GEOMETRÍA E INFORMACIÓN

OPTATIVO

Mariela Adelina Portesi
Pedro Walter Lamberti
Steeve Zozor

Facultad de Ciencias Exactas

UNIVERSIDAD NACIONAL DE LA PLATA

edulp
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 epģrafe 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. ...

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. ...

Mariela A. Portesi
Grenoble, Junio de 2016

Índice

CAPÍTULO 1
Elementos de teoría de probabilidades

Mariela A. Portesi

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)

1.1 Introducción

1.2 Probabilidades

El concepto de *probabilidad* es importante en situaciones donde el resultado (o *outcome*) 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 puede obtenerse 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 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 tienen un dado atributo, la probabilidad asociada a dicho atributo en un dado procesos 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.) [1].

Los axiomas de Kolmogorov [2] 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 $\mathcal{A}$ una colección de conjuntos de Ω, donde los elementos de $\mathcal{A}$ son llamados *eventos*. Por ejemplo, Ω puede ser las caras de un dado de 6 caras (los números naturales del 1 al 6, o las letras *a, b, c, d, e, f*, u otro etiquetado), $\mathcal{A}$ teniendo los eventos A "es un número natural par" y B indicando "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 análogas a las de la teoría de conjuntos (ej. (Spiegel, 1976; Brémaud, 1988; Mandel & Wolf, 1995; Sierpiński, 1975, 1976; Borel, 1898, 1909)):

- Combinación o unión de eventos: $A \cup B$, implicando que se da A, ó B, o ambos (ej. por un dado, A eventos "cara par" y B evento "cara menor o igual a 3" tal que $A \cup B = \{1\,,\,2\,,\,3\,,\,4\,,\,6\}$); Según la literatura, se denota a veces $A + B$ o $A \wedge B$.

- Intersección de eventos: $A \cap B$, implicando que se dan ambos A y B (con el ejemplo precediente, $A \cap B = \{2\}$); Se denota a veces (A, B) o $A \vee B$.

- Complemento de un evento: $\bar{A}$ e indica que no se da A; Se denota a veces $-A$ o A^c (con el ejemplo precediente, $\bar{A} = \{1\,,\,3\,,\,5\}$).

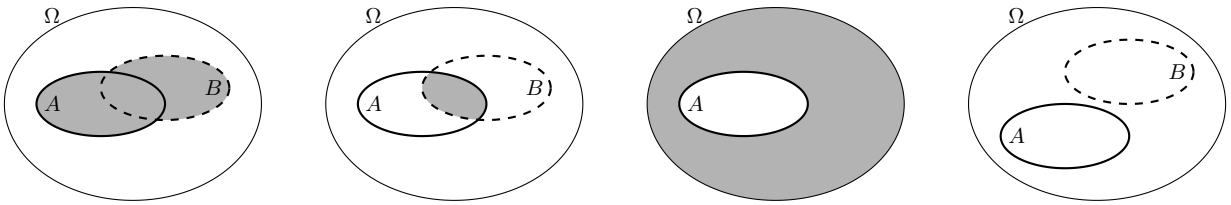
- Eventos *disjuntos* o *mutuamente excluyentes* o *o incompatibles*: son aquellos que no se superponen, se anota $A \cap B = \emptyset$ donde $\emptyset = \bar{\Omega}$ denota el evento nulo (evento que no puede ocurrir, es el complemento de Ω, por ejemplo A "cara par" y B "cara impar").

---

[1] A pesar de que la noción de azar (viniendo del arabe) o de alea (en latin) es muy antiguo, el italiano Gerolamo Cardano es "probablemente" un de los primeros tratando matematicamente del concepto de probabilidad en el siglo XVI, escribiando un libro sobre los juegos de azar en 1564 (D.Bellhouse, 2005) o (Hald, 1990, Cap. 4). Entre los numerosos matematicos desarollando la teoria de las probabildades, (en particular los franceses Pierre de Fermat y Blaise Pascal (Hald, 1990, Cap. 5)) hay que mencionar el suizo Jacob Bernoulli y el francés Abraham de Moivre, quizas un de los primeros llevandos un aporte importante al desarollo de la teoria de las probabilidades en el siglo XVIII a través de este punto de vista "frequencista" y combinatorial (Bernoulli, 1713, en latin) o ((E. D. Sylla, Translator), 1713; DeMoivre, 1756) y (Hald, 1990, Cap. 13, 15 & 22).

[2] Un paso importante es debido a Kolmogorov en 1933 que se apoyó sobre trabajos de Richard von Mises (von Mises, 1932) y también sobre la teoria de la medida y de la integraciíon debido entre otros a Emile Borel y Henri-Léon Lebesgues (Borel, 1898, 1909; Lebesgue, 1904, 1918; Halmos, 1950) para formalizar analíticamente la teoria de las probabilidades (Kolmogorov, 1956; Barone & Novikoff, 1978; Jacob & Protters, 2003).

Eso es ilustrado en la figura Fig. 1-1. La unión e intersección satisfacen a las mismas reglas que en la teoria ensemblista, es decir cada una es comutativa $A \cup B = B \cup A,\ A \cap B = B \cap A,$ asociativa $(A \cup B) \cup C = A \cup (B \cup C),\ (A \cap B) \cap C = A \cap (B \cap C),$ distributiva con respeto a la otra $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ y $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (ver ej. (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)).



(a) (b) (c) (d)

Figura 1-1: Ilustración de la operaciones de unión $A \cup B$ (a), intersección $A \cap B$ (b), complemento $\bar{A}$ (c), enventos excluyentes $A \cap B = \emptyset$ (d). A es representado en linea llena, B en linea discontinua; (a)-(c) el resultado de la operación es la zona en grise. A veces, esta representación ensemblista se denota *diagrama de Venn o de Euler*.

Formalmente, se define de manera abstracta un espacio medible $(\Omega, \mathcal{A})$ de la manera siguiente (Halmos, 1950; ?, ?; Feller, 1971; Brémaud, 1988; Ibarrola et al., 1997; 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 de un espacio muestral* Ω *y una colección* $\mathcal{A}$ *de conjuntos de* Ω *es llamado espacio medible si satisface a los requisitos*

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

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

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

Las propiedades de la probabilidad P de un dado evento quedan determinadas por los siguientes (ej. (Spiegel, 1976; Kolmogorov, 1956; Shafer & Vovk, 2006; von Plato, 2005)):
Axiomas de Kolmogorov

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

3. $P(\Omega) = 1$

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

Definición 1-2 (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* $\{A_i\}$ *de elementos mutualmente 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 acotada por arriba,* $\mu(\Omega) < +\infty$*, la medida es dicha finita y el espacio tambiń es dicho finito. Además, si*

$$P \equiv \mu, \quad P(\Omega) = 1,$$

la medida es dicha medida de probabilidad, $\mu \equiv P$*. En este caso, el espacio* $(\Omega, \mathcal{A}, P)$ *es llamado espacio probabilístico.*

(ver también (Kolmogorov & Fomin, 1961, Cap. 5 & 6)).

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: $P(\emptyset) = 0$;
- 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 $\{A_i\}_{i=1}^{n}$ es dicho *conjunto completo de eventos posibles excluyentes entre sí* y es ilustrado figure 1-2;
- si A es subconjunto de B, lo que escribiremos $A \subset B$, es decir si B se realiza, A se realiza también (pero no necesariamente al revés), entonces $P(A) \leq P(B)$; Es ilustrado figure 1-2;

Nota: 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: $0 \leq P(A \cap B) \leq \text{mín}\{P(A), P(B)\}$; (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 $\{A \cap B_j\}$ mutuamente excluyentes y $\bigcup_j (A \cap B_j) = A \cap \left(\bigcup_j B_j\right) = A \cap \Omega = A$).

(a) (b)

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

En el caso de eventos no necesariamente mutuamente excluyentes, se prueba que la *ley de composición* o *formula 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 de 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

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

Similarmente, 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

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

Se puede preguntarse de cual es la probabilidad de un evento A, si sabemos que tenemos un evento B, dado. Por ejemplo, por un dado de 6 caras equilibriado, cual es la probabilidad de tener un número par sabiendo que tenemos un numero menor a igual a 3. La respuesta es en la noción de *probabilidad condicional* (Hausdorff, 1901; Jeffrey, 1948, 1973; Brémaud, 1988; Mandel & Wolf, 1995; Jacob & Protters, 2003; Shafer & Vovk, 2006):

Definición 1-3 (Probabilidad condicional)**.** *Por definición, la probabilidad condicional de A dado B es la razón entre la probabilidad del evento conjunto y la probabilidad de que se dé B (cuando éste es un evento no nulo):*

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

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

Es fácil demostrar que esta cantidad toma valores entre 0 y 1, con $P(\Omega|B) = 1$, y que es aditiva para una unión de eventos mutuamente excluyentes referidos al cumplimiento de B. Luego, $P(A|B)$ es una medida de probabilidad [3]; Por eso, a veces en la literatura se la denota $P_B(A)$. Diversas situaciones de probabilidades condicionales son ilustradas en la figura siguiente, Fig. 1-3.

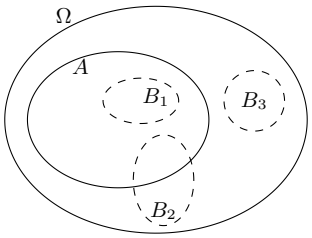


Figura 1-3: Ilustración de la probabilidad condicional con A interior del elipse en linea llena y unos B_i interiores de los elipses en lineas discontinuas. $\omega \in B_1 \Rightarrow \omega \in A$ así que $P(A|B_1) = 1$. Al revés, $\omega \in B_3 \Rightarrow \omega \notin A$ así que $P(A|B_3) = 0$. Entre estas situaciones extremas, si $P(\bar{A} \cap B_2) \neq 0$ y $P(A \cap B_2) \neq 0$ tenemos $0 < P(A|B_2) < 1$ (ej. con probabilidades iguales a las superficias relativas de los conjuntos).

Algunas propiedades interesantes son las siguientes:

- condición de normalización: $\sum_{i=1}^{n} P(A_i|B) = 1$, siendo $\{A_i\}_{i=1}^{n}$ un conjunto completo de resultados posibles mutuamente excluyentes;
- 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;
- *fórmula de probabilidades totales*: si $\{B_j\}$ es un conjunto completo de eventos no nulos mutuamente excluyentes, entonces
$$P(A) = \sum_j P(A|B_j) P(B_j)$$
Viene de $A = A \cap \left(\bigcup_j B_j\right) = \bigcup_j (A \cap B_j)$ donde los $A \cap B_j$ son mutuamente excluyentes, y $P(A \cap B_j) = P(A|B_j)P(B_j)$.
- *fórmula de Bayes*: si $\{B_j\}$ es un conjunto completo de eventos no nulos mutuamente excluyentes, entonces
$$P(B_i|A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}.$$
(ver (Brémaud, 1988; Jacob & Protters, 2003; Bayes, 1763; Barnard, 1958)).

Terminamos esta sección por la noción de independencia entre dos eventos. Por ejemplo, si dos dados son tirado sobre dos mesas diferentes, no hay ninguna razón de que la muestra de uno “influye” la del otro. Dicho de otra manera, dos eventos son independientes si conciendo uno no lleva ninguna “información” sobre el otro (Brémaud, 1988; Mandel & Wolf, 1995; Hausdorff, 1901; Jacob & Protters, 2003; Borel, 1909):

[3] Se puede definir un espacio de probabilidad $(\Omega_B, \mathcal{A}_B, P_B)$ donde $P_B(A) \equiv P(A|B)$.

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 estadísticamente *mutuamente* 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. Por ejemplo, tiramos 2 dados independientemente y consideramos los eventos A_i: el dado i es par y A_3 la suma es impar. Es claro de que A_1 y A_2 son independientes y ademas $P(A_1 \cap A_3) = \frac{1}{4} = P(A_1)P(A_3)$ (es tener par y impar), mientras que $P(A_1 \cap A_2 \cap A_3) = 0 \neq \frac{1}{8}$: los eventos son independientes pares, pero no son mutuamente independientes (Hogg et al., 2013).

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 o continua.*

Formalmente, la noción de variable aleatoria se apoya sobre la noción de función medible (Kolmogorov & Fomin, 1961; Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

Definición 1-5 (Función medible)**.** *Sean $(\Omega, \mathcal{A})$ y $(\Upsilon, \mathcal{B})$ dos espacios medibles. Una función $f : \Omega \mapsto \Upsilon$ es dicha $(\mathcal{A}, \mathcal{B})$-medible si*

$$\forall\, B \in \mathcal{B}, \quad A \equiv f^{-1}(B) = \{\omega \in A : \ f(\omega) \in B\} \ \in \mathcal{A}$$

Dicho de otra manera, la pre-imagen de un elemento dada de $\mathcal{B}$ (elemento medible) partenece a $\mathcal{A}$ (elemento medible). A veces, se dice más simplemente de que $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ es medible por abuso de escritura.

Además, saliendo de un espacio de medida y una función f medible, se puede definir una medida imagen sobre el espacio de llegada (Athreya & Lahiri, 2006; Bogachev, 2007a; Cohn, 2013):

Teorema 1-1 (Teorema de la medida imagen)**.** *Sean $(\Omega, \mathcal{A}, \mu)$ un espacio de medida, $(\Upsilon, \mathcal{B})$ un espacio medible y una función $f : (\Omega, \mathcal{A}) \mapsto (\Upsilon, \mathcal{B})$ 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). μ_f es dicha medida imagen de μ por f.

Demostración. Por definición, claramente $\mu_f \geq 0$ y por definición de una función, $f^{-1}(\emptyset) = \emptyset$ dando $\mu_f(\emptyset) = \mu(\emptyset) = 0$. Luego, si para un conjunto numerable $\{B_j\}$ de elementos de $\mathcal{B}$ disjuntos entre sí, las preimagenes de los B_j son disjuntos también entre sí (para $k \neq j$ no se puede tener $\omega \in f^{-1}(B_j) \cap f^{-1}(B_k)$ si no ω tendría dos imagenes distinctas por f). Entonces $f^{-1}\left(\bigcup_j B_j\right) = \bigcup_j f^{-1}(B_j)$. Eso implica de que $\mu_f\left(\bigcup_j B_j\right) = \mu\left(f^{-1}\left(\bigcup_j B_j\right)\right) = \bigcup_j \mu\left(f^{-1}(B_j)\right) = \sum_j \mu\left(f^{-1}(B_j)\right) = \sum_j \mu_f(B_j)$. Finalmente, necesariamente $f^{-1}(\Upsilon) = \Omega$ (es incluida y $f(\Omega)$ siendo en Υ son necesariamente iguales) lo que cierra la prueba [4]. □

Un espacio jugando un rol particular es $\mathbb{R}$, a lo cual se puede asociar $\mathcal{B}(\mathbb{R})$ la σ-álgebra más pequeña generada por los intervalos $(-\infty\,;\,b]$ (equivalentemente, por los abiertos de $\mathbb{R}$, o también por los intervalos $(a\,;\,b]$), *i. e.,* uniones numerables, intersecciones numerables, complementos de estos intervalos (Athreya & Lahiri, 2006; Bogachev, 2007a, 2007b; Cohn, 2013). $\mathcal{B}(\mathbb{R})$ es llamada *Borelianos de* $\mathbb{R}$ o *σ-álgebra de Borel de* $\mathbb{R}$.

Con estas definiciones, tenemos todo lo necesario para introducir la definición de una variable aleatoria real (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988):

Definición 1-6 (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 los eventos

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

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 de que los numeros serán la imagen de Ω por X (ej. $X(\omega_j) = j, \quad j = 1, \ldots, 6$).

Fijense de 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 $(-\infty\,;\,b]$. Eso da lugar a la definición de la función de repartición (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

---

[4] De hecho, se puede sencillamente probar que la preimagen de una unión numerable (que sean disjuntos o no) es la unión de las preimagenes; lo mismo occure para la intersección y además la preimagen del complemento es el complemento de la preimagen. Eso es conocido como *leyes de de Morgan* (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 (Función de repartición). *Por definición, la función de repartición F_X de una variable aleatoria es definida por*

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

A veces, por abuso de terminologia, se denomina F_X como ley de la variable aleatoria. Se encuentra también el la literatura la terminologia de función cumulativa (cdf por cumulative density function en inglés).

Naturalmente, de las propiedades de una medida de probabilidad,

- $0 \leq F_X(x) \leq 1$;
- $\lim_{x\to-\infty} F_X(x) = 0$ 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 (viene de que $x_1 \leq x_2 \Rightarrow (-\infty\,;\,x_1] \subseteq (-\infty\,;\,x_2]$);
- F_X no es necesariamente continua (lo vamos a ver más adelante), pero en cada punto x es continua a su derecha (ver punto 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 $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ donde $\cdot^t$ denota la transpuesta, caracterizado por d-uplas de variables aleatorias reales. Como en el caso univariado, se define este vector de la manera siguiente: (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988)

Definición 1-8 (Vector aleatorio real). *Un variable 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 del vector aleatorio X. Como en el caso escalar,

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

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 $(-\infty\,;\,b_1] \times \cdots \times (-\infty\,;\,b_d]$, *i. e.,* la función de repartición multivariada (Athreya & Lahiri, 2006; Cohn, 2013; Brémaud, 1988; Hogg et al., 2013):

Definición 1-9 (Función de repartición multivariada). *Por definición, la función de repartición F_X de un vector aleatorio es definida en $x = (x_1, \ldots, x_d)$ por*

$$F_X(x) = P_X((-\infty\,;\,x_1] \times \cdots \times (-\infty\,;\,x_d]) = P\left(\bigcap_{i=1}^{d}(X_i \leq x_i)\right)$$

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

- $0 \leq F_X(x) \leq 1$;

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

Al final, 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 de que

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

(viene de que $\bigcap_{j=1}^k (X_{i_j} \leq x_{i_j}) = \left(\bigcap_{j=1}^k (X_{i_j} \leq x_{i_j})\right) \bigcap \left(\bigcap_{i \notin I_k} (X_i \in \mathbb{R})\right)$). Esta función es dicha *función de repartición marginale* de F_X.

Cerramos estas generalidades con el caso de variables independientes:

Definición 1-10 (Independencia)**.** *Sean d variables aleatorias X_i y $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$. Los X_i son mutualmente independientes si y solamente si, para cualiquier ensemble de conjuntos B_i, los eventos $(X_i \in \mathcal{B}_i)$ son mutualmente independientes, i. e.,*

$$P_X(\times_{i=1}^d B_i) = \prod_{i=1}^d P_{X_i}(B_i)$$

donde $\times$ denota el producto cartesiano entre elementos de $\mathcal{B}(\mathbb{R})$. Es equivalente a

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

La ley del vector aleatorio se factoriza.

Es importante notar de que no es equivalente a tener la independencia por pares, como ilustrado en el fin de la seción precediente.

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. La meta de las susecciones siguientes es estudiar las particularidades de cada caso.

Par fijar unas 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 razones de simplificaciones, se considera $\mathcal{X}$ como siendo el espacio muestral y se olvida de 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 preimagen.

1.3.1 Variable aleatoria discreta

Definición 1-11 (Variable aleatoria discreta)**.** *Una variable aleatoria es dicha discreta cuando $\mathcal{X} = X(\Omega)$ es discreto, finito o infinito numerable. En lo que sigue, denotaremos por $|\mathcal{X}|$ el cardinal de $\mathcal{X}$, posiblemente infinito.*

En otras palabras, los posibles valores de una variable aleatoria discreta X consisten en un conjunto contable (finito o infinito numerable) de números reales: $\mathcal{X} = \{x_j\}_j$ (Athreya & Lahiri, 2006; Hogg et al., 2013). Fijense de que Ω no es necesariamente discreto. Por ejemplo, si ω es la posición de un punto sobre una linea, y $X(\omega) = 0$ si ω es a la izquierda de un umbral, y $X(\omega) = 1$ si ω es a su derecha, $\mathcal{X} = \{0\,;\,1\}$ mientras de que Ω no es discreto.

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

Definición 1-12 (Función de masa de probabilidad)**.** *Por definición, la función de masa de probabilidad de X, variable aleatoria discreta tomando sus valores sobre $\mathcal{X}$ es dada por*

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

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

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

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

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

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

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

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

lo que da, tratando de 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 inmediato de que F_X es una función discontinua, con saltos finitos (en x_j, salto de altura $p_X(x_j)$). Eso es ilustrado figura Fig. 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: $p_X(x) = \mathbb{1}_{\{x_k\}}(x)$ o $p_X = \mathbb{1}_k$, donde

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{si} \quad x \in A \\ 0 & \text{si no} \end{cases}$$

es la función indicator y $\mathbb{1}_k$ es el vector (posiblemente de dimensión infinita) de componentes k-esima $\delta_{jk} = \mathbb{1}_{\{k\}}(j)$ símbolo *delta de Kronecker*,

$$\delta_{jk} = \begin{cases} 1 & \text{si } j = k \\ 0 & \text{si no} \end{cases}$$

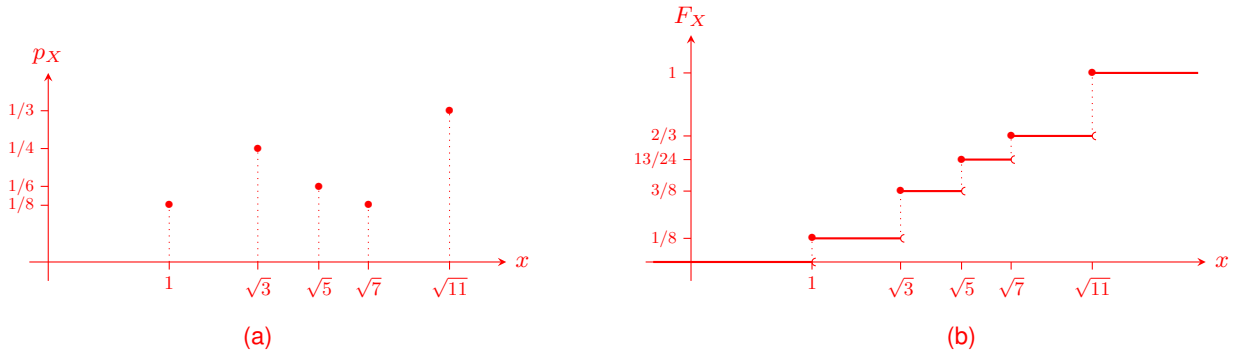


Figura 1-4: Ilustración de una distribución de probabilidad discreta (a), y la función de repartición asociada (b), con $\mathcal{X} = \{1, \sqrt{3}, \sqrt{5}, \sqrt{7}, \sqrt{11}\}$ y $p_X = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{6} & \frac{1}{8} & \frac{1}{3} \end{bmatrix}^t$.

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

Reordenamiento y relación de mayorización : a ver como trasladar lo que ya estaba en en cap. 2 mas los ejemplos. Todo comentado por ahora

1.3.2 Variable aleatoria continua

En varios contextos, puede tomar valores en un conjunto no numerable, por ejemplo cualesquiera de los números en un dado intervalo de la recta real. No son variables discretas más. En las variables que no son discretas, el caso particular de interés es el de variables continuas (Athreya & Lahiri, 2006; Hogg et al., 2013):

Definición 1-13 (Variable aleatoria continua)**.** *Una variable aleatoria X es dicha continua si su función de repartición F_X es continua sobre $\mathbb{R}$ (on respeto a la medida de Lebesgue).*

Cuando se puede, es conveniente asociar una *función densidad de probabilidad* (comúnmente anotada por su sigla en inglés: pdf por *probability density function*):

Definición 1-14 (Variable aleatoria continua admitiendo una densidad de probabilidad)**.** *Sea X variable aleatoria continua y P_X su medida de probabilidad y F_X su función de repartición. Si existe una función no negativa p_X sobre $\mathbb{R}$ tal que*

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

(la integral debe ser entendido como en el sentido de Lebesgue [5]), entonces X es dicha admitiendo una

[5] **Decir un poquito sobre la medida de Lebesgue y ref. (Lebesgue, 1904, 1918; Kolmogorov & Fomin, 1961; Athreya & Lahiri,**

densidad y p_X es llamada densidad de probabilidad de X. En particular,

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

Dicho de otra manera, si F_X es (continua y) derivable sobre $\mathbb{R}$ (con respeto a la medida de Lebesgue), por lo menos por partes, X admite una densidad de probabilidad y

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

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

Fijense de que $\mathcal{X}$ es el soporte de p_X, i. e., $p_X(\mathcal{X}) \neq 0$. En lo que sigue, denotaremos $|\mathcal{X}|$ el volumen (o medida de Lebesgue) de $\mathcal{X}$, posiblemente infinito.

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 claramente

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

$\{x\}$ es dicho de medida P_X nula (es el caso de todos conjuntos numerable de $\mathbb{R}$). Fijense de que si $0 \leq F_X \leq 1$, no 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, $p_X(x) = 2\mathbb{1}_{[\frac{1}{2}\,;\,+\infty)}(x)$. No es contradictorio en el sentido de que p_X no es una probabilidad, sino que $p_X(x)\,dx$ es esquematicamente la probabilidad de hallar a la variable con valores en el "intervalo infinitesimal entre x y $x + dx$". Al final, la condición de normalización se escribe

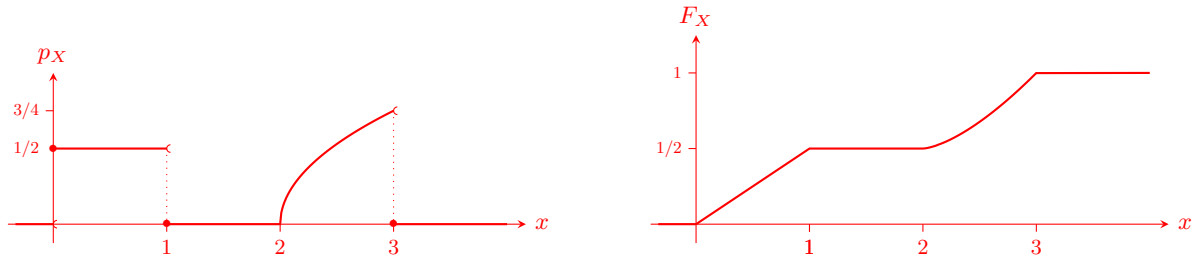
$$\int_{\mathcal{X}} p_X(x)\,dx = \int_{\mathbb{R}} p_X(x)\,dx = 1.$$

En la figura Fig. 1-5-(a) se muestra una representación gráfica de una función densidad de probabilidad para una variable continua y en Fig. 1-5-(b) la función cumulativa correspondiente.

Fijense de que una variable aleatoria puede ser ni continua, ni discreta:

- Sean U y V variables continuas, independientes, de densidad de probabilidad $p_U = p_V = \mathbb{1}_{[0\,;\,1)}$ (U y V son dichas uniformes sobre $[0\,;\,1)$) y sea $X = V\mathbb{1}_{U<\frac{1}{2}}$, es decir $X(\omega) = V(\omega)$ si $U(\omega) < \frac{1}{2}$ y 0 si no. Entonces de la formula 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.*, $F_X(x) = \frac{1}{2}P\left((V \leq x)\,\middle|\,(U < \frac{1}{2})\right) + P\left((0 \leq x)\,\middle|\,(U \geq \frac{1}{2})\right)$. Ahora, de la independencia de U y V, tenemos $F_X(x) = \frac{1}{2}F_V(x) + \frac{1}{2}\mathbb{1}_{\mathbb{R}_+}(x) = \frac{x+1}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{[1\,;\,+\infty)}(x)$. Esta función de repartición es representada figura Fig. 1-6: es ni discreta, ni continua. Entonces, a pesar de que $\mathcal{X} = [0\,;\,1)$ sea un intervalo, X no es continua (y tampoco no puede ser discreta).

2006; Bogachev, 2007a; Cohn, 2013).



(a) (b)

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

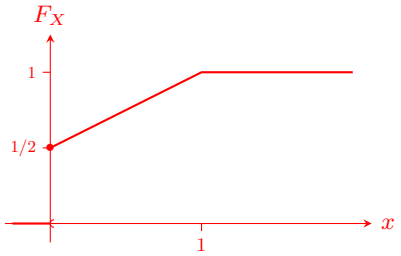


Figura 1-6: Función de repartición $F_X(x) = \frac{x}{2}\mathbb{1}_{[0\,;\,1)}(x) + \mathbb{1}_{\mathbb{R}_+}(x)$ asociada a $X = V\mathbb{1}_{U<\frac{1}{2}}$ con U y V variables continuas uniformes sobre $\mathcal{X} = [0\,;\,1)$. No es tipo escalon, así que X no es discreta. A pesar de que $\mathcal{X} = [0\,;\,1)$ sea un intervalo, de la presencia del salto en $x = 0$, tampoco X no es continua.

- Sea U variable continua uniforme sobre $[0\,;\,1)$ y $X = \mathbb{1}_{U\notin\mathbb{Q}}$. Claramente X no es continua, pero $\mathcal{X} = [0\,;\,1)\backslash\mathbb{Q}$ siendo no numerable, X tampoco es discreta [6].

De hecho, en el caso continu, discreto, o cualquiera, se conserva la forma integral de la medida de probabilidad en una forma tipo $P_X(B) = \int_B dP_X(x)$ basado sobre la teoria de la medida y de la integración (Lebesgue, 1904; Kolmogorov & Fomin, 1961; Athreya & Lahiri, 2006; Bogachev, 2007a, 2007b; Cohn, 2013). En el caso discreto cierto $X = x_0$, la distribución discreta es dada por $p_X(x) = \mathbb{1}_{\{x_0\}}(x) = \mathbb{1}_{\{0\}}(x - x_0)$. En este caso, P_X es dicha *medida de Dirac* y $dP_X(x)$ es denotado $\delta_{x_0}(x)$ o $\delta(x - x_0)$ también llamado *delta de Dirac*. Se puede ver este Dirac como una densidad de probabilidad $p_X(x)$ pero no es una función "ordinaria" pero una función generalizada o distribución de Schwartz [7]. En particular, $F_X(x) = \mathbb{1}_{\mathbb{R}_+}(x - x_0)$ y en el sentido de las

[6] $X = 1$ **casi siempre...**

[7] La teoría de la distribuciones valió a Laurent Schwarz la meddala Field en 1950. Entre otros en el trabajo de Schwartz, se probó que el Dirac, visto como distribución de Schwartz, o función generalizada, tiene una "representación integral" $\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}dt$ o más rigurosamente transformada de Fourier de $x \mapsto 1$ en el sentido de las funciones generalizadas o distribuciones. Eso muestra claramente su caracter no ordinario (la integral siendo divergente en el sentido usual). Eso va más allá de la meta del capítulo y el lector se podrá

distribuciones, $\frac{dF_X}{dx} = \delta_{x_0}$. Además, se usan en general las propiedades, para cualquier function f,

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

pero hay que entender la integración a través de la medida Dirac (esta notación es un abuso de escritura, ej. (Gel'fand & Shilov, 1964)).

Usando las medidas de Dirac, se puede unificar el tratamiento de las variables aleatorias discretas con las continuas (entre otros): si una variable aleatoria discreta toma los valores x_j con probabilidades $P(X = x_j)$ respectivamente, entonces formalmente se puede describir mediante una variable aleatoria continua X con "función densidad de probabilidad" $p_X(x) = \sum_j p_j\,\delta(x - x_j)$ donde $p_j = P(X = x_j)$.

1.3.3 Vector aleatorio discreto

Un ejemplo de vector aleatorio discreto puede verse a través de un conjunto de dados (que podrían ser dependientes si son ligados por un hilo por ejemplo).

Definición 1-15 (Vector aleatorio discreto)**.** *Un vector aleatorio d-dimensional $X = \begin{bmatrix} X_1 & \cdots & X_d \end{bmatrix}^t$ y $\mathcal{X} = X(\Omega) = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ donde $\mathcal{X}_i = X_i(\Omega)$. X es dicho discreto cuando $\mathcal{X} \subseteq \mathbb{N}^d$, es discreto, finito o infinito numerable. En lo que sigue, denotaremos también por $|\mathcal{X}|$ el cardinal de $\mathcal{X}$, posiblemente infinito.*

Obviamente, la medida de probabilidad en los $x = (x_1\,,\,\ldots\,,\,x_d) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ caracterisa completamente este vector aleatorio:

Definición 1-16 (Función de masa de probabilidad conjunta)**.** *Por definición, la función de masa de probabilidad de X, vector aleatorio discreto tomando sus valores sobre $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ es dada por*

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

Se la llama también función de masa de probabilidad conjunta de los X_i, o, por abuso de denominación, la llamaremos todavía p_X distribución de probabilidad (conjunta). En el caso multivariado, la notación vectorial es más delicada a usar: p_X sería un "tensor" d-dimensional (una matriz para $d = 2$,...). Pero queda posible unsar una notación vectorial, recordandose de que $\mathbb{N}^d$ puede ser en biyección con $\mathbb{N}$ y una biyección elegida, usar la para etiquetar los componentes de p_X puesto 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 $j = \sum_{i=1}^{d-1}(j_i - 1)\prod_{k=i+1}^{d}\alpha_k + j_d$-esima componente del vector p_X.

Como en el caso escalar, la función de repartición de un vector aleatorio discreto d-dimensional es echo de hyperplanos d-dimensionales constantes. Además, las componentes son mutualmente independientes si y

---

referir a (Schwartz, 1966; Gel'fand & Shilov, 1964, 1968) por ejemplo.

solamente si la función de répartición se factoriza, o equivalentemente la función de masa se factoriza, *i. e.,*

$$X_i \text{ mutualmente independientes} \quad \Leftrightarrow \quad p_X = p_{X_1} \otimes \cdots \otimes p_{X_d}$$

donde $\otimes$ denota el producto tensorial o externo, $p_{X_j} \otimes p_{X_k}(x_j, x_k) = p_{X_j}(x_j) p_{X_k}(x_k)$. Cuando los α_i son finito y la notación vectorial de la definición es adoptada, esta expresión queda valide donde $\otimes$ representa el producto de Kronecker [8]

Al final, de la formula de calculo de función de repartición marginales visto pagina 24, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1, \ldots, d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ la probabilidad marginal o distribución marginale de X_{I_k} es dada por

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

1.3.4 Vector aleatorio continuo

Definición 1-17 (Vector aleatorio continuo y densidad de probabilidad multivariada)**.** *Un vector aleatorio X es dicho continuo si su función de repartición F_X es continua sobre $\mathbb{R}^d$ (on respeto a la medida de Lebesgue). Si existe una función no negativa p_X sobre $\mathbb{R}^d$ tal que*

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

(la integral multiple debe ser entendido como en el sentido de Lebesgue [9] y $dx = dx_1 \cdots dx_d$), entonces X es dicha admitiendo una densidad y p_X es llamada densidad de probabilidad de X, o también densidad de probabilidad conjunta de los X_i. En particular,

$$F_X(x) = \int_{(-\infty\,;\,x_1]\times\cdots\times(-\infty\,;\,x_d]} p_X(u)\, du$$

o, equivalentemente, para F_X es (continua y) derivable sobre $\mathbb{R}^d$ (con respeto 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 todavía la terminología (por abuso) de distribución de probabilidad; $\mathcal{X}$ es el soporte de p_X, denotaremos todavía $|\mathcal{X}|$ el volumen (o medida de Lebesgue) de $\mathcal{X}$, posiblemente infinito.

---

[8] Para $p = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix}^t$ y $q = \begin{bmatrix} q_1 & \cdots & q_m \end{bmatrix}^t$ el producto de Kronecker es dado por $p \otimes q$ vector de tamaño nm de componente $(j-1)m + k$-esima el producto $p_j q_k$, $1 \leq j \leq n$, $1 \leq k \leq m$. Fijense de que este producto es asociativo pero no es comutativo.

[9] Ver nota de pie 5, pagina 27.

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

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

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

$$X_i \text{ mutualmente independientes} \quad \Leftrightarrow \quad p_X(x) = p_{X_1}(x_1) \ldots p_{X_d}(x_d)$$

Terminamos esta sección mencionando que, de la formula de calculo de función de repartición marginales visto pagina 24, para un subconjunto $I_k = (i_1, \ldots, i_k)$ de $1 \leq k \leq d$ elementos de $\{1, \ldots, d\}^k$, $X_{I_k} = \begin{bmatrix} X_{i_1} & \cdots & X_{i_k} \end{bmatrix}^t$ la *densidad de probabilidad marginal* de X_{I_k} es dada por

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

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.

Reordenamiento y relación de mayorización : a ver como trasladar lo que ya estaba en en cap. 2 mas los ejemplos.

1.3.5 Transformación de variables y vectores aleatorios

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

Teorema 1-2 (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 es 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. g siendo medible, 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) asociado al espacio de salida de g, el resultado es consecuencia del teorema 1-1. □

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

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

No se todavía si será útil tratar del caso de limite de series de funciones medibles (quizas, tratando de los momentos/integración).

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

Teorema 1-3 (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 es relacionada a 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 de 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 $\mathcal{X}$ pueder no discreto.

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

Lema 1-1 (Continuidad y caracter 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 recordandose de 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 al 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 añandido $d - d'$ transformaciones. Por ejemplo, con $d = 2$ si nos interesamos a $X_1 + X_2$, se puede considerar $\begin{bmatrix} X_1 + X_2 & X_2 - X_1 \end{bmatrix}^t$ y llegar a la variable de interés por calculo 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 admitiendo una densidad de probabilidad, una pregunta natural es entonces de saber si se conserva la continuidad y la existencia de una densidad, así que su forma. La respuesta es dada por el teorema siguiente (Brémaud, 1988; Jacob & Protters, 2003; Athreya & Lahiri, 2006; Cohn, 2013; Hogg et al., 2013):

Teorema 1-4 (Densidad de probabilidad por transformación continua inyectiva diferenciable)**.** *Sean X, vector aleatorio d-dimensional continuo y admitiendo una densidad de probabilidad p_X, $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ una función continua, inyectiva y diferenciable tal que $|\mathrm{J}_g| > 0$, donde J_g denota la matriz de componentes $\frac{\partial g_i}{\partial x_j}$, matriz Jacobiana de la transformación $g \equiv \begin{bmatrix} g_1(x_1, \ldots, x_d) & \cdots & g_d(x_1, \ldots, x_d) \end{bmatrix}^t$) y $|\cdot|$ representa el valor absoluto del determinante de la matriz. Sea $Y = g(X)$. Entonces Y es continua admitiendo una densidad de probabilidad p_Y 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, X admitiendo una densidad y g siendo medible,

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

Por cambio de variable $x = g^{-1}(y)$ (g siendo inyectiva, el antecedante es único por definicón),

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

lo que cierra la prueba. □

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

Corolario 1-1. *Sean X, variable aleatoria continua y admitiendo 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 admitiendo una densidad de probabilidad p_Y tal que*

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

Una forma alternativa de derivar este resultado es partir de la función de repartición, notando de que g es necesariamente monotona:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \begin{cases} P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) & \text{si } \ g \ \text{ es creciente} \\ P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{si } \ g \ \text{ es decreciente} \end{cases}$$ [10]

y calcular 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, particiendo $\mathbb{R}^d$ en conjuntos donde g es inyectiva, dando

Teorema 1-5. *Sean X, vector aleatorio d-dimensional continuo y admitiendo una densidad de probabilidad p_X, $g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \mapsto (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ 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 todos $k \geq 1$, $g : \mathcal{X}_{[k]} \mapsto \mathcal{Y}$ sea inyectiva y tal que $|\mathrm{J}_g(y)| > 0$. Suponemos de que $\mathcal{X}_{[0]}$ sea de medida de Lebesgue nula, notamos g_k^{-1} la función inversa de g sobre $g(\mathcal{X}_{[k]})$ (rama k-esima de la función multivaluada g^{-1}), $\mathrm{J}_{g_k^{-1}}$ su matriz Jacobiana y $I(y) = \{k,\, y \in g(\mathcal{X}_{[k]})\}$ los indices tales que y tiene un inverso por g_k. Eso es ilustrado figura Fig. 1-7 para $d = 1$. Entonces Y es continua admitiendo 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 $d = 1$ eso se formula

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

Demostración. Sufice escribir $B = \bigcup_{k=0}^m (B \cap g(\mathcal{X}_k))$ unión de borelianos disjuntos, notar de que por consecuencia $g^{-1}(B) = \bigcup_{k=0}^m g^{-1}(B \cap g(\mathcal{X}_k))$ unión de borelianos disjuntos y por linearidad escribir la integración sobre $g^{-1}(B)$ como la suma de integrales sobre $g^{-1}(B \cap g(\mathcal{X}_k))$. Se cierra la prueba notando de que $g^{-1}(B \cap g(\mathcal{X}_0))$ es necesario de medida de Lebesgue nula, siendo la integral nula y de que $g^{-1}(B \cap g(\mathcal{X}_k)) = g_k^{-1}(B \cap g(\mathcal{X}_k))$. □

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

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

(Ver figura 1-7).

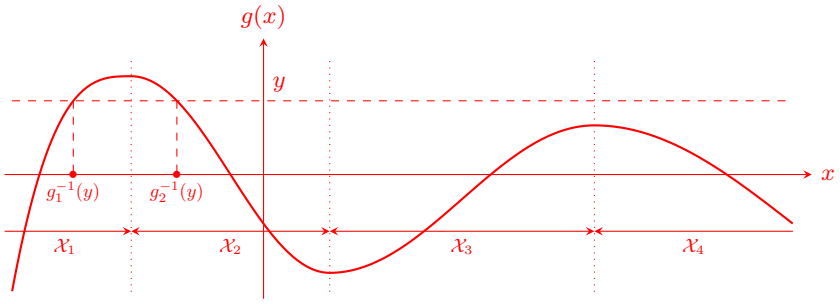


Figura 1-7: (a): Ilustración de una transformación g no inyectiva, tal que $\mathcal{X}_{[0]} = \{x,\, g'(x) = 0\}$, representado por las lineas punteadas (x correspondiente), es de medida de Lebesgue nula. Los $\mathcal{X}_{[k]}$ son descrito debajo de cada dominio. La linea discontinua da un nivel y y los puntos en el eje x representan $g_k^{-1}(y),\ k \in I(y)$; en el ejemplo, $I(y) = \{1\,;\, 2\}$.

Es importante notar de que la condición $\mathcal{X}_{[0]}$ de medida nula es importante. El el caso contrario, Y no queda continua. Por ejemplo, considera X uniforme sobre $\mathcal{X} = (3\,;\, 3)$ y $Y = g(X)$ 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 es representado figura Fig. 1-8-(a). Claramente, g es continua y diferenciable sobre $\mathcal{X}$, pero con $\mathcal{X}_{[0]} = [-2\,;\, 1]$ que no es de medida nula. Saliendo de $F_Y(y) = P(g(X) \leq y)$ se calcula sencillamente $F_Y(y) = \frac{2}{3}\left(1 - \frac{1}{\pi}\operatorname{acos}(y-1)\right) \mathbb{1}_{[0\,;\,2)} + \mathbb{1}_{[2\,;\,+\infty)}(y)$, ilustrada figura Fig. 1-8-(b). Claramente F_Y es discontinua en $y = 2$: Y no es continua.

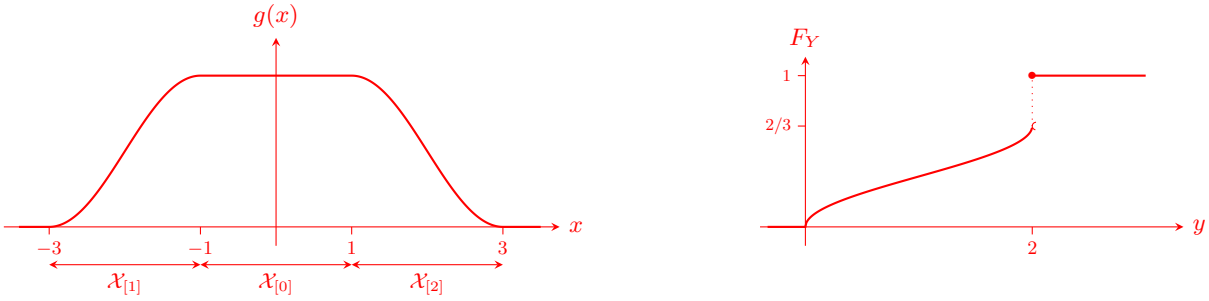


Figura 1-8: En (a) se dibuja $g(x) = \left(1 + \cos\left((1 - |x|)\frac{\pi}{2}\right)\right) \mathbb{1}_{(1\,;\,3)}(|x|) + 2\mathbb{1}_{[0\,;1]}(|x|)$. Suponiendo de que $\mathcal{X} = (-3\,;\, 3)$, claramente $\mathcal{X}_{[0]} = [-1\,;\, 1]$ no es de medida nula, dando para X uniforme sobre $\mathcal{X}$ la variable $Y = g(X)$ no continua de función de repartición representenda en (b).

Sea X una variable aleatoria (continua, en general) definida en el intervalo $[x_m, x_M]$ con función densidad de probabilidad $p(x)$. Sea $Y = \Psi(X)$ una función real de X, luego Y toma los valores $y = \Psi(x)$ en el intervalo

$[y_m, y_M]$. La función densidad de probabilidad $q(y)$ para la variable aleatoria transformada Y se obtiene de la siguiente manera, dependiendo de la forma de la transformación:

- Si Ψ es inversible, con inversa (única), se tiene $x = \Phi(y)$, con $\Phi = \Psi^{-1}$. A partir de la propiedad de conservación de la probabilidad

  $$|q(y)\, dy| = |p(x)\, dx|$$

  para una correspondencia biunívoca entre x e y, se obtiene la pdf transformada

  $$q(y) = p(x) \left| \frac{dx}{dy} \right| = p\left(\Phi(y)\right)\ |\Phi'(y)| = \frac{p\left(\Phi(y)\right)}{|\Psi'(\Phi(y))|}.$$

  Una forma alternativa de derivar este resultado es partir de la función de repartición:

  $$F_Y(y) = P(Y \leq y) = P(\Psi(X) \leq y) = P(X \leq \Psi^{-1}(y)) = F_X(\Phi(y))$$

  y calcular las derivadas del primer y último términos respecto de la variable transformada y.

- Si la inversa de Ψ es multivaluada, cada valor de y se corresponde con un conjunto de valores de x, digamos $\{x_k = \Phi_k(y),\ k = 1, 2, \ldots\}$. Debido a que estas soluciones son mutuamente excluyentes, las probabilidades se suman, de modo que

  $$q(y) = \sum_k p(x_k) \left| \frac{dx_k}{dy} \right| = \sum_k \frac{p\left(\Phi_k(y)\right)}{|\Psi'(\Phi_k(y))|},$$

  que formalmente se puede expresar como $q(y) = \int p(x)\, \delta(y - \Psi(x)\,)\, dx$, donde se usa la expansión de la función delta en términos de sus ceros: $\delta(y - \Psi(x)\,) = \sum_k \delta(x - x_k)/|\Psi'(x_k)|$.
  Por ejemplo, para la transformación de variables $Y = X^2$ se tiene $Y = \Psi(X) = X^2$ cuyas inversas son $X_1 = \Phi_1(Y) = +\sqrt{Y}$ y $X_2 = \Phi_2(Y) = -\sqrt{Y}$; luego $q(y) = \frac{p(\sqrt{y})}{2\sqrt{y}} + \frac{p(-\sqrt{y})}{|-2\sqrt{y}|}$, para $y > 0$.

Consideramos ahora el caso de un vector aleatorio $\mathbf{X} = \{X^1, \ldots, X^d\}$ con función densidad de probabilidad conjunta $p(x^1, \ldots, x^d)$. Se define otro vector aleatorio $\mathbf{Y} = \{Y^1, \ldots, Y^d\}$, por medio de las transformaciones $Y^j = \Psi^j(X^1, \ldots, X^d),\ j = 1, \ldots, d$. Suponiendo que las funciones Ψ^j tienen inversa (única), se puede escribir $X^j = \Phi^j(Y^1, \ldots, Y^d)$ para cada j. La función densidad de probabilidad conjunta $q(y^1, \ldots, y^d)$ para $\mathbf{Y}$ se puede obtener a partir de la propiedad de conservación de la probabilidad

$$|q(y^1, \ldots, y^d)\ dy^1 \cdots dy^d| = |p(x^1, \ldots, x^d)\ dx^1 \cdots dx^d|.$$

Para una correspondencia biunívoca entre $\mathbf{x}$ e $\mathbf{y}$, se obtiene la pdf transformada

$$q(y^1, \ldots, y^d) = |\mathrm{J}_\Phi|\ p(x^1, \ldots, x^d)$$

donde $\mathrm{J}_\Phi = \frac{\partial(\Phi^1, \ldots, \Phi^d)}{\partial(y^1, \ldots, y^d)}$ es el Jacobiano de la transformación.

Una *variable aleatoria compleja* $Z = X + iY$ puede interpretarse en términos de las dos variables aleatorias reales X e Y. La pdf asociada $P(z) = p(x, y)$ está dada por la función densidad de probabilidad conjunta de las variables reales. La condición de normalización se escribe

$$\int P(z)\, d^2z = 1$$

donde $d^2z = dx\, dy$.

Tratar el caso de leyes condicionales

1.4 Esperanza, momentos y funciones generadoras

introducción...

1.4.1 Momentos de una distribución

Una variable aleatoria continua X tiene asociado un *promedio* o *media* (también llamado *valor esperado o de expectación*) que se obtiene pesando cada valor de x con la probabilidad asociada a ese valor, $p(x)\, dx$, e integrando sobre el rango permitido de x:

$$E[X] = \langle x \rangle = \int_\Omega x\, p(x)\, dx \equiv \mu$$

si la integral existe. La *esperanza* de la variable aleatoria X representa el valor medio que puede tomar entre todos los eventos de una prueba. Una variable aleatoria X se dice *integrable* cuando $E[|X|] < \infty$.

En general, si X es una variable aleatoria, cualquier función $f(X)$ también lo es, y su valor de expectación, si existe, está dado por

$$E[f(X)] = \langle f(x) \rangle = \int_\Omega f(x)\, p(x)\, dx$$

En particular, para el monomio $f(x) = x^r$ siendo $r \in \mathbb{N}$, se obtiene el *r-ésimo momento (ordinario)* de X:

$$\nu_r \equiv E[X^r] = \langle x^r \rangle = \int_\Omega x^r\, p(x)\, dx$$

que tiene unidades de X^r. Se puede incluir el caso $r = 0$, que corresponde a la condición de normalización: $\nu_0 = \int_\Omega p(x)\, dx = 1$. La media es el primer momento: $\nu_1 = \langle x \rangle = \mu$. Es fácil probar que $\langle x^2 \rangle \geq \langle x \rangle^2$. Típicamente, los primeros momentos son más relevantes que los de órdenes mayores, para la caracterización de una distribución.

Por ejemplo, para la distribución uniforme $p(x) = \frac{1}{b-a}$ en el intervalo $[a, b]$, resulta: $\nu_1 = \langle x \rangle = \frac{1}{2}(b + a)$, $\nu_2 = \langle x^2 \rangle = \frac{1}{3}(b^2 + ab + b^2)$, $\ldots$, $\nu_r = \frac{b^{r+1} - a[r+1}{(r+1)(b-a)}$.

Cuando una pdf $p(x)$ tiene soporte (semi)infinito, necesariamente la función p debe tender a 0 cuando $|x| \to \infty$. Si $p(x)$ 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 $p(x) = \frac{\gamma}{\pi}\frac{1}{\gamma^2+(x-x_0)^2}$ para $x \in (-\infty, \infty)$, con $\gamma > 0$ y x_0 fijos, no tiene momentos finitos de orden $r \geq 1$.

En el caso de una variable aleatoria discreta X que toma valores en $\Omega = \{x_1, \ldots, x_N\}$, la esperanza de la variable viene dada por $E[X] = \sum_{n=1}^{N} x_n\, p(x_n)$. Consideraremos que el espacio muestral es $\mathbb{N}$, luego resulta

$$E[X] = \langle n \rangle = \sum_{n \geq 1} n\, p_n,$$

que se puede obtener también como $E[X] = \sum_{j=0}^{\infty} \Pr(X > j)$. Para una función f definida sobre el conjunto $\{0, 1, 2, \ldots\}$ se tiene

$$E[f(X)] = \langle f(n) \rangle = \sum_{n \geq 0} f(n)\, p_n,$$

y se define el r-ésimo momento (ordinario) de n como

$$\nu_r \equiv E[X^r] = \langle n^r \rangle = \sum_{n=1}^{\infty} n^r\, p_n.$$

En el caso de variables discretas sobre $\mathbb{N}$, resulta útil introducir el r-ésimo *momento factorial* de n mediante

$$\langle n^{(r)} \rangle \equiv \langle n(n-1)\cdots[n-(r-1)] \rangle = \sum_{n=r}^{\infty} n(n-1)\cdots(n-r+1)\, p_n.$$

Los *momentos centrales* se definen alrededor de $x = \langle x \rangle$, como el valor de expectación de potencias de la *desviación* $\Delta x \equiv x - \langle x \rangle$:

$$\mu_r \equiv \langle (x - \langle x \rangle)^r \rangle = \int_{\Omega} (x - \langle x \rangle)^r\, p(x)\, dx.$$

Se deduce que si la densidad de probabilidad $p(x)$ es una función simétrica respecto a la media, entonces todos los momentos centrales impares son nulos. Los momentos centrales brindan medidas que caracterizan la distribución:

1. el primer momento central es idénticamente nulo para toda pdf:

$$\mu_1 = \langle x - \langle x \rangle \rangle = 0;$$

2. el segundo momento central se conoce como *varianza*, *dispersión* o también *desviación cuadrática media*:

$$\mu_2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = \mathrm{Var}(X) \equiv \sigma^2, \tag{1}$$

y es una medida del cuadrado del ancho efectivo de una pdf, es no negativo y se anula sólo cuando $p(x) = \delta(x)$, esto es, cuando no hay incerteza sobre el resultado. La varianza está bien definida si X es una variable aleatoria de cuadrado integrable, esto es, cuando $E[X^2] < \infty$. El *ancho* de una distribución está dado por la *desviación estándar* $\sigma = \sqrt{\mu_2}$, 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 $\frac{\sigma}{\langle x \rangle} = \sqrt{\frac{\langle x^2 \rangle}{\langle x \rangle^2} - 1}$ cuando $\langle x \rangle \neq 0$;

3. el tercer momento central permite definir el *coeficiente de asimetría*:

$$\alpha_3 \equiv \frac{\mu_3}{\sigma^3},$$

que resulta adimensional y puede tener signo positivo o negativo, anulándose para distribuciones que son simétricas respecto del valor medio;

4. el cuarto momento central da lugar a la *curtosis*:

$$\alpha_4 \equiv \frac{\mu_4}{\sigma^4},$$

que posibilita diferenciar entre distribuciones altas y angostas (con $\alpha_4 < 3$), de otras bajas y anchas (con $\alpha_4 > 3$)

La relación entre los momentos centrales y los momentos ordinarios se obtiene directamente de las definiciones:

$$\mu_r = \int (x - \langle x \rangle)^r \, p(x) \, dx = \sum_{s=0}^{r} \binom{r}{s} (-\langle x \rangle)^{r-s} \int x^s \, p(x) \, dx = \sum_{s=0}^{r} \binom{r}{s} \nu_s (-\nu_1)^{r-s}$$

para cualquier $r = 1, 2, \ldots$, siendo $\nu_0 = 1$. Por ejemplo, $\mu_2 = \nu_2 - \nu_1^2$ como en la Ec. (1), mientras que $\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$.

Dada una variable aleatoria X con una distribución de probabilidad $p(x)$, teniendo en cuenta que los dos primeros momentos dan las características más importantes de la pdf, puede resultar conveniente hacer una transformación de variable aleatoria a la llamada *forma estándar*: $Y \equiv \frac{X - \langle X \rangle}{\sigma}$, que entonces tiene media igual a 0 y desviación estándar igual a 1.

Mencionamos algunas propiedades de $E[X]$ y de $E[X^2]$.

Proposición: Sean X e Y dos variables aleatorias integrables, y sean $a, b \in \mathbb{R}$ arbitrarios. Entonces la variable aleatoria $Z = aX + bY$ es integrable, siendo $E[Z] = aE[X] + bE[Y]$.

Proposición: Sean X e Y dos variables aleatorias integrables. Si X e Y son independientes, entonces $E[XY] = E[X]E[Y]$.

Teorema: Sean X e Y dos variables aleatorias reales Las variables X e Y 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 en $\mathbb{R}$, continuas y acotadas.

Proposición: Sea X una variable aleatoria de cuadrado integrable, y sea $\mathrm{Var}(X) = E[(X - \langle X \rangle)^2] \equiv \sigma^2$ su varianza. Luego:

1. $\mathrm{Var}(X) = E[X^2] - (E[X])^2$
2. $\forall \, a \in \mathbb{R} : \; \mathrm{Var}(X + a) = \mathrm{Var}(X), \; \mathrm{Var}(aX) = a^2 \, \mathrm{Var}(X)$
3. Si Y es otra variable aleatoria de cuadrado integrable, e independiente de X, entonces: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Desigualdades de Chebyshev y de Bienaymé–Chebyshev

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.

Desigualdad de Chebyshev:

Sea X una variable aleatoria real con función densidad de probabilidad $p(x)$. Sea $g(x) \geq 0 \; \forall \, x \in \mathbb{R}$, con $g(x) \geq K \; \forall \, x \in D \subset \mathbb{R}$, para algún $K > 0$. Entonces por un lado

$$\Pr[g(X) \geq K] = \Pr[X \in D] = \int_D p(x)\, dx$$

y por otro lado

$$\langle g(X) \rangle = \int_{\mathbb{R}} g(x) p(x)\, dx \geq \int_D g(x) p(x)\, dx \geq K \int_D p(x)\, dx,$$

luego se tiene la desigualdad:

$$\Pr[g(X) \geq K] \leq \frac{\langle g(X) \rangle}{K}. \tag{2}$$

Desigualdad de Bienaymé–Chebyshev:

Sea X una variable aleatoria real de esperanza μ y varianza σ^2 finita. Entonces, $\forall \, \epsilon > 0$ se tiene la desigualdad:

$$\Pr[|X - \mu| > \epsilon] \leq \frac{\sigma^2}{\epsilon^2}.$$

En forma equivalente, se puede plantear la probabilidad de que X se aparte de su valor medio en más de cierto número η de desviaciones estándar: tomando $g(x) = (\Delta x)^2 = (x - \mu)^2$ en la Ec. (2), resulta la desigualdad :

$$\Pr[|\Delta X| \geq \eta \sigma] = \Pr[(\Delta X)^2 \geq \eta^2 \sigma^2] \leq \frac{1}{\eta^2} \tag{3}$$

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; por ejemplo, la desigualdad (3) indica que la probabilidad de encontrar una fluctuación superior a $\eta = 3$ desviaciones estándar alrededor de la media, está por debajo de $1/9$; el cálculo para una distribución típica como la Gaussiana ajusta dicha probabilidad por debajo de 0.003.

Momentos para varias variables aleatorias

En el caso de varias variables aleatorias $X, Y, Z, \ldots$ con pdf conjunta $p(x, y, z, \ldots)$ se define el *momento central de orden* $r, s, t, \ldots$ como (Mandel & Wolf, 1995; Cover & Thomas, 2006)

$$\mu_{r,s,t,\ldots} \equiv \langle (\Delta x)^r (\Delta y)^s (\Delta z)^t \ldots \rangle = \int (x - \langle x \rangle)^r (y - \langle y \rangle)^s (z - \langle z \rangle)^t \cdots \; p(x, y, z, \ldots)\, dx\, dy\, dz \cdots .$$

Por ejemplo, para $\begin{pmatrix} X \\ Y \end{pmatrix} \sim p(x, y)$ los momentos centrales de orden lineal resultan: $\mu_{1,0} = \mu_{0,1} = 0$, y los momentos centrales de orden cuadrático están dados por las varianzas de cada variable y por la llamada

covarianza: $\mu_{2,0} = {\sigma_X}^2$, $\mu_{0,2} = {\sigma_Y}^2$, y $\mu_{1,1} = \langle \Delta X \Delta Y \rangle$. Estos últimos se pueden acomodar en una matriz, con propiedades interesantes como veremos a continuación.

Sea $X^1, \ldots, X^d$ un conjunto de d variables aleatorias. La *covarianza* entre X^i y X^j se define como

$$\mu^{ij} \equiv \langle \Delta x^i \Delta x^j \rangle = \mu^{ji}$$

para $i, j = 1, \ldots, d$. Las $d(d+1)/2$ cantidades de este tipo se disponen en un arreglo (simétrico) de $d \times d$, la *matriz de covarianza* Σ, cuya diagonal son las varianzas $(\sigma^i)^2$. Por ejemplo, si $d = 2$ se tiene

$$\begin{pmatrix} X^1 \\ X^2 \end{pmatrix} \sim p(x^1, x^2) \ : \qquad \Sigma = \begin{pmatrix} (\sigma^1)^2 & \mu^{12} \\ \mu^{21} & (\sigma^2)^2 \end{pmatrix}.$$

Proposición:

$$|\mu^{ij}|^2 \leq \mu^{ii}\mu^{jj}$$

La demostración de esta proposición involucra la desigualadad de Cauchy–Schwarz .....

Se define el *coeficiente de correlación* que es adimensional y toma valores entre -1 (variables completamente anticorrelacionadas) y 1 (variables completamente correlacionadas) como: $\rho^{ij} = \rho^{ji} \equiv \frac{\mu^{ij}}{\sigma^i \sigma^j}$.
Como ejemplo, dadas X^1 y $X^2 = aX^1 + b$ que fluctúan en fase ($a > 0$) o al revés ($a < 0$), se tiene $\Delta x^2 = a\Delta x^1$, luego $\rho^{12} = \frac{a}{|a|} = \pm 1$.

....

1.4.2 Funciones generatrices

Se definen un conjunto de funciones que permiten hallar fácilmente los distintos momentos de una distribución de probabilidad. Se llaman *funciones generadoras* o *funciones generatrices*, y están dadas como valores de expectación de funciones de la variable aleatoria (discreta o continua), con un parámetro real o complejo.

La *función generadora de momentos* (MGF, *moment generating function*) se define como

$$M(\xi) \equiv \langle e^{\xi X} \rangle = \int e^{\xi x} p(x)\, dx, \quad \xi \in \mathbb{R}$$

en el caso de una variable aleatoria continua X con pdf $p(x)$. Se tiene $M(0) = \int p(x)\, dx = 1$ (que corresponde a la condición de normalización). Si la variable X es positiva y se toma $\xi = -s$ con $s > 0$, se interpreta en términos de la transformada de Laplace de la función p.
Si existe, la MGF posibilita obtener fácilmente los momentos (ordinarios) de X a distintos órdenes, mediante los coeficientes del desarrollo de M en serie de potencias de ξ:

$$M(\xi) = \sum_{r=0}^{\infty} \frac{\xi^r}{r!} \int x^r p(x)\, dx = 1 + \sum_{r=1}^{\infty} \frac{\nu_r}{r!} \xi^r$$

o, alternativamente, mediante las sucesivas derivadas de M respecto de ξ en 0:

$$\nu_r = \left. \frac{d^r M(\xi)}{d\xi^r} \right|_{\xi=0}, \quad r = 1, 2, \ldots; \quad \nu_0 \equiv 1.$$

En el caso de una variable aleatoria discreta, suponiendo que el espacio muestral es $\mathbb{N}$, se definen dos funciones: la *función generadora de momentos (ordinarios)* (MGF) dada por

$$M(\xi) \equiv \langle e^{\xi N} \rangle = \sum_{n \geq 0} e^{\xi n} p_n,$$

y la *función generadora de momentos factoriales* (FMGF, *factorial moment generating function*) como

$$F(\xi) \equiv \langle (1+\xi)^N \rangle = \sum_{n \geq 0} (1+\xi)^n p_n$$

para $\xi \in \mathbb{R}$ en ambos casos. Se verifica $M(0) = F(0) = \sum_{n=0}^{\infty} p_n = 1$. Se muestra simplemente que

$$M(\xi) = \sum_{r=0}^{\infty} \frac{\langle n^r \rangle}{r!} \xi^r,$$

lo que permite obtener los momentos de la distribución para cualquier orden $r \geq 1$. Por otro lado, el desarrollo de la FMGF da

$$F(\xi) = \sum_{n=0}^{\infty} \sum_{r=0}^{n} \binom{n}{r} \xi^r p_n = \sum_{r=0}^{\infty} \sum_{n=r}^{\infty} \frac{n(n-1)\cdots(n-r+1)}{r!} \xi^r p_n = \sum_{r=0}^{\infty} \frac{\langle n^{(r)} \rangle}{r!} \xi^r$$

teniendo en cuenta en las dobles sumas que $0 \leq r \leq n$, con n hasta $n_{\text{máx}}$ ó ∞. Se ve entonces que F permite obtener los momentos factoriales de orden r arbitrario.

Dada una variable aleatoria a valores naturales, la función $G(\xi) = \sum_{n=0}^{\infty} p_n \xi^n$, con $-1 \leq \xi \leq 1$, es también una función generatriz. Por ejemplo, si G admite derivadas primera y segunda en $\xi = 1$ se obtienen: $\langle N \rangle = G'(1)$, $\langle N(N-1) \rangle = G''(1)$, $\mathrm{Var}(N) = G''(1) + G'(1) - [G'(1)]^2$; además, se obtiene la ley de distribución evaluando derivadas de G en $\xi = 0$: $p_n = \frac{G^{(n)}(0)}{n!}$. (François, 2009)

La *función característica* (CF, *characteristic function*) tiene argumento complejo: (Lukacs, 1961)

$$C_X(\xi) \equiv \langle e^{i\xi X} \rangle = \int e^{i\xi x} p(x)\, dx.$$

La importancia de esta función reside en que siempre existe y está bien definida, dado que es la transformada de Fourier de una función absolutamente integrable (i.e. $\int |f(x)|\, dx < \infty$) (Golberg, 1961)

Si la pdf $p(x)$ es de cuadrado integrable, entonces

$$p(x) = \frac{1}{2pi} \int e^{-i\xi x} C_X(\xi)\, d\xi.$$

El requisito para esta importante relación es que $\int_{-\infty}^{\infty} |p(x)|^2\, dx < \infty$; sin embargo, aún es válida para distribuciones con una contribución tipo δ. Por otro lado los momentos, si existen, se obtienen derivando la función C tal como expresa la siguiente proposición:

Proposición: La variable aleatoria X admite momento de orden r si y sólo si la función característica C es r veces derivable en $\xi = 0$, siendo

$$\langle X^r \rangle = (-i)^r C_X^{(r)}(0).$$

Por ejemplo, en el caso de la distribución de Cauchy–Lorentz resulta

$$C(\xi) = \frac{\gamma}{\pi} \int_{-\infty}^{\infty} \frac{e^{i\xi x}}{\gamma^2 + (x - x_0)^2} dx = e^{-\gamma|\xi| e^{i x_0 \xi}}$$

tomando $\gamma > 0$. Esta función está definida para todo ξ, pero no es derivable en $\xi = 0$, lo que coincide con el hecho de que no están definidos los momentos para esta pdf.

Para una variable aleatoria compleja $Z = X + iY$, usando la noción de transformada de Fourier bidimensional, se define:

$$C_Z(\mu) \equiv \int e^{\mu^* z - \mu z^*} p(z)\, d^2 z.$$

Resumimos algunas propiedades importantes de la función característica:

1. $C(0) = 1$
2. $|C(\xi)| \leq C(0)$
3. $C(\xi)$ es una función continua en $\mathbb{R}$ (aún si la pdf $p(x)$ tiene discontinuidades)
4. $C(-\xi) = C(\xi)*$
5. $C(\xi)$ es definida no negativa, de tal forma que para un conjunto arbitrario de N números reales $\xi_1, \ldots, \xi_N$ y N números complejos $a_1, \ldots, a_N$, se cumple

$$\sum_{i,j=1}^{N} a_i^* a_j C(\xi_j - \xi_i) \geq 0.$$

6. $C(\xi) = M(i\xi) = F(e^{i\xi} - 1)$, si M y F existen; $F(\xi) = M(\ln(1 + \xi))$

Teorema 1-6. *(Bochner, Goldberg)....*

Proposición: Sean X e Y dos variables aleatorias reales independientes, cuyas funciones características son C_X y C_Y. Entonces $C_{X+Y} = C_X C_Y$.

Cumulant generating function ....

Extendemos la definición de función característica para un vector aleatorio. ...

....

1.5 Algunos ejemplos de distribuciones de probabilidad

introducción...

1.5.1 Distribuciones de variable discreta

Variable con certeza

...

Ley de Bernoulli

...

Ley geométrica

...

Distribuciń binomial

...

Distribución de Poisson

...

Estadística de los números de ocupación de niveles energéticos: distribuciones de Maxwell–Boltzmann, de Fermi–Dirac, y de Bose–Einstein

...

Leyes de los grandes números

1.5.2 Distribuciones de variable continua

Distribución uniforme sobre un intervalo

...

Distribución exponencial

...

Distribución normal o Gaussiana

...

Distribución Gamma

...

Teorema del límite central

...

EPÍLOLOGO

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. ...

Los autores

Referencias

Aczél, J. & Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. New-York: Academic Press.

Ali, S. M. & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society B*, *28*(1), 131–142.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, *65*(331), 1248–1255.

Arimoto, S. (1971). Information-theoretical considerations on estimation problems. *Information and control*, *19*(3), 181–194.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Arndt (2001). *Information Measures: Information and its Description in Sciences and Ingineering*. Berlin: Springer Verlag.

Athreya, K. B. & Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. New-York: Springer.

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.

Barron, A. R. (1984). Monotonic central limit theorem for densities. Technical report no. 50, Department of Statistics, Stanford University.

Barron, A. R. (1986). Entropy and the central limit theorem. *The Annals of Probability*, *14*(1), 336–342.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, *18*(4), 349–369.

Basseville, M. (2013). Divergence measures for statistical data processing – an annotated bibliography. *Signal Processing*, *93*(4), 621–633.

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.

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, *50*(4), 495–510.

Ben-Tal, A., Bornwein, J. M., & Teboulle, M. (1992). Spectral estimation via convex programming. In F. Y.

Phillips & J. J. Rousseau (Eds.), *Systems and Management Science by Extremal Methods* chapter 18, (pp. 275–290). Springer.

Ben-Tal, A., Charnes, A., & Teboulle, M. (1989). Entropic means. *Journal of Mathematical Analysis and Applications*, *139*(2), 537–551.

Bengtsson, I. & Życzkowski, K. (2006). *Geometry of Quantum States: An Introduction to Quantum Entanglement*. Cambridge: Cambridge University Press.

Bercher, J.-F. (2012). On a (β, q)-generalized Fisher information and inequalities invoving q-Gaussian distributions. *Journal of Mathematical Physics*, *53*(6), 063303.

Bercher, J.-F. (2013). On multidimensional generalized Cramér-Rao inequalities, uncertainty relations and characterizations of generalized q-Gaussian distributions. *Journal of Physics A*, *46*(9), 095303.

Berlekamp, E. R. (Ed.). (1974). *Key Papers in the Development of Coding Theory*. IEEE Press.

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. (1997). *Matrix Analysis*. New-York: Springer Verlag.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99–109.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, *7*(4), 401–406.

Blachman, N. M. (1965). The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, *11*(2), 267–271.

Boekee, D. E. & van der Lubbe, J. C. A. (1979). Some aspects of error bounds in feature selection. *Pattern Recognition*, *11*(5-6), 353–360.

Boekee, D. E. & van der Lubbe, J. C. A. (1980). The R-norm information measure. *Information and Control*, *45*(2), 136–155.

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). *Leçons 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).

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problem in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*(3), 200–217.

Brémaud, P. (1988). *An Introduction to Probabilitic Modeling*. New-York: Springer.

Burbea, J. & Rao, C. R. (1982). On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, *28*(3), 489–495.

Burg, J. P. (1967). Maximum entropy spectral analysis. In *Proceedings of the 37th Meeting of the Society of Exploration Geophysicists*, Oklahoma City, Oklahoma.

Burg, J. P. (1972). The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, *37*(2), 375–376.

Burg, J. P. (1975). *Maximum entropy spectral analysis*. PhD thesis, Department of Geophysics, Stanford University, Stanford University, Stanford, CA.

Cambini, A. & Martein, L. (2009). *Generalized Convexity and Optimization: Theory and Applications*. Heidelberg: Springer Verlag.

Cauchy, A.-L. (1821). *Cours d'analyse de l'école royale polytechnique*, volume 1: analyse algébrique. Paris: Imprimerie royale (digital version, Cambrige, 2009).

Chenciner, A. (2017). La force d'une idée simple. *Gazette de la Société de Mathématiques Française*, *152*, 16–22.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, *23*(4), 493–507.

Chong, K. M. (1974). Some extensions of a theorem of Hardy, Littlewood and Pólya and their applications. *Journal canadien de mathématiques*, *26*, 1321–1340.

Clavier, A. G. (1948). Evaluation of transmission efficiency according to Hartley's expression of information content. *Technical Journal of the International Telephone and Telegraph Corporation and Associate Companies*, *25*(4), 414–420.

Cohen, M. (1968). The Fisher information and convexity. *IEEE Transactions on Information Theory*, *14*(4), 591–592.

Cohn, D. L. (2013). *Measure Theory* (2nd ed.). New-York: Springer.

Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Cramér, H. (1946). *Mathematical Methods of Statistics*. New-York: Princeton University Press.

Cressie, N. & Pardo, L. (2000). Minimum ϕ-divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica*, *10*(3), 867–884.

Cressie, N. & Read, L. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, *46*(3), 440–464.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, *8*(1-2), 85–108.

Csiszàr, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 299–318.

Csiszár, I. (1974). Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory*, volume B, (pp. 73–86)., Prague, 18-23 august.

Csiszàr, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, *19*(4), 2031–2066.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, *68*(1-2), 161–186.

Csiszàr, I. & Matúš, F. (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythgorean identities. *Kybernetica*, *48*(4), 637–689.

Csiszár, I. & Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends^{TM} in Communications and Information Theory*, *1*(4), 417–528.

Darmois, G. (1935). Sur les lois de probabilités à estimation exhaustive. *Comptes rendus de l'Acadéie des Sciences*, *200*, 1265–1966.

Darmois, G. (1945). Sur les limites de la dispersion de certaines estimations. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *13*(1/4), 9–15.

Daróczy, Z. (1970). Generalized information functions. *Information and Control*, *16*(1), 36–51.

Daróczy, Z. & Járai, A. (1979). On the measurable solution of a functional equation arising in information theory. *Acta Mathematica Academiae Scientiarum Hungaricae*, *34*(1-2), 105–116.

D.Bellhouse (2005). Decoding cardano's liber de ludo aleae. *Historia Mathematica*, *32*(2), 180–202.

Dembo, A., Cover, T. M., & Thomas, J. A. (1991). Information theoretic inequalities. *IEEE Transactions on Information Theory*, *37*(6), 1501–1518.

DeMoivre, A. (1756). *The Doctrine of Chances : or, a method for calculating the probabilities of events in play* (3rd ed.). London: AMS Chelsea Publishing.

Doob, J. L. (1936). Statistical estimation. *Transactions of the American Mathematical Society*, *39*(3), 410–421.

(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.

Ebeling, W., Molgedey, L., Kurths, J., & Schwarz, U. (2000). Entropy, complexity, predictability and data analysis of time series and letter sequences. In *Theory of Disaster* (A. Bundle and H.-J. Schellnhuber ed.). Berlin: Springer Verlag.

Edgeworth, F. Y. (1908). On the probable errors of frequency-constants. *Journal of the Royal Statistical Society*, *71*(3, 6 & 7), 381–397, 499–512 & 499–512.

Elias, P. (1957). List decoding for noisy channels. Technical Report 335, Research Laboratory of Electronics, MIT, MIT, Cambridge, MA.

Endres, D. & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, *49*(7), 1858–1860.

Esteban, M. D. (1997). A general class of entropy statistics. *Applications of Mathematics*, *42*(3), 161–169.

Fadeev, D. K. (1956). On the concept of entropy of a finite probabilistic scheme (russian). *Uspekhi Matematicheskikh Nauk*, *11*(1(67)), 227–231.

Fadeev, D. K. (1958). *Fundations in Information Theory*, chapter On the concept of entropy of a finite probabilistic scheme (English traduction). New-York: McGraw-Hill.

Fano, R. M. (1949). The transmission of information. Technical Report 65, Research Laboratory of Electronics, MIT, MIT, Cambridge, MA.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2. New-York: John Wiley & Sons, Inc.

Ferreri, C. (1980). Hypoentropy and related heterogeneity, divergence and information measures. *Statistica*, *2*, 155–167.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, *222*(594-604), 309–368.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Flandrin, P. & Rioul, O. (2016). Laplume, sous le masque.

François, O. (2009). Notes de cours de probabilités appliquées. Ensimag.

Fréchet, M. (1943). Sur l'extension de certaines evaluations statistiques au cas de petits echantillons. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *11*(3/4), 182–205.

Frieden, B. R. (2004). *Science from Fisher Information: A Unification*. Cambridge, UK: Cambridge University Press.

Frigyik, B. A., Srivastava, S., & Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Infomation Theory*, *54*(11), 5130–5139.

Gallager, R. (1978). Variations on a theme by Huffman. *IEEE Transactions on Information Theory*, *24*(6), 668–674.

Gallager, R. (2001). Claude E. Shannon: a retrospective on his life, work, and impact. *IEEE Transactions on Information Theory*, *47*(7), 2681–2695.

Gelfand, I. M. & Fomin, S. V. (1963). *Calculus of Variations*. Englewood Cliff, NJ, USA: Prentice Hall.

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.

Guo, D., Shamai, S., & Verdú, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, *51*(4), 1261–1282.

Gupta, H. C. & Sharma, B. D. (1976). On non-additive measures of inaccuracy. *Czechoslovak Mathematical Journal*, *26*(4), 584–595.

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.

Haghighatshoar, S., Abbe, E., & Telatar, I. E. (2014). A new entropy power inequality for integer-valued random variables. *IEEE Transactions on Information Theory*, *60*(7), 3787–3796.

Hald, A. (1990). *History of Probability and Statistics and Their Applications before 1750*. John Wiley & Sons, Inc.

Halmos, P. R. (1950). *Measure Theory*. New-York: Springer.

Hardy, G., Littlewood, J. E., & Pólya, G. (1952). *Inequalities* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1929). Some simple inequalities satisfied by convex functions. *Messenger of Mathematics*, *58*, 145–152.

Harremoës, P. & Vignat, C. (2003). An entropy power inequality for the binomial family. *Journal of Inequalities in Pure and Applied Mathematics*, *4*(5), 93.

Hartley, R. V. L. (1928). Transmission of informations. *The Bell System Technical Journal*, *7*(3), 535–563.

Hausdorff, F. (1901). Beiträge zur wahrscheinlichkeitsrechnung. *Berichte über die Verhandlungen der Köeniglich Sächsische Akademie der Wissenschaften zu Leipzig*, *53*(1), 152–178.

Havrda, J. & Charvát, F. (1967). Quantification method of classification processes: Concept of structural α-entropy. *Kybernetika*, *3*(1), 30–35.

Hellinger, E. (1909). Neue bergründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die riene und angewandte Mathematik*, 210–271.

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.

Holevo, A. (2011). *Probabilistic and statistical aspects of quantum theory* (2nd ed.)., volume 1 of *Quaderni Monographs*. Pisa: Edizioni Della Normale.

Holevo, A. S. (1973). Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, *9*(3), 177–183.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101.

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. (1957a). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, *108*(2), 171–190.

Jaynes, E. T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, *33*(5), 391–398.

Jaynes, E. T. (1968). Prior probabilities. *IEEE transactions on systems science and cybernetics*, *4*(3), 227–241.

Jaynes, E. T. (1982). On the rational of maximum-entropy methods. *Proceedings of the IEEE*, *70*(9), 939–952.

Jeffrey (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal*

Society A, *186*(1007), 453–461.

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, O. (2004). *Information Theory and The Central Limit Theorem*. London: Imperial college Press.

Johnson, O. & Yu, Y. (2010). Monotonicity, thinning, and discrete versions of the entropy power inequality. *IEEE Transactions on Information Theory*, *56*(11), 5387–5395.

Kafka, P., Österreicher, F., & Vincze, I. (1991). On powers of f-divergences defining a distance. *Studia Scientiarum Mathematicarum Hungarica*, *24*(4), 415–422.

Kagan, A. (2001). A discrete version of the stam inequality and a characterization of the Poisson distributions. *Journal of Statistical Planning and Inference*, *92*(1-2), 7–12.

Kagan, A. & Smith, P. J. (1999). A stronger version of matrix convexity as applied to functions of Hermitian matrices. *Journal of Inequalities and Applications*, *3*(2), 143–152.

Kagan, A. & Yu, T. (2008). Some inequalities related to the Stam inequality. *Applications of Mathematics*, *53*(3), 195–205.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications*, *15*(1), 52–60.

Kaniadakis, G. (2001). Non-linear kinetics underlying generalized statistics. *Physica A*, *296*(3-4), 405–425.

Kapur, J. N. (1967). Generalized entropy of order α and type β. *The Mathematical Seminar*, *4*, 78–94.

Kapur, J. N. (1989). *Maximum Entropy Model in Sciences and Engineering*. New-Dehli: Wiley Eastern Limited.

Kapur, J. N. & Kesavan, H. K. (1992). *Entropy Optimization Principle with Applications*. San Diego: Academic Press.

Karamata, J. (1932). Sur une inegalité relative aux fonctions convexes. *Publications Mathématiques de l'Université de Belgrade*, *1*, 145–148.

Karush, J. (1961). A simple proof of an inequality of McMillan. *IEEE Transactions on Information Theory*, *7*(2), 118–118.

Kay, S. M. (1993). *Fundamentals for Statistical Signal Processing: Estimation Theory*. vol. 1. Upper Saddle River, NJ: Prentice Hall.

Kendall, D. G. (1964). Functional equations in information theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *2*(3), 225–229.

Khinchin, A. I. (1957). *Mathematical foundations of information theory*. New-York: Dover Publications.

Kolmogorov, A. N. (1930). Sur la notion de la moyenne. *Atti della Reale Accademia Nazionale dei Lincei*, *12*,

388–391.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability* (2nd ed.). New-York: Chelsea Publishing Company.

Kolmogorov, A. N. (1991). On the notion of mean. In V. M. Tikhomirov (Ed.), *Selected Works of A. N. Kolmogorov*, volume I: Mathematics and Mechanics (pp. 144–146). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, *39*(3), 399–399.

Kraft Jr, L. G. (1949). A device for quantizing, grouping, and coding amplitude-modulated pulses. Master's thesis, Department of Electrical Engineering, MIT, Massachusetts Institute of Technology.

Krajči, S., Liu, C.-F., Mikeš, L., & Moser, S. M. (2015). Performance analysis of Fano coding. In *2015 IEEE International Symposium on Information Theory (ISIT)*, (pp. 1746–1750)., Hong-Kong, China.

Kuczma, M. (2009). *An Introduction to the Theory of Functional Equations and Inequalities: Cauchy's Equation and Jensen's Inequality* (2nd ed.). Basel: Birkhäuser.

Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications.

Kullback, S. & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kumar, P. & Chhina, S. (2005). A symmetric information divergence measure of the Csiszár's f-divergence class and its bounds. *Computers and Mathematics with Applications*, *49*(4), 575–588.

Laplume, J. (1948). Sur le nombre de signaux discernables en présence de bruit erratique dans un système de transmission à bande passance limitéee. *Comptes Rendus de l'Academie des Sciences*, *226*, 1348–1349. Séance du 26 avril.

Lebesgue, H. (1904). *Leçons 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.

Lee, P. M. (1964). On the axioms of information theory. *The Annals of Mathematical Statistics*, *35*(1), 415–418.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New-York: Springer-Verlag.

Lieb, E. H. (1975). Some convexity and subadditvity properties of entropy. *Bulletin of the American Mathematical Society*, *81*(1), 1–13.

Lieb, E. H. (1978). Proof of an entropy conjecture of Wehrl. *Communications in Mathematical Physics*, *62*(1), 35–41.

Lieb, E. H. & Loss, M. (2001). *Analysis* (2nd ed.). Providence, Rhode Island: American Mathematical Society.

Liese, F. & Vajda, I. (2006). On divergence and informations in statistics and information theory. *IEEE Transactions on Information Theory*, *52*(10), 4394–4412.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Lindhard, J. & Nielsen, V. (1971). Studies in statistical mechanics. *Det Kongelige Danske Videnskabernes Selskab Matematisk-fysiske Meddelelser*, *38*(9), 1–42.

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.

Lundheim, L. (2002). On Shannon and "Shannon's formula". *Telektronikk*, *98*(1), 20–29.

Lutwak, E., Lv, S., Yang, D., & Zhang, G. (2012). Extension of Fisher information and Stam's inequality. *IEEE Transactions on Information Theory*, *58*(3), 1319–1327.

Lutwak, E., Yang, D., & Zhang, G. (2005). Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory*, *51*(2), 473–478.

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.

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.

McMillan, B. (1956). Two inequalities implied by unique decipherability. *IEEE Transactions on Information Theory*, *2*(4), 115–116.

Menéndez, M. L., Morales, D., Pardo, L., & Salicrú, M. (1997). (h, ϕ)-entropy differential metric. *Applications of Mathematics*, *42*(1-2), 81–98.

Menéndez, M. L., Morales, D., Pardo, L., & Vajda, I. (1977). Testing in stationary models based on divergences of observed and theoretical frequencies. *Kybernetica*, *33*(5), 465–475.

Merhav, N. (2010). Statistical physics and information theory. *Foundations and Trends® in Communications and Information Theory*, *6*(1-2), 1–212.

Merhav, N. (2018). *Statistical Physics for Electrical Engineering*. Springer.

Miller, R. E. (2000). *Optimization: Foundations and Applications*. New-York: John Wiley & Sons, inc.

Minkowski, H. (1910). *Geometrie der Zahlen*. Leipzig, Germany: Teubner.

Mittal, D. P. (1975). On additive and non-additive entropies. *Kybernetica*, *11*(4), 271–276.

Montagné, J.-C. B. (2008). *Transmissions. L'histoire des moyens de communication à distance depuis l'Antiquité jusqu'au milieu du xxe siècle*. Bagneux, JCB Montagné.

Morimoto, T. (1963). Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, *18*(3), 328–331.

Nagumo, M. (1930). Über eine klasse der mittelwerte. *Japanese journal of mathematics: transactions and abstracts*, *7*, 71–79.

Nielsen, F. & Boltz, S. (2011). The Burbea-Rao and Bhattacharyya centroids. *IEEE Transactions on Information Theory*, *57*(8), 5455–5466.

Nielsen, F. & Nock, R. (2017). Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Processing Letters*, *24*(8), 1123–1127.

Ohya, M. & Petz, D. (1993). *Quantum Entropy and Its Use*. Berlin: Springer Verlag.

Onicescu, O. (1966). Energie informationnelle. *Comptes rendus de l'académie des Sciences. série 1, mathématiques*, *263*(3), 841–842.

Orsak, G. C. & Paris, B.-P. (1995). On the relationship between measures of discrimination and the performance of suboptimal detectors. *IEEE Transactions on Information Theory*, *41*(1), 188–203.

Osán, T. M., Bussandri, D. G., & Lamberti, P. W. (2018). Monoparametric family of metrics derived from classical Jensen-Shannon divergence. *Physica A*, *495*, 336–344.

Österreicher, F. (1996). On a class of perimeter-type distances of probability distributions. *Kybernetika*, *32*(4), 389–393.

Österreicher, F. & Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, *55*(3), 639–653.

Palomar, D. P. & Verdú, S. (2006). Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, *52*(1), 141–154.

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Boca Raton, FL, USA: Chapman & Hall.

Pardo, M. C. (1999). On Burbea-Rao divergence based goodness-of-fit tests for multinomial models. *Journal of Multivariate Analysis*, *69*(1), 65–87.

Payaró, M. & Palomar, D. P. (2009). Hessian and concavity of mutual information differential entropy, and entropy power in linear vector gaussian channels. *IEEE Transactions on Information Theory*, *55*(8), 3613–3628.

Pearson, K. & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. on the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London A*, *191*, 229–311.

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.

Petz, D. (2007). Bregman divergence as relative operator entropy. *Acta Mathematica Hungarica*, *116*(1-2), 127–131.

Phillips, F. Y. & Rousseau, J. J. (Eds.). (1992). *Systems and Management Science by Extremal Methods*. Springer.

Pigeon, S. (2003). Huffman coding. In K. Sayood (Ed.), *Lossless Compression Handbook* chapter 4, (pp. 79–99). San Diego, CA: Academic Press.

Planck, M. (2015). *Eight Lectures on Theoretical Physics*. New-York: Columbia University Press.

Poor, H. V. (1988). Fine quantization in signal detection and estimation. *IEEE Transactions on Information Theory*, *34*(5), 960–972.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, *37*(3), 81–91.

Rao, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Foundations and Basic Theory*, volume I (pp. 235–247). New York: Springer.

Rao, C. R. & Wishart, J. (1947). Minimum variance and the estimation of several parameters. *Mathematical Proceedings of the Cambridge Philosophical Society*, *43*(2), 280–283.

Rathie, P. N. (1991). Unified (r, s)-entropy and its bivariate measures. *Information Sciences*, *54*(1-2), 23–39.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Contributions to the Theory of Statistics, (pp. 547–561). University of California Press, Berkeley, CA.

Rioul, O. (2007). *Théorie de l'information et du codage*. Paris: Lavoisier.

Rioul, O. (2011). Information theoretic proofs of entropy power inequalities. *IEEE Transactions on Information Theory*, *57*(1), 33–55.

Rioul, O. (2017). Yet another proof of the entropy power inequality. *IEEE Transactions on Information Theory*, *63*(6), 3595–3599.

Rioul, O. & Flandrin, P. (2017). Le dessein de laplume. In *Colloque GRETSI*, Juan-les-Pins, France.

Rioul, O. & Magossi, J. (2014). On Shannon's formula and Hartley's rule: Beyond the mathematical coincidence. *Entropy*, *16*(12), 4892–4910.

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.

Salicrú, M. (1987). Funciones de entropía asociada a medidas de Csiszár. *Qüestiió*, *11*(3), 3–12.

Salicrú, M. (1994). Measures of information associated with Csiszár's divergences. *Kybernetica*, *30*(5), 563–573.

Salicrú, M., Menéndez, M. L., Morales, D., & Pardo, L. (1993). Asymptotic distribution of (h, ϕ)-entropies. *Communications in Statistics – Theory and Methods*, *22*(7), 2015–2031.

Sayood, K. (Ed.). (2003). *Lossless Compression Handbook*. San Diego, CA: Academic Press.

Schur, I. (1923). Über eine klasse von mittelbildungen mit anwendungen auf die determinantentheorie. *Sitzungsberichte der Berliner Mathematischen Gesellschaft*, *22*, 9–20.

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.

Shafer, G. & Vovk, V. (2006). The sources of Kolmogorov's grundbegriffe. *Statistical Science*, *21*(1), 70–98.

Shamai, S. & Wyner, A. (1990). A binary analog to the entropy-power inequality. *IEEE Transactions on Information Theory*, *36*(6), 1428–1430.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(4), 623–656.

Shannon, C. E. & Weaver, W. (1964). *The Mathematical Theory of Communication*. Urbana, USA: The University of Illinois Press.

Sharma, B. D. & Mittal, D. P. (1975). New non-additive measures of entropy for discrete probability distributions. *Journal of Mathematical Sciences*, *10*, 28 –40.

Sharma, B. D. & Taneja, I. J. (1975). Entropy of type (α, β) and other generalized measures in information theory. *Metrika*, *22*(1), 205–215.

Sharma, N., Das, S., & Muthukrishnan, S. (2011). Entropy power inequality for a family of discrete random variables. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, (pp. 1945–1949)., Saint Petersburg, Russia.

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.

Spiegel, M. (1976). *Probabilidad y Estadística*. México: McGrew Hill.

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, *2*(2), 101–112.

Steele, J. M. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge: Cambridge University Press.

Stix, G. (1991). Profile: Davis a. Huffman. *Scientific American*, *265*(3), 54–58.

Teboulle, M. (1992). On Φ-divrgence and its applications. In F. Y. Phillips & J. J. Rousseau (Eds.), *Systems and Management Science by Extremal Methods* chapter 17, (pp. 255–273). Springer.

Toranzo, I. V., Zozor, S., & Brossier, J.-M. (2018). Generalization of the de Bruijn identity to general ϕ-entropies and ϕ-fisher informations. *IEEE Transactions on Information Theory*, on press.

Tribus, M. & McIrvine, E. C. (1971). Energy and information. *Scientific American*, *225*(3), 179–188.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, *52*(1-2), 479–487.

Tverberg, H. (1958). A new derivation of the information function. *Mathematica Scandinavica*, *6*, 297–298.

Vajda, I. (1968). Axioms for a-entropy of a generalized probability scheme. *Kybernetika*, *4*(2), 105–112.

Vajda, I. (1972). On the f-divergence and singularity of probability measures. *Periodica Mathematica Hungarica*, *2*(1-4), 223–234.

Vajda, I. (2009). On metric divergences of probability measures. *Kybernetika*, *45*(6), 885–900.

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 Brunt, B. (2004). *The Calculus of Variations*. New-York: Springer Verlag.

van den Bos, A. (2007). *Parameter Estimation for Scientists and Engineers*. Hoboken, New Jersey: John Wiley & Sons.

Varma, R. S. (1966). Generalization of Rényi's entropy of order α. *Journal of Mathematical Sciences*, *1*, 34–48.

Verdu, S. (1998). Fifty years of Shannon theory. *IEEE Transactions on Information Theory*, *44*(6), 2057–2078.

Verdú, S. & Guo, D. (2006). A simple proof of the entropy-power inequality. *IEEE Transactions on Information Theory*, *52*(5), 2165–2166.

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.

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal end the Machine* (2nd ed.). Cambridge, MA: MIT Press.

Wong, A. K. C. & You, M. (1985). Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *7*(5), 599–609.

Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, *44*(3), 1246–1250.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, *16*(1), 159–195.

Zozor, S., Puertas-Centeno, D., & Dehesa, J. S. (2017). On generalized Stam inequalities and Fisher–Rényi complexity measures. *Entropy*, *19*(9), 493.

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.