2: Spaces of Fundamental and Generalized Functions. New-York: Academic Press.

Gibbs, J. W. (1902). *Elementary Principle in Statistical Mechanics*. Cambridge, USA: University Press - John Wilson and son.

Golberg, R. R. (1961). *Fourier Transforms*. Cambridge University Press.

Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, *34*(1), 152–177.

Gradshteyn, I. S. & Ryzhik, I. M. (2015). *Table of Integrals, Series, and Products* (8th ed.). San Diego: Academic Press.

Gupta, A. K. & Nagar, D. K. (1999). *Matrix Variate Distributions*. Chapman and Hall.

Gupta, R. D. & Richards, D. S. P. (2001). The history of the Dirichlet and Liouville distributions. *International Statistical Review*, *69*(3), 433–446.

Hadamard, J. (1893). Etude sur les propriétés des fonctions entières et en particulier d’une function considérée par Riemann. *Journal de Mathématiques Pures et Appliquées*, *58*(9), 171–215.

Hald, A. (1984). Nicholas Bernoulli’s theorem. *International Statistical Review*, *52*(1), 93–99.

Hald, A. (1990). *History of Probability and Statistics and Their Applications before 1750*. John Wiley & Sons, Inc.

Hald, A. (2006). *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. New-York, USA: Springer.

Halmos, P. R. (1950). *Measure Theory*. New-York: Springer.

Harville, D. A. (2008). *Matrix Algebra From a Statistician’s Perspective*. Springer-Verlag.

Hausdorff, F. (1901). Beiträge zur wahrscheinlichkeitsrechnung. *Berichte über die Verhandlungen der Köeniglich Sächsische Akademie der Wissenschaften zu Leipzig*, *53*(1), 152–178.

Helmert, F. R. (1875). Über die bestimmung des wahrscheinlichen fehlers aus einer endlichen anzahl wahrer beobachtungsfehler. *Zeitschrift für Mathematik und Physik*, (8), 300–303.

Helmert, F. R. (1876). Die genauigkeit der formel von peters zur berechnung des wahrscheinlichen beobachtungsfehlers directer beobachtungen gleicher genauigkeit. *Astronomische Nachrichten*, *88*(8-9), 113–131.

Hogg, R. V., McKean, J. W., & Craig, A. (2013). *Introduction to Mathematical Statistics* (7th ed.). Boston: Pearson.

Hölder, O. (1889). Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, *2*, 38–47.

Horn, R. A. & Johnson, C. R. (2013). *Matrix Analysis* (2nd ed.). Cambridge University Press.

Humbert, P. (1922). The confluent hypergeometric functions of two variables. *Proceedings of the Royal Society of Edinburgh*, *41*, 73–96.

Hurst, S. (1995). The characteristic function of the student-t distribution. Technical Report Financial Mathematics Research Report No. FMRR006-95, Statistics Research Report No. SRR044-95, Center for Financial Mathematics, School of Mathematical Sciences, Australian National University, Canberra, Australia.

Huygens, C. (1657). De ratiociniis in ludo aleae. In *printed in Exercitationum Mathematicarum by F. Van Schooten*. Leiden, The Netherland: Elsevirii.

Ibarrola, P., Pardo, L., & Quesada, V. (1997). *Teoría de la Probabilidad*. Madrid: Síntesis.

Jacob, J. & Protters, P. (2003). *Probability Essentials* (2nd ed.). Berlin: Springer.

Jaynes, E. T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, *33*(5), 391–398.

Jeffrey, H. (1948). *Theory of Probability* (2nd ed.). Oxford: Clarendon.

Jeffrey, H. (1973). *Scientific Inference* (3rd ed.). Cambridge: Cambridge University Press.

Jensen, J. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, *30*(1), 175–193.

Jessen, B. (1931a). Bemærkninger om konvekse funktioner og uligheder imellem middelærdier. I. *Matematisk Tidsskrift. B*, 17–28.

Jessen, B. (1931b). Bemærkninger om konvekse funktioner og uligheder imellem middelærdier. II. *Matematisk Tidsskrift. B*, 84–95.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995a). *Continuous Univariate Distributions* (2nd ed.)., volume 1. New-York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995b). *Continuous Univariate Distributions* (2nd ed.)., volume 2. New-York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New-York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions* (2nd ed.). New-York: John Wiley & Sons.

Kay, S. M. (1993). *Fundamentals for Statistical Signal Processing: Estimation Theory*. vol. 1. Upper Saddle River, NJ: Prentice Hall.

Kibria, B. M. G. & Joarder, A. H. (2006). A short review of multivariate t-distribution. *Journal of Statistical Research*, *40*(1), 59–72.

Knuth, D. E. (1997). *The Art of Computer Programming* (3rd ed.)., volume 1 / fundamental algorithms. Reading: Addison Wesley Longman.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability* (2nd ed.). New-York: Chelsea Publishing Company.

Kolmogorov, A. N. & Fomin, S. V. (1957). *Elements of the Theory of Function and Functional Analysis*, volume 1: Metric and Normed Spaces. Rochester, NY, USA: Graylock Press.

Kolmogorov, A. N. & Fomin, S. V. (1961). *Elements of the Theory of Function and Functional Analysis*, volume 2: Measure. The Lebesgue Integral. Hilbert Space. Rochester, NY, USA: Graylock Press.

Kondo, T. (1930). A theory of the sampling distribution of standard deviations. *Biometrika*, *22*(1-2), 36–64.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–399.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous Multivariate Distributions* (2nd ed.)., volume 1: Models and Applications. New-York: John Wiley & Sons.

Langius, J. C. (1712). *Nvclevs Logicae Weisianae*. Giessen: Henningius Müllerus.

Lapidoth, A. (2017). *A Foundation in Digital Communication* (2nd ed.). Cambridge, UK: Cambridge University Press.

Laplace, P. S. (1809a). Mémoire sur les approximations des formules qui sont fonctions de très grand nombres et sur leur application aux probabilités. *Mémoires de l'académie des sciences de Paris*, *Ière série T. X.*, 353–415.

Laplace, P. S. (1809b). Supplément aux mémoire sur les approximations des formules qui sont fonctions de très grand nombres et sur leur application aux probabilités. *Mémoires de l'académie des sciences de Paris*, *Ière série T. X.*, 559–565.

Le Cam, L. (1986). The central limit theorem around 1935. *Statistical Science*, *1*(1), 78–91.

Lebesgue, H. (1904). *Le0.98628cons sur l'Intégration et la recherche des Fonctions Primitives*. Paris: Gauthier-Villars et fils.

Lebesgue, H. (1918). Remarques sur les théories de la mesure et de l'intégration. *Annales Scientifiques de l'Ecole Normale Supérieure*, *35*, 191–250.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris, France: Firmin Didot, Libraie pour les Mathématiques, la Marine, L'architecture, et les Édition stéréotypes.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New-York: Springer-Verlag.

Lejeune-Dirichlet (1839). Sur une nouvelle méthode pour la détermination des intégrales multiples. *Journal de Mathématiques Pures et Appliquées*, 164–168.

Lieb, E. H. & Loss, M. (2001). *Analysis* (2nd ed.). Providence, Rhode Island: American Mathematical

Society.

Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung Mathematische Zeitschrift*, *15*, 211–225.

Liouville (1839). Note sur quelques intégrales définies. *Journal de Mathématiques Pures et Appliquées*.

Lord, R. (1954). The use of the Hankel transform in statistics I. General theory and examples. *Biometrika*, *41*(1/2), 44–55.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.

Lukacs, E. (1961). Recent developments in the theory of characteristic functions. In *Proceeding of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 2: Contributions to Probability Theory, (pp. 307–335). University of California Press, Berkeley, CA.

Lüroth, J. (1876). Vergleichung von zwei werthen des wahrscheinlichen fehlers. *Astronomische Nachrichten*, *87*(14), 209–220.

Magnus, J. R. & Neudecker, H. (1979). The commutation matrix: Some properties and applications. *The Annals of Statistics*, *7*(2), 381–394.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3rd ed.). New-York: John Wiley & Sons.

Mandel, L. & Wolf, E. (1995). *Optical coherence and quantum optics*. Cambridge University Press.

Markov, A. (1884). *On certain applications of algebraic continued fractions*. PhD thesis, University of Saint Petersburg, St. Petersburg, Russia.

Marquis de Condorcet (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris, France: Imprimerie Royale de Paris.

Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of Majorization and Its Applications* (2nd ed.). New-York: Springer Verlag.

Maxwell, J. C. (1867). On the dynamical theory of gases. *Philosophical Transactions of the Royal Society of London*, *157*, 49–88.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference* (5th ed.)., volume 162 of *"Statistics: textbooks and monographs"*. New-York: Marcel Dekker.

Navarro, J. (2013). A very simple proof of the multivariate Chebyshev's inequality. *Communications in Statistics - Theory and Methods*, *45*(12), 3458–3463.

Neudecker, H. & Wansbeek, T. (1983). Some results on commutation matrices, with statistical applications. *Canadian Journal of Statistics*, *11*(3), 221–231.

Nikodym, O. (1930). Sur une généralisation des intégrales de M. J. Radon. *Fundamenta Mathematicae*, *15*(1), 131–179.

Olkin, I. & Pratt, J. W. (1958). A multivariate tchebycheff inequality. *The Annals of Mathematical Statistics*, *29*(1), 226–234.

Olkin, I. & Rubin, H. (1964). Multivariate Beta distributions and independence properties of the Wishart distribution. *The Annals of Mathematical Statistics*, *35*(1), 261–269.

Pascal, B. (1679). *Varia opera Mathematica D. Petri de Fermat Senatoris Tolosae*, chapter Lettre de Monsieur Pascal à M. de Fermat, (pp. 184–188). Toulouse, France: Joannem Pech, Comitiorum Fuxenfiul Typographum, juxta Collegium PP. Societatis Jesu.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *186*, 343–414.

Pearson, K. (1905). "das fehlergesetz und seine verallgemeinerungen durch fechner und pearson.". A rejoinder. *Biometrika*, *4*(1/2), 169–212.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25–45.

Pearson, K. (1924). Historical note on the origin of the normal curve of errors. *Biometrika*, *16*(3/4), 402–404.

Pearson, K., de Moivre, A., & Archibald, R. C. (1926). A rare pamphlet of Moivre and some of its discoveries. *ISIS A Journal of The History of Science Society*, *8*(4), 671–683.

Peddada, S. D. & Richards, D. S. P. (1991). Proof of a conjecture of M. L. Eaton on the characteristic function of the Wishart distribution. *The Annals of Probability*, *19*(2), 868–874.

Perlman, M. D. (1974). Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *Journal of Multivariate Analysis*, *4*(1), 52–65.

Pfanzagl, J. (1996). Studies in the history of probability and statistics XLIV. a forerunner of the t-distribution. *Biometrika*, *83*(4), 891–898.

Phillips, P. C. B. (1988). The characteristic function of the Dirichlet and multivariate F distribution. discussion paper 985, Cowles Foundation for Research in Economics Yale University.

Picinbono, B. (1996). Second-order complex random vectors and normal distributions. *IEEE Transactions on Signal Processing*, *44*(10), 2637–2640.

Pinsky, M. A. (2009). *Introduction to Fourier Analysis and Wavelets*, volume 102. Providence, Rhode Island, USA: American Mathematical Society.

Poisson, S. D. (1837). *Recherches sur la probabilit'e des jugemnts en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Paris, France: Bachelier, Imprimeur-Librairie pour les mathématiques, la physique, etc.

Polya, G. (1920). Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, *8*(3-4), 171–181.

Rényi, A. (2007). *Probability Theory*. Mineola, New-York: Dover Publications INC.

Robert, C. P. (2007). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation* (2nd ed.). New-York: Springer.

Rudin, W. (1991). *Functional Analysis* (2nd ed.). New-York: McGraw-Hill.

Samorodnitsky, G. & Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes. Stochastic Models*

with infinite Variance. New-York: Chapman & Hall.

Sasvári, Z. (2013). *Multivariate Characteristic Functions and Correlations Functions*. Berlin, Germany: Walter De Gruyter GmbH.

Schreier, P. & Scharf, L. (2003). Second-order analysis of improper complex random vectors and processes. *IEEE Transactions on Signal Processing*, *51*(3), 714–725.

Schwartz, L. (1966). *Théorie des distributions*. Paris: Hermann.

Schwarz, H. A. (1888). Ueber ein die flächen kleinsten flächeninhalts betreffendes problem der variationsrechnung. *Acta societatis scientiarum Fennicæ*, *15*, 315–362.

Seber, G. (2004). *Multivariate Observations*. Hoboken, New-Jersey, USA: John Wiley & Sons.

Serrano Marugán, E. (2000). Etimología de algunos términos matemáticos. *Suma*, *35*, 87–96.

Shafer, G. & Vovk, V. (2006). The sources of Kolmogorov's grundbegriffe. *Statistical Science*, *21*(1), 70–98.

Sheynin, O. (1995). Helmert's work in the theory of errors. *Archive for History of Exact Sciences*, *49*(1), 73–104.

Shohat, J. (1929). Inequalities for moments of frequency functions and for various statistical constants. *Biometrika*, *21*(1-4), 361–375.

Sierpiński, W. (1918). Sur les définitions axiomatiques des ensembles mesurables. *Bulletin international de l'Académie des sciences de Cracovie: Série A. Classe des sciences mathématiques et naturelles – Sciences mathématiques*, 29–34.

Sierpiński, W. (1975). *Oeuvres choisies, Tome II: Théorie des ensembles et ses applications*. Warszawa, Poland: PWM Éditions scientifiques de Pologne.

Sierpiński, W. (1976). *Oeuvres choisies, Tome III: Théorie des ensembles et ses applications*. Warszawa, Poland: PWM Éditions scientifiques de Pologne.

Simpson, T. (1740). *The Nature and Laws of Chance. The Whole afetr a new, general and conspicuous Manner , and illustrated with a great Variety of Examples*. London, UK: Edward Cave.

Song, D.-K., Park, H.-J., & Kim, H.-M. (2014). A note on the characteristic function of multivariate t distribution. *Communications for Statistical Applications and Methods*, *21*(1), 81–91.

Spiegel, M. (1976). *Probabilidad y Estadística*. México: McGrew Hill.

Srivastava, H. M. & Karlsson, P. W. (1985). *Multiple Gaussian Hypergeometric Series*. Chichester: John John Wiley & Sons.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge: Cambridge University Press.

Stein, E. M. & Weiss, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press.

Stellato, B., Van Parys, B. P. G., & Goulart, P. J. (2017). Multivariate Chebyshev inequality with estimated mean and variance. *The American Statistician*, *71*(2), 123–127.

Stirling, J. (1730). *Methodus Differentialis sive Tractatus de Summatione et Interpolatione Serierum*

Infinitarum. Londini: Typis Gul. Bowyer; impensis G. Strahan.

Student (1907). On the error of counting with a haemacytometer. *Biometrika*, *5*(3), 351–360.

Student (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.

Sultan, S. & Tracy, D. S. (1996). Moments of Wishart distribution. *Stochastic Analysis and Applications*, *14*(2), 237–243.

Sutradhar, B. C. (1986). On the characteristic function of multivariate Student t-distribution. *Canadian Journal of Statistics*, *14*(4), 329–337.

Sylvester, J. (1851). On the relation between the minor determinants of linearly equivalent quadratic functions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *1*(4), 295–305.

Tchébichev, P. (1867). Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, *12*, 177–184.

van Brakel, J. (1976). Some remarks on the prehistory of the concept of statistical probability. *Archive for History of Exact Sciences*, *16*(2), 119–136.

van den Bos, A. (1995). The multivariate complex normal distribution-a generalization. *IEEE Transactions on Information Theory*, *41*(2), 537–539.

Venn, J. M. A. (1880). I. On the diagrammatic and mechanical representation of propositions and reasonings. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *10*(59), 1–18.

von Mises, R. (1932). Théorie des probabilités. fondements et applications. *Annales de l'institut Henri Poincaré*, *3*(2), 137–190.

von Plato, J. (2005). A.N. Kolmogorov, Grundbegriffe der wahrscheinlichkeitsrechnung (1933). In *Landmark Writings in Western Mathematics 1640-1940* chapter 75, (pp. 960–969). Elsevier.

Wang, L. & Madiman, M. (2004). Beyond the entropy power inequality via rearrangements. *IEEE Transactions on Information Theory*, *60*(9), 5116–5137.

Wendl, M. C. (2016). Pseudonymous fame. *Science*, *351*(6280), 1406.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905-2014 R.I.P. *The American Statistician*, *68*(3), 191–195.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, *20A*(1/2), 32–52.

Los autores

Lamberti, Pedro Walter

Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea). Este es un párrafo Normal con texto simulado, (Arial 10, interlineado de 1,5 líneas, sin sangría en la primera línea).

Portesi, Mariela Adelina

Obtuvo el título de Licenciada en Física en la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata, y el grado de Doctora en Física en la misma casa de altos estudios. Es Investigador Independiente del Consejo Nacional de Investigaciones Científicas y Técnicas, con lugar de trabajo en el Instituto de Física La Plata. Su especialidad es la teoría y geometría de la información en mecánica cuántica. Posee cargo docente de Profesor Adjunto en el Departamento de Matemática de la Facultad de Ciencias Exactas de la UNLP, desempeñándose desde 2013 como integrante del Equipo Coordinador de la asignatura Análisis Matemático II (CiBEx). cursos de grado avanzados y de posgrado en la Facultad de Ciencias Exactas de la UNLP y en la Facultad de Matemática, Astronomía, Física y Computación de la Universidad Nacional de Córdoba. También ha participado en el dictado del curso de grado “Probabilidades” como Profesor Visitante de la Université Grenoble-Alpes en Francia.

Zozor, Steeve

Nació en 1972 en Colmar, Francia. Obtuvo el título de Ingeniero, de Licenciada, el grado de Doctor y la “Habilitation à diriger de Recherches”, respectivamente en 1995, 1999 y 2012, ambos del Instituto Nacional Politécnico de Grenoble (Grenoble INP), Francia. En 2001, paso varios meses en el Laboratorio de Procesamiento de Señales de la Escuela Politénica Federal de Lausanne (EPFL), Suiza como postdoctorante. Pasó un año en el Instituto de Física de La Plata (IFLP) de la Universidad Nacional de La Plata (UNLP), Argentina (2012-2013) así que varios estancias desde 2010 como profesor visitante. En 2001 ingresó al Centro National de la Investigación Cientifica (CNRS), equivalente Francés del CONICET, como “Chargé de Recherche” (cargado de investigación) y es “Directeur de Recherches” (director de investigación) desde 2017, ambos en el Laboratorio de Imagenes, Palabras, Señales y Automática de Grenoble (GIPSA-Lab), Francia. Desde 2015 es editor asociado de la revista IEEE Signal Processing Letters. Sus temas de investigación incluyen el procesamiento no lineal de señales, el estudio del efecto de resonancia estocástica, el estudio de procesamiento de datos en contextos α-estables y/o

de distribuciones de probabilidad elípticas, la téoria de la información (medidades informacionales generalizadas clásicas y cúanticas) con aplicaciones en procesamientos de datos, mecanica cúantica o ingeniería biomédica. Es a cargo de docencia en varias escuelas de Grenoble-INP de matemática para el ingeniero, probabilidades aplicadas, procesamiento estadístico de señales, métodos bayesianos. Da regularmente un mini-curso sobre los básicos de la teoría de la información en la Facultad de Ciencias Exactas de la UNLP.