Introducción a la Probabilidad

Francisco Montes Suay

Departament d'Estadística i Investigació Operativa
Universitat de València

Departament d'Estadística i Investigació Operativa
Universitat de València
46100-Burjassot
Spain

Índice general

Capítulo 1

Espacio de probabilidad

1.1. Causalidad y aleatoriedad

A cualquiera que preguntemos cuanto tiempo tardaríamos en recorrer los 350 kilómetros que separan Valencia de Barcelona, si nos desplazamos con velocidad constante de 100 kms/hora, nos contestará sin dudar que 3 horas y media. Su actitud será muy distinta si, previamente a su lanzamiento, le preguntamos por la cara que nos mostrará un dado. Se trata de dos fenómenos de naturaleza bien distinta,

- *el primero* pertenece a los que podemos denominar **deterministas**, aquellos en los que la relación causa-efecto aparece perfectamente determinada. En nuestro caso concreto, la conocida ecuación $e = v \cdot t$, describe dicha relación,
- *el segundo* pertenece a la categoría de los que denominamos **aleatorios**, que se caracterizan porque aun repitiendo en las mismas condiciones el experimento que lo produce, el resultado variará de una repetición a otra dentro de un conjunto de posibles resultados.

La Teoría de la Probabilidad pretende emular el trabajo que los físicos y, en general, los científicos experimentales han llevado a cabo. Para entender esta afirmación observemos que la ecuación anterior, $e = v \cdot t$, es un resultado experimental que debemos ver como *un modelo matemático* que, haciendo abstracción del móvil concreto y del medio en el que se desplaza, describe la relación existente entre el espacio, el tiempo y la velocidad. La Teoría de la Probabilidad nos permitirá la obtención de *modelos aleatorios o estocásticos* mediante los cuales podremos conocer, en términos de probabilidad, el comportamiento de los fenómenos aleatorios.

1.2. Experimento, resultado, espacio muestral y suceso

Nuestro interlocutor sí que será capaz de responder que el dado mostrará una de sus caras. Al igual que sabemos que la extracción al azar de una carta de una baraja española pertenecerá a uno de los cuatro palos: oros, copas, espadas o bastos. Es decir, el *experimento* asociado a nuestro fenómeno aleatorio[1] da lugar a un *resultado*, ω, de entre un conjunto de posibles resultados. Este conjunto de posibles resultados recibe el nombre de *espacio muestral*, Ω. Subconjuntos de resultados con una característica común reciben el nombre de *sucesos aleatorios* o, simplemente,

[1] Una pequeña disquisición surge en este punto. La aleatoriedad puede ser inherente al fenómeno, lanzar un dado, o venir inducida por el experimento, extracción al azar de una carta. Aunque conviene señalarlo, no es este lugar para profundizar en la cuestión

sucesos. Cuando el resultado del experimento pertenece al suceso A, decimos que **ha ocurrido** o **se ha realizado** A.

A continuación mostramos ejemplos de experimentos aleatorios, los espacios muestrales asociados y algunos sucesos relacionados.

Lanzamiento de dos monedas.- Al lanzar dos monedas el espacio muestral viene definido por Ω ={CC,C+,+C,++}. Dos ejemplos de sucesos en este espacio pueden ser:

A ={Ha salido una cara}={C+,+C},

B ={Ha salido más de una cruz}={++}.

Elegir un punto al azar en el círculo unidad.- Su espacio muestral es Ω ={Los puntos del círculo}. Ejemplos de sucesos:

$A = \{\omega;\ d(\omega, centro) < 0{,}5\}$,

$B = \{\omega;\ 0,3 < d(\omega, centro) < 0{,}75\}$.

1.2.1. Sucesos, conjuntos y σ-álgebra de sucesos

Puesto que los sucesos no son más que subconjuntos de Ω, podemos operar con ellos de acuerdo con las reglas de la teoría de conjuntos. Todas las operaciones entre conjuntos serán aplicables a los sucesos y el resultado de las mismas dará lugar a nuevos sucesos cuyo significado debemos conocer. Existen, por otra parte, sucesos cuya peculiaridad e importancia nos lleva a asignarles nombre propio. De estos y de aquellas nos ocupamos a continuación:

Suceso cierto o seguro: cuando llevamos a cabo cualquier experimento aletorio es seguro que el resultado pertenecerá al espacio muestral, por lo que Ω, en tanto que suceso, ocurre siempre y recibe el nombre de *suceso cierto* o *seguro*.

Suceso imposible: en el extremo opuesto aparece aquel suceso que no contiene ningún resultado que designamos mediante $\emptyset$ y que, lógicamente, no ocurre nunca, razón por la cual se le denomina *suceso imposible*.

Sucesos complementarios: la ocurrencia de un suceso, A, supone la no ocurrencia del suceso que contiene a los resultados que no están en A, es decir, A^c. Ambos sucesos reciben el nombre de *complementarios*.

Unión de sucesos: la unión de dos sucesos, $A \cup B$, da lugar a un nuevo suceso que no es más que el conjunto resultante de dicha unión. En consecuencia, $A \cup B$ *ocurre* cuando el resultado del experimento pertenece a A, a B o ambos a la vez.

Intersección de sucesos: la intersección de dos sucesos, $A \cap B$, es un nuevo suceso cuya *realización* tiene lugar si el resultado pertenece a ambos a la vez, lo que supone que ambos ocurren simultáneamente.

Sucesos incompatibles: Existen sucesos que al no compartir ningún resultado su intersección es el suceso imposible, $A \cap B = \emptyset$. Se les denomina, por ello, *sucesos incompatibles*. Un suceso A y su complementario A^c, son un buen ejemplo de sucesos incompatibles.

Siguiendo con el desarrollo emprendido parece lógico concluir que todo subconjunto de Ω será un suceso. Antes de admitir esta conclusión conviene una pequeña reflexión: la noción de suceso es un concepto que surge con naturalidad en el contexto de la experimentación aleatoria pero, aunque no hubiera sido así, la necesidad del concepto nos hubiera obligado a *inventarlo*. De

la misma forma, es necesario que los sucesos poseean una mínima *estructura* que garantice la estabilidad de las operaciones *naturales* que con ellos realicemos, entendiendo por naturales la *complementación*, la *unión* y la *intersección*. Esta dos últimas merecen comentario aparte para precisar que no se trata de uniones e intersecciones en número cualquiera, puesto que más allá de la numerabilidad nos movemos con dificultad. Bastará pues que se nos garantice que uniones e intersecciones numerables de sucesos son estables y dan lugar a otro suceso. Existe una estructura algebraica que verifica las condiciones de estabilidad que acabamos de enumerar.

Definición 1.1 (σ-álgebra de conjuntos) *Una familia de conjuntos $\mathcal{A}$ definida sobre Ω decimos que es una σ-álgebra si:*

1. $\Omega \in \mathcal{A}$.
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$.
3. $\{A_n\}_{n\geq 1} \subset \mathcal{A} \Rightarrow \bigcup_{n\geq 1} A_n \in \mathcal{A}$.

La familia de las partes de Ω, $\mathcal{P}(\Omega)$, cumple con la definición y es por tanto una σ-álgebra de sucesos, de hecho la más grande de las existentes. En muchas ocasiones excesivamente grande para nuestras necesidades, que vienen determinadas por el núcleo inicial de sucesos objeto de interés. El siguiente ejemplo permite comprender mejor este último comentario.

Ejemplo 1.1 *Si suponemos que nuestro experimento consiste en elegir al azar un número en el intervalo [0,1], nuestro interés se centrará en conocer si la elección pertenece a cualquiera de los posibles subintervalos de [0,1]. La σ-álgebra de sucesos generada a partir de ellos, que es la menor que los contiene, se la conoce con el nombre de* σ-álgebra de Borel en [0,1]*, $\beta_{[0,1]}$, y es estrictamente menor que $\mathcal{P}([0,1])$.*

En resumen, el espacio muestral vendrá acompañado de la correspondiente σ-álgebra de sucesos, la más conveniente al experimento. La pareja que ambos constituyen, $(\Omega, \mathcal{A})$, recibe el nombre de *espacio probabilizable*.

Señalemos por último que en ocasiones no es posible economizar esfuerzos y $\mathcal{A}$ coincide con $\mathcal{P}(\Omega)$. Por ejemplo cuando el espacio muestral es numerable.

1.3. Probabilidad

Ya sabemos que la naturaleza aleatoria del experimento impide *predecir* de antemano el resultado que obtendremos al llevarlo a cabo. Queremos conocer si cada suceso de la σ-álgebra se realiza o no. Responder de una forma categórica a nuestro deseo es demasiado ambicioso. Es imposible predecir en cada realización del experimento si el resultado va a estar o no en cada suceso. En Probabilidad la pregunta se formula del siguiente modo: ¿qué posibilidad hay de que tenga lugar cada uno de los sucesos? La respuesta exige un tercer elemento que nos proporcione esa información: Una función de conjunto P, es decir, una función definida sobre la σ-álgebra de sucesos, que a cada uno de ellos le asocie un valor numérico que exprese la mayor o menor probabilidad o posibilidad de producirse cuando se realiza el experimento. Esta función de conjunto se conoce como *medida de probabilidad* o simplemente *probabilidad*. Hagamos un breve incursión histórica antes de definirla formalmente.

El concepto de probabilidad aparece ligado en sus orígenes a los juegos de azar, razón por la cual se tiene constancia del mismo desde tiempos remotos. A lo largo de la historia se han hecho muchos y muy diversos intentos para formalizarlo, dando lugar a otras tantas *definiciones* de probabilidad que adolecían todas ellas de haber sido confeccionadas *ad hoc*, careciendo por tanto de la generalidad suficiente que permitiera utilizarlas en cualquier contexto. No por ello el

interés de estas definiciones es menor, puesto que supusieron sucesivos avances que permitieron a Kolmogorov enunciar su conocida y definitiva axiomática en 1933. De entre las distintas aproximaciones, dos son las más relevantes:

Método frecuencialista.- Cuando el experimento es susceptible de ser repetido en las mismas condiciones una infinidad de veces, la probabilidad de un suceso A, $P(A)$, se define como el límite[2] al que tiende la *frecuencia relativa de ocurrencias* del suceso A.

Método clásico (Fórmula de Laplace).- Si el experimento conduce a un espacio muestral finito con n resultados posibles, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, todos ellos igualmente probables, la probabilidad de un suceso A que contiene m de estos resultados se obtiene mediante la fórmula

$$P(A) = \frac{m}{n},$$

conocida como *fórmula de Laplace*, que la propuso a finales del siglo XVIII. La fórmula se enuncia como *el cociente entre el número de casos favorables y el número de casos posibles*. Obsérvese la incorreción formal de esta aproximación en la medida que exige equiprobabilidad en los resultados para poder definir, precisamente, la probabilidad, lo cual implica un conocimiento previo de aquello que se quiere definir.

Las anteriores definiciones son aplicables cuando las condiciones exigidas al experimento son satisfechas y dejan un gran número de fenómenos aleatorios fuera de su alcance. Estos problemas se soslayan con la definición axiomática propuesta por A.N.Kolmogorov en 1933:

Definición 1.2 (Probabilidad) *Una función de conjunto, P, definida sobre la σ-álgebra $\mathcal{A}$ es una probabilidad si:*

1. $P(A) \geq 0$ *para todo* $A \in \mathcal{A}$.
2. $P(\Omega) = 1$.
3. P *es numerablemente aditiva, es decir, si* $\{A_n\}_{n\geq 1}$ *es una sucesión de sucesos disjuntos de* $\mathcal{A}$*, entonces*

$$P(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} P(A_n).$$

A la terna $(\Omega, \mathcal{A}, P)$ la denominaremos *espacio de probabilidad*.

Señalemos, antes de continuar con algunos ejemplos y con las propiedades que se derivan de esta definición, que los axiomas propuestos por Kolmogorov no son más que la generalización de las propiedades que posee la frecuencia relativa. La definición se apoya en las aproximaciones previas existentes y al mismo tiempo las incluye como situaciones particulares que son.

Ejemplo 1.2 (Espacio de probabilidad discreto) *Supongamos un espacio muestral, Ω, numerable y como σ-álgebra la familia formada por todos los posibles subconjuntos de Ω. Sea p una función no negativa definida sobre Ω verificando: $\sum_{\omega\in\Omega} p(\omega) = 1$. Si definimos $P(A) = \sum_{\omega\in A} p(\omega)$, podemos comprobar con facilidad que P es una probabilidad. Hay un caso particular de especial interés, el llamado* espacio de probabilidad discreto uniforme*, en el que Ω es finito, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, y $p(\omega_i) = \frac{1}{n}$, $\forall \omega_i \in \Omega$. Entonces, para $A = \{\omega_{i_1}, \omega_{i_2}, \ldots, \omega i_m\}$ se tiene*

$$P(A) = \frac{m}{n},$$

[2]Debemos advertir que no se trata aquí de un límite puntual en el sentido habitual del Análisis. Más adelante se introducirá el tipo de convergencia al que nos estamos refiriendo

que es la fórmula de Laplace, obtenida ahora con rigor. El nombre de uniforme *se justifica porque la masa de probabilidad está uniformemente repartida al ser constante en cada punto.*

Un ejemplo de espacio de probabilidad discreto uniforme es el que resulta de lanzar dos dados. El espacio muestral, $\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$*, está formado por las* 6×6 *posibles parejas de caras. Si los dados son correctos, cualquiera de estos resultados tiene la misma probabilidad, 1/36. Sea ahora* $A = \{$*ambas caras son pares*$\}$*, el número de puntos que contiene A son 9, por lo que aplicando la fórmula de Laplace,* $P(A) = 9/36 = 1/4$*.*

Ejemplo 1.3 (Probabilidad discreta) *Si el espacio probabilizable,* $(\Omega, \mathcal{A})$*, es arbitrario pero la probabilidad P verifica que existe un subconjunto numerable de* Ω*,* $D = \{\omega_k,\ k \geq 1\}$*, y una sucesión de valores no negativos,* $\{p_k\}_{k\geq 1}$*, tal que* $P(A) = \sum_{\omega_k \in A\cap D} p_k$*, se dice que la* probabilidad es discreta*. Observemos que debe verificarse* $P(D) = \sum_{\omega_k \in D} p_k = 1$*.*

Supongamos, por ejemplo, que nuestro espacio probabilizable es $(\mathcal{R}, \beta)$*, y consideremos* $\omega_k = k$ *para* $k = 0,\ldots,n$ *y* $p_k = \binom{n}{k} p^k (1-p)^{n-k}$*. Tenemos una probabilidad discreta que recibe el nombre de* binomial*.*

Ejemplo 1.4 (Espacio de probabilidad uniforme continuo) *Al elegir un punto al azar en un círculo de radio 1, el espacio muestral resultante estará formado por todos los puntos del círculo* $\Omega = \{(\omega_1, \omega_2);\ \omega_1^2 + \omega_2^2 \leq 1\}$*. La elección al azar implica que la masa de probabilidad se distribuye uniformemente en todo el círculo, lo que significa que cualquiera de sus puntos puede ser igualmente elegido. Las características del espacio no permiten afirmar, como hacíamos en el caso finito, que cada punto tiene la misma probabilidad. Ahora la uniformidad se describe afirmando que la probabilidad de cualquier suceso A es directamente proporcional a su área,* $P(A) = k|A|$*. Pero* $P(\Omega) = 1 = k|\Omega| = k\pi$*, de donde* $k = 1/\pi$*, y de aquí,*

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{\pi},\ \forall A \subset \Omega.$$

Por ejemplo, si $A = \{$*puntos que distan del centro menos de 1/2*$\}$*, A será el círculo de radio 1/2 y*

$$P(A) = \frac{\pi/4}{\pi} = \frac{1}{4}.$$

El concepto de espacio de probabilidad uniforme continuo se puede generalizar fácilmente a cualquier subconjunto de Borel acotado en R^k*, sustituyendo el área por la correspondiente medida de Lebesgue.*

1.3.1. Propiedades de la probabilidad

De la definición de probabilidad se deducen algunas propiedades muy útiles.

La probabilidad del vacío es cero.- $\Omega = \Omega \cup \emptyset \cup \emptyset \cup \ldots$ y por la aditividad numerable, $P(\Omega) = P(\Omega) + \sum_{k\geq 1} P(\emptyset)$, de modo que $P(\emptyset) = 0$.

Aditividad finita.- Si $A_1,\ldots,A_n$ son elementos disjuntos de $\mathcal{A}$, aplicando la σ-aditividad, la propiedad anterior y haciendo $A_i = \emptyset,\ i > n$ tendremos

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

Se deduce de aquí fácilmente que $\forall A \in \mathcal{A}$, $P(A^c) = 1 - P(A)$.

Monotonía.- Si $A, B \in \mathcal{A},\ A \subset B$, entonces de $P(B) = P(A) + P(B - A)$ se deduce que $P(A) \leq P(B)$.

Probabilidad de una unión cualquiera de sucesos (fórmula de inclusión-exclusión).- Si $A_1, \ldots, A_n \in \mathcal{A}$, entonces

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n). \tag{1.1}$$

Para su obtención observemos que si $n = 2$ es cierta, pues

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2 - A_1) = P(A_1) + P(A_2 - A_1) + P(A_1 \cap A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2). \end{aligned}$$

El resto se sigue por inducción.

Subaditividad.- Dados los sucesos $A_1, \ldots, A_n$, la relación existente entre la probabilidad de la unión de los A_i y la probabilidad de cada uno de ellos es la siguiente:

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

En efecto, sean $B_1 = A_1$ y $B_i = A_i - \bigcup_{j=1}^{i-1} A_j$ para $i = 2, \ldots, n$. Los B_i son disjuntos y $\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i$. Por la aditividad finita y la monotonía de $\mathcal{P}$ se tiene

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

Si se trata de una sucesión de sucesos, $\{A_n\}_{n\geq 1}$, se comprueba, análogamente, que

$$P\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} P(A_n).$$

Continuidad de la probabilidad.- Sea $\{A_n\}_{n\geq 1}$ una sucesión monótona creciente de sucesos y sea A su límite. Es decir, $A_n \subset A_{n+1},\ \forall n$ y $\bigcup_{n\geq 1} A_n = A$ (que en lo que sigue denotaremos mediante $A_n \uparrow A$). Si a partir de la sucesión inicial definimos $B_n = A_n - \bigcup_{i=1}^{n-1} A_i = A_n - A_{n-1}$, para $n > 1$, y $B_1 = A_1$, se tiene

$$P(A) = P\left(\bigcup_{n\geq 1} A_n\right) = P\left(\bigcup_{n\geq 1} B_n\right) = \sum_{n\geq 1} P(B_n) =$$

$$\lim_{n\to+\infty} \sum_{j=1}^{n} P(B_j) = \lim_{n\to+\infty} P\left(\bigcup_{j=1}^{n} B_j\right) = \lim_{n\to+\infty} P(A_n),$$

propiedad que se conoce como *continuidad desde abajo.*

Si la sucesión es decreciente y $A_n \downarrow A$, la sucesión de complementarios será creciente y aplicando la continuidad desde abajo y que $P(B^c) = 1 - P(B),\ \forall B \in \mathcal{A}$, tendremos que $P(A_n) \downarrow P(A)$, que se conoce como *continuidad desde arriba.*

1.4. Probabilidad condicionada. Teorema de Bayes

Si compramos un número para una rifa que se celebra anualmente durante las fiestas de verano en nuestro pueblo y que está compuesta por 100 boletos numerados del 1 al 100, sabemos que nuestra probabilidad ganar el premio, suceso que designaremos por A, vale

$$P(A) = \frac{1}{100}$$

Supongamos que a la mañana siguiente de celebrase el sorteo alguien nos informa que el boleto premiado termina en 5. Con esta información, ¿continuaremos pensando que nuestra probabilidad de ganar vale 10^{-2}? Desde luego sería absurdo continuar pensándolo si nuestro número termina en 7, porque evidentemente la *nueva* probabilidad valdría $P'(A) = 0$, pero aunque terminara en 5 también nuestra probabilidad de ganar habría cambiado, porque los números que terminan en 5 entre los 100 son 10 y entonces

$$P'(A) = \frac{1}{10},$$

10 veces mayor que la inicial.

Supongamos que nuestro número es el 35 y repasemos los elementos que han intervenido en la nueva situación. De una parte, un suceso original $A = \{$*ganar el premio con el número 35*$\}$, de otra, un suceso $B = \{$*el boleto premiado termina en 5*$\}$ de cuya ocurrencia se nos informa *a priori*. Observemos que $A \cap B = \{$*el número 35*$\}$ y que la nueva probabilidad encontrada verifica,

$$P'(A) = \frac{1}{10} = \frac{1/100}{10/100} = \frac{P(A \cap B)}{P(B)},$$

poniendo en evidencia algo que cabía esperar, que la *nueva* probabilidad a depende de $P(B)$. Estas propiedades observadas justifican la definición que damos a continuación.

Definición 1.3 (Probabilidad condicionada) *Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y sean A y B dos sucesos, con $P(B) > 0$, se define la* probabilidad de A condicionada a B *mediante la expresión,*

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

A la anterior expresión se la denomina *probabilidad* con toda justicia, porque verifica los tres axiomas que definen el concepto de probabilidad, como fácilmente puede comprobarse. De entre los resultados y propiedades que se derivan de este nuevo concepto, tres son especialmente relevantes: el teorema de factorización, el teorema de la probabilidad total y el teorema de Bayes.

1.4.1. Teorema de factorización

A partir de la definición de probabilidad condicionada, la probabilidad de la intersección de dos sucesos puede expresarse de la forma $P(A \cap B) = P(A|B)P(B)$. El teorema de factorización extiende este resultado para cualquier intersección finita de sucesos.

Consideremos los sucesos $A_1, A_2, \ldots, A_n$, tales que $P(\cap_{i=1}^{n} A_i) > 0$, por inducción se comprueba fácilmente que

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P\left(A_n | \cap_{i=1}^{n-1} A_i\right) P\left(A_{n-1} | \cap_{i=1}^{n-2} A_i\right) \ldots P(A_2|A_1)P(A_1).$$

Ejemplo 1.5 *En una urna que contiene 5 bolas blancas y 4 negras, llevamos a cabo 3 extracciones consecutivas sin reemplazamiento. ¿Cuál es la probabilidad de que las dos primeras sean blancas y la tercera negra?*

Cada extracción altera la composición de la urna y el total de bolas que contiene. De acuerdo con ello tendremos (la notación es obvia)

$$\begin{aligned} &P(B_1 \cap B_2 \cap N_3) = \\ &\quad = P(N_3|B_1 \cap B_2)P(B_2|B_1)P(B_1) = \tfrac{4}{7} \cdot \tfrac{4}{8} \cdot \tfrac{5}{9} \end{aligned}$$

1.4.2. Teorema de la probabilidad total

Si los sucesos $A_1, A_2, \ldots, A_n$ constituyen un partición del Ω, tal que $P(A_i) > 0$, $\forall i$, tendremos que cualquier suceso B podrá particionarse de la forma, $B = \cup_{i=1}^{n} B \cap A_i$ y tratándose de una unión disjunta podremos escribir

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i). \tag{1.2}$$

Este resultado se conoce con el nombre de *teorema de la probabilidad total*.

Encuesta sobre cuestiones *delicadas*: una aplicación del teorema de la probabilidad total

Es bien conocida la reticencia de la gente a contestar cualquier encuesta, reticencia que se convierte en clara desconfianza y rechazo si el cuestionario aborda lo que podríamos denominar *temas delicados*: situación económica, creencias religiosas, afinidades políticas, costumbres sexuales, consumo de estupefacientes, ... El rechazo y la desconfianza están casi siempre basados en la creencia de una insuficiente garantía de anonimato. Es comprensible, por tanto, el afán de los especialistas en convencer a los encuestados de que el anonimato es absoluto. El teorema de la probabilidad total puede ayudar a ello.

Supongamos que un sociólogo está interesado en conocer el consumo de drogas entre los estudiantes de un Instituto de Bachillerato. Elige 100 estudiantes al azar y para garantizar la confidencialidad de las respuestas, que sin duda redundará en un resultado más fiable, diseña una estrategia consistente en que cada estudiante extrae al azar un bola de un saco o urna que contiene 100 bolas numeradas del 1 al 100, conservándola sin que nadie la vea,

- si el número de la bola elegida está entre el *1* y el *70*, contesta a la pregunta *¿has consumido drogas alguna vez?*,
- si el número de la bola elegida está entre el *71* y el *100*, contesta a la pregunta *¿es par la última cifra de tu DNI?*.

En ambos casos la respuesta se escribe sobre un trozo de papel sin indicar, lógicamente, a cuál de las dos preguntas se está contestando.

Realizado el proceso, las respuestas afirmativas han sido 25 y para estimar la proporción de los que alguna vez han consumido droga aplicamos (1.2),

$$\begin{aligned} P(sí) \;=\; & P(sí|pregunta\ delicada)P(pregunta\ delicada) + \\ & P(sí|pregunta\ intrascendente)P(pregunta\ intrascendente) \end{aligned}$$

Sustituyendo,

$$0{,}25 = P(sí|pregunta\ delicada) \times 0{,}7 + 0{,}5 \times 0{,}3,$$

y despejando,

$$P(sí|pregunta\ delicada) = \frac{0{,}25 - 0{,}15}{0{,}7} \approx 0{,}14$$

Es obvio que $P(pregunta\ intrascendente)$ ha de ser conocida muy aproximadamente, como en el caso de la terminaciones del DNI, que por mitades deben de ser pares o impares.

1.4.3. Teorema de Bayes

Puede tener interés, y de hecho así ocurre en muchas ocasiones, conocer la probabilidad asociada a cada elemento de la partición dado que ha ocurrido B, es decir, $P(A_i|B)$. Para ello, recordemos la definición de probabilidad condicionada y apliquemos el resultado anterior.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

Este resultado, conocido como el *teorema de Bayes*, permite conocer el cambio que experimenta la probabilidad de A_i como consecuencia de haber ocurrido B. En el lenguaje habitual del Cálculo de Probabilidades a $P(A_i)$ se la denomina probabilidad *a priori* y a $P(A_i|B)$ probabilidad *a posteriori*, siendo la ocurrencia de B la que establece la frontera entre el antes y el después. ¿Cuál es, a efectos prácticos, el interés de este resultado? Veámoslo con un ejemplo.

Ejemplo 1.6 *Tres urnas contienen bolas blancas y negras. La composición de cada una de ellas es la siguiente:* $U_1 = \{3B, 1N\}$, $U_2 = \{2B, 2N\}$, $U_3 = \{1B, 3N\}$. *Se elige al azar una de las urnas, se extrae de ella una bola al azar y resulta ser blanca. ¿Cuál es la urna con mayor probabilidad de haber sido elegida?*

Mediante U_1, U_2 *y* U_3, *representaremos también la urna elegida. Estos sucesos constituyen una partición de* Ω *y se verifica, puesto que la elección de la urna es al azar,*

$$P(U_1) = P(U_2) = P(U_3) = \frac{1}{3}.$$

Si B=$\{$la bola extraída es blanca$\}$, *tendremos*

$$P(B|U_1) = \frac{3}{4}, \quad P(B|U_2) = \frac{2}{4}, \quad P(B|U_3) = \frac{1}{4}.$$

Lo que nos piden es obtener $P(U_i|B)$ *para conocer cuál de las urnas ha originado, más probablemente, la extracción de la bola blanca. Aplicando el teorema de Bayes a la primera de las urnas,*

$$P(U_1|B) = \frac{\frac{1}{3} \cdot \frac{3}{4}}{\frac{1}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot \frac{2}{4} + \frac{1}{3} \cdot \frac{1}{4}} = \frac{3}{6},$$

y para las otras dos, $P(U_2|B) = 2/6$ *y* $P(U_3|B) = 1/6$. *Luego la primera de las urnas es la que con mayor probabilidad dió lugar a una extracción de bola blanca.*

El teorema de Bayes es uno de aquellos resultados que inducen a pensar que *la cosa no era para tanto*. Se tiene ante él la sensación que produce lo trivial, hasta el punto de atrevernos a pensar que lo hubiéramos podido deducir nosotros mismos de haberlo necesitado, aunque afortunadamente el Reverendo Thomas Bayes se ocupó de ello en un trabajo titulado *An Essay towards solving a Problem in the Doctrine of Chances*, publicado en 1763. Conviene precisar que Bayes no planteó el teorema en su forma actual, que es debida a Laplace.

1.4.4. El teorema de Bayes en términos de apuestas (*odds*): el valor de la evidencia

Cuando apostamos en una carrera de caballos es lógico que lo hagamos a aquel caballo que creemos ganador, es decir, aquél que tiene mayor probabilidad de ganar. Pero el mundo de las apuestas tiene un lenguaje propio y no se habla en él de probabilidad de ganar, utilizando en su lugar expresiones del tipo: las apuestas están "5 a 2 a favor" de un determinado caballo o "6 a 1 en contra" de que el Valencia gane la Liga.

¿Qué significa que las apuestas están "3 a 2 en contra" de que *Lucero del Alba* gane el Grand National? La expresión resume el hecho de que 2 de cada 5 apostantes lo hacen por dicho caballo como vencedor. Si habláramos de "5 a 2 a favor" estaríamos afirmando que 5 de cada 7 apostantes lo consideran ganador. Si queremos expresar estas afirmaciones en términos de probabilidad y denotamos por G el suceso *Lucero del Alba gana*, $P(G)$ no es más que la proporción de apostantes que piensan que ganará, es decir, $P(G) = 2/5$ o $P(G^c) = 3/5$ en el primer caso y $P(G) = 5/7$ o $P(G^c) = 2/7$ en el segundo.

Podemos establecer una sencilla relación entre ambas formas de expresar la misma idea. Si por O (del inglés *odds*) denotamos las apuestas en contra expresadas en forma de fracción, podemos escribir

$$O = \frac{P(G^c)}{P(G)},$$

que no es más que el cociente entre la probabilidad de no ganar y la de hacerlo. A su vez, como $P(G^c) = 1 - P(G)$, fácilmente se obtiene la expresión de la probabilidad de ganar en términos de las apuestas

$$P(G) = \frac{1}{O+1}. \tag{1.3}$$

Volvamos de nuevo al teorema de Bayes. Dados dos sucesos A y B escribíamos

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

Si reemplazamos A por su complementario, A^c, tenemos

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

Al dividir ambas expresiones obtenemos

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \times \frac{P(A)}{P(A^c)}, \tag{1.4}$$

expresión que se conoce como el *teorema de Bayes en forma de apuestas (odds)*. Comentemos el significado de los tres cocientes que aparecen en (1.4).

La izquierda de la igualdad representa las apuestas a favor de A, dado el suceso B. El segundo factor de la derecha son esas mismas apuestas obtenidas sin la información que supone conocer que ha ocurrido B. Por último, el primer factor de la parte derecha de la igualdad es el cociente entre las probabilidades de un mismo suceso, B, según que A haya ocurrido o no. Es lo que se denomina *razón de verosimilitud*.

Para ver el interés que (1.4) tiene, vamos a utilizarla en un contexto forense. Se trata de obtener el valor de una evidencia (Ev) en la discusión sobre la culpabilidad (C) o inocencia (C^c) de un sospechoso. La expresión (1.4) nos permite adaptar las apuestas *a priori* (antes

de la presentación de la evidencia Ev) a favor de su culpabilidad y convertirlas en apuestas *a posteriori*, llevando a cabo dicha conversión mediante el factor

$$V = \frac{P(Ev|C)}{P(Ev|C^c)}, \tag{1.5}$$

al que se conoce como *valor de la evidencia*. Es importante destacar el hecho de que para su cálculo necesitamos dos probabilidades: las de Ev tanto si el sospechoso es culpable[3] como si es inocente.

El ejemplo que sigue ilustra el papel que este concepto puede jugar durante un juicio en la valoración de las pruebas y la consecuente ayuda que para juez o jurado supone.

Harvey contra el Estado (Alaska, 1999)

En 1993 Kimberly Esquivel, una adolescente de 14 años que vivía con su madre y su padrastro, quedó embarazada y se sometió a una operación de aborto. Poco después del aborto acusó a su padrastro, Patrick Harvey, de ser el padre[4]. Se llevó a cabo un análisis del DNA de los dos implicados y de una muestra del tejido del feto que el cirujano había conservado, obteniéndose el resultado que recoge la tabla.

	locus DQ-alpha	**locus D1S80**
P. Harvey	1.1,1.3	18,24
K. Esquivel	4.0,4.0	24,25
Feto	1.1,4.0	18,24,25

De acuerdo con estos resultados el laboratorio emitió durante el juicio un informe en el que se afirmaba:

- "... da un índice de paternidad de 6,90. Esto significa que las apuestas genéticas en favor de la paternidad son 6,90 veces más probables a favor de que Harvey sea el padre biológico de que lo sea un varón aleatoriamente elegido entre la población caucásica norteamericana".
- "... usando un valor neutral del 50 % para las apuestas no genéticas en favor de la paternidad, obtenemos una probabilidad de paternidad del 87,34 %".

¿Cómo se obtuvieron estas cifras? Si denotamos mediante H=\{*Harvey es el padre biológico*\} y H^c=\{*Harvey NO es el padre biológico*\}, de acuerdo con las leyes de la genética y teniendo en cuenta que las frecuencias en la población de los alelos 1.1 y 18 son 13,7 % y 26,5 %, respectivamente, se obtiene

$$P(1.1 \text{ y } 18|H) = 0{,}5 \times 0{,}5 = 0{,}25,$$

y

$$P(1.1 \text{ y } 18|H^c) = 0{,}137 \times 0{,}265 = 0{,}0365,$$

donde \{*1.1 y 18*\}=\{*el feto posee los alelos 1.1 y 18*\}.

Lo que el informe denomina índice de paternidad no es más que el valor de la evidencia del fenotipo encontrado, es decir,

$$PI = \frac{P(1.1 \text{ y } 18|H)}{P(1.1 \text{ y } 18|H^c)} = \frac{0{,}25}{0{,}0365} = 6{,}90.$$

[3]Se entiende aquí *culpable* en el sentido de haber realizado verdaderamente la acción punible, no el hecho de serlo declarado por un juez o jurado

[4]El ejemplo está sacado de las notas del curso *Probability and Statistics for Law*, impartido por D. H. Kaye en la Universitat Pompeu Fabra de Barcelona en marzo de 2000

El valor neutral al que se refiere el informe supone asignar una probabilidad a priori 0,5 a H, lo que se traduce en que las apuestas *a priori* a favor de la paternidad de Harvey son de 1 a 1. Aplicando (1.4) para obtener las apuestas *a posteriori*

$$\frac{P(H|1.1 \text{ y } 18)}{P(H^c|1.1 \text{ y } 18)} = PI \times \frac{P(H)}{P(H^c)} = 6{,}90 \times 1 = 6{,}90.$$

La probabilidad de paternidad de Harvey, teniendo en cuenta la evidencia que los fenotipos aportan, puede calcularse mediante (1.3),

$$P(H|1.1 \text{ y } 18) = \frac{1}{\frac{1}{6{,}90} + 1} = \frac{6{,}90}{7{,}90} = 0{,}873,$$

valor aportado en el informe en forma de porcentaje.

Comentario acerca del informe del laboratorio.- El informe del laboratorio es incorrecto porque contiene dos errores que merecen ser comentados.

- El primero se refiere a la confusión entre el *índice de paternidad* y las *apuestas a favor de la paternidad*. Como ya hemos dicho el índice no es más que el valor de la evidencia, el cociente entre la probabilidad de que fuera Harvey quien aportara sus alelos y la probabilidad de que una extracción al azar de la población de genes aportara los alelos. Esta confusión es otra manera de presentarse la *falacia del fiscal*.
- La anterior objeción tiene una salvedad, ambos conceptos coinciden cuando las apuestas *a priori* a favor de la paternidad de Harvey son de 1 a 1, como ocurre en este caso. Pero para conseguirlo se ha asignado el valor 0,5 a $P(H)$, que el propio informe califica como *neutral* cuando *arbitrario* sería un calificativo más apropiado (asignar una probabilidad de 0,5 equivale, como ya dijimos anteriormente, a decidir la paternidad a cara o cruz). Un experto no necesita escoger un valor particular para las apuestas a priori. En su lugar debe dar una tabla de resultados como la que sigue, cuya valoración dejará en manos del juez o del jurado.

P(H)	***P(H|1.1 y 18)***
0,10	0,433
0,30	0,633
0,50	0,873
0,70	0,941
0,90	0,984

1.5. Independencia

La información previa que se nos proporcionó sobre el resultado del experimento modificó la probabilidad inicial del suceso. ¿Ocurre esto siempre? Veámoslo.

Supongamos que en lugar de comprar un único boleto, el que lleva el número 35, hubiéramos comprado todos aquellos que terminan en 5. Ahora $P(A) = 1/10$ puesto que hemos comprado 10 boletos, pero al calcular la probabilidad condicionada a la información que se nos ha facilitado, $B = \{$*el boleto premiado termina en 5*$\}$, observemos que $P(A \cap B) = 1/100$ porque al intersección de ambos sucesos es justamente el boleto que está premiado, en definitiva

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/100}{10/100} = \frac{1}{10},$$

la misma que originalmente tenía A. Parecen existir situaciones en las que la información previa no modifica la probabilidad inicial del suceso. Observemos que este resultado tiene una consecuencia inmediata,

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

Esta es una situación de gran importancia en probabilidad que recibe el nombre de *independencia* de sucesos y que generalizamos mediante la siguiente definición.

Definición 1.4 (Sucesos independientes) *Sean A y B dos sucesos. Decimos que A y B son* independientes *si $P(A \cap B) = P(A)P(B)$.*

De esta definición se obtiene como propiedad,

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

y su simétrica $P(B|A) = P(B)$.

En ocasiones se define la independencia de dos sucesos a partir de este resultado, obteniéndose entonces como propiedad la que nosotros hemos dado como definición. Existe equivalencia entre ambas definiciones, aunque a fuerza de ser rigurosos, hay que matizar que definir el concepto a partir de la probabilidad condicional exige añadir la condición de que el suceso condicionante tenga probabilidad distinta de cero. Hay además otra ventaja a favor de la definición basada en la factorización de $P(A \cap B)$, pone de inmediato en evidencia la *simetría* del concepto.

El concepto de independencia puede extenderse a una familia finita de sucesos de la siguiente forma.

Definición 1.5 (Independencia mutua) *Se dice que los sucesos de la familia $\{A_1, \ldots, A_n\}$ son* mútuamente independientes *cuando*

$$P(A_{k_1} \cap \ldots \cap A_{k_m}) = \prod_{i=1}^{m} P(A_{k_i}) \tag{1.6}$$

siendo $\{k_1, \ldots, k_m\} \subset \{1, \ldots, n\}$ y los k_i distintos.

Conviene señalar que la independencia mutua de los n sucesos supone que han de verificarse $\binom{n}{n} + \binom{n}{n-1} + \ldots \binom{n}{2} = 2^n - n - 1$ ecuaciones del tipo dado en (1.6).

Si solamente se verificasen aquellas igualdades que implican a dos elementos diríamos que los sucesos son *independientes dos a dos*, que es un tipo de independencia menos restrictivo que el anterior como pone de manifiesto el siguiente ejemplo. Solo cuando $n = 2$ ambos conceptos son equivalentes.

Ejemplo 1.7 *Tenemos un tetraedro con una cara roja, una cara negra, una cara blanca y la cuarta cara pintada con los tres colores. Admitimos que el tetraedro está bien construido, de manera que al lanzarlo sobre una mesa tenemos la misma probabilidad de que se apoye sobre una cualquiera de las cuatro caras, a saber, $p = \frac{1}{4}$. El experimento consiste en lanzar el tetraedro y ver en que posición ha caido. Si*

$R =$\{*el tetraedro se apoya en una cara con color* rojo\}

$N =$\{*el tetraedro se apoya en una cara con color* negro\}

$B =$\{*el tetraedro se apoya en una cara con color* blanco\},

se comprueba fácilmente que son independientes dos a dos pero no son mútuamente independientes.

El tipo de independencia habitualmente exigida es la mutua, a la que nos referiremos simplemente como *independencia*.

Digamos por último, que si la *colección de sucesos es infinita* diremos que son independientes cuando cualquier subcolección finita lo sea.

1.5.1. Independencia de clases de sucesos

Si A y B son sucesos independientes, ni A ni B nos proporcionan información sobre la ocurrencia del otro. Parece lógico que tampoco nos digan mucho sobre los complementarios respectivos. La pregunta es ¿son A y B^c independientes? La respuesta afirmativa la deducimos a continuación.

$$\begin{aligned} &P(A \cap B^c) = \\ &\quad = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c). \end{aligned}$$

Del mismo modo se comprueba que A^c es independiente tanto de B como de B^c. Resulta así que las conjuntos de sucesos $\{A, A^c\}$ y $\{B, B^c\}$ son *independientes* en el sentido que al tomar un suceso de cada una de ellos, los sucesos son independientes. De forma más general podemos hablar de clases independientes de sucesos.

Definición 1.6 (Clases independientes de sucesos) *Las clases de sucesos $\mathcal{A}_1, \ldots, \mathcal{A}_n \subset \mathcal{A}$ se dicen* independientes*, si al tomar A_i en cada $\mathcal{A}_i$, $i = 1, \ldots, n$, los sucesos de la familia $\{A_1, \ldots, A_n\}$ son independientes.*

Notemos que en la definición no se exige que los elementos de cada clase $\mathcal{A}_i$ sean independientes entre sí. De hecho A y A^c sólo lo son si $P(A) = 0$ ó $P(A) = 1$.

Para una *colección infinita de clases de sucesos* la anterior definición se extiende con facilidad. Diremos que $\{\mathcal{A}_n\}_{n \geq 1} \subset \mathcal{A}$ son *independientes* si cualquier subcolección finita lo es.

1.6. Probabilidades geométricas

La probabilidades definidas sobre los espacios de probabilidad descritos en los ejemplos 1.3 y 1.4 forman parte de una familia de probabilidades más general conocida como *probabilidades geométricas*. En efecto, si sobre cada uno de los espacios definimos la medida natural, λ_0, medida de conteo en el caso de un conjunto discreto y λ_2, medida de Lebesgue en $\mathcal{R}^2$ en el caso del círculo unidad, la probabilidad definida en ambos casos se reduce a

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}, \quad \forall A \in \mathcal{A}, \tag{1.7}$$

donde μ es la correspondiente medida. Que (1.7) es una probabilidad es consecuencia inmediata de las propiedades de las medidas. La probabilidad (1.7) se extiende de inmediato a cualquier $\Omega \subset \mathcal{R}^k$, tal que $\Omega \in \beta^k$ y $\lambda_k(\Omega) < +\infty$.

Le Comte de Buffon

El origen de la probabilidad geométrica se debe a George Louis Leclerc, Conde de Buffon, eminente naturalista francés, quién en el año 1777 publicó una obra titulada *Essai d'Arithmétique Morale* con el objeto de reivindicar a la Geometría como ciencia capaz de aportar soluciones a los problemas de azar, capacidad hasta entonces exclusivamente atribuida a la Aritmética. En ella plantea y resuelve el siguiente problema:

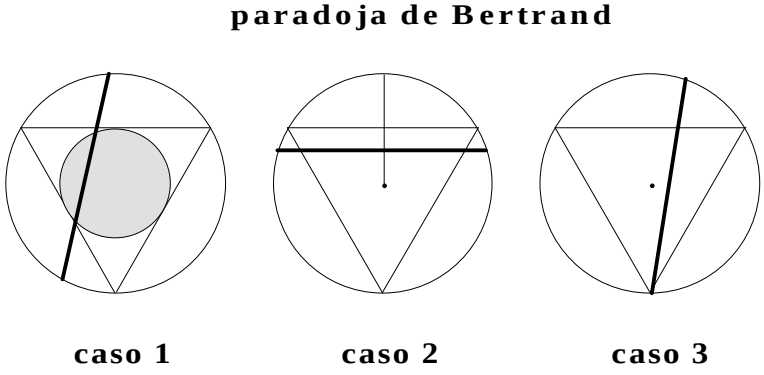
¿Cuál es la probabilidad de que una aguja que se lanza al azar sobre un entramado de líneas paralelas equidistantes, con distancia entre ellas mayor que la longitud de la aguja, corte a alguna de las paralelas?

Con el tiempo se convirtió en una referencia clásica y pasó a ser conocido en la literatura probabilística como *el problema de la aguja de Buffon*. En la solución aportada por Buffon se ponía en evidencia que el problema requería *medir*, a diferencia de lo que ocurría con las soluciones a los problemas relacionados con juegos de azar discretos en los que bastaba *contar*. Nos ocuparemos de su solución más tarde, cuando hayamos introducido el concepto de vector aleatorio. Para ilustrar la aplicación de las probabilidades geométricas utilizaremos otro problema clásico conocido como la paradoja de Bertrand.

1.6.1. La paradoja de Bertrand

Un ejemplo clásico de probabilidad geométrica es el que se conoce como *paradoja de Bertrand*[5]. Como veremos a continuación la paradoja reside en el hecho de existir, aparentemente, varias soluciones al problema que se plantea. Veamos su enunciado y cómo se resuelve la paradoja.

La paradoja de Bertrand.- Elegimos una cuerda al azar en el círculo unidad. ¿Cuál es la probabilidad de que su longitud supere la del lado del triángulo equilátero inscrito en el círculo?



[5]Joseph Louis François Bertrand (1822-1900) fue profesor de l'École Polytechnique de Paris y del Collège de France. Aunque es más conocido por la conjetura de teoría de números que lleva su nombre y que fue demostrada por Chebyshev en 1850 (para todo $n > 3$, existe siempre un número primo entre n y $2n - 2$), Bertrand trabajó también en geometría diferencial y en teoría de la probabilidad.

Solución.- La paradoja estriba en que la respuesta parece no ser única. En efecto, el valor de la probabilidad que se nos pide depende del significado que demos a *la elección al azar*. La longitud de una cuerda en un círculo puede calcularse a partir de,

1. la distancia al centro de su punto medio,
2. la posición de su punto medio sobre un radio cualquiera,
3. la posición de uno de sus extremos sobre la circunferencia, supuesto fijo el otro extremo.

Cada una de estas interpretaciones supone una elección al azar sobre espacios muestrales distintos. Así,

Caso 1.- El espacio muestral es todo el círculo unidad, C_1 y sólo las cuerdas cuyos puntos medios caen en el círculo inscrito en el triángulo equilátero, $C_{1/2}$, tienen longitud mayor que $\sqrt{3}$. Es sabido que este círculo tiene radio $1/2$ y recurriendo a la probabilidad geométricas, si A=$\{$*la cuerda tiene longitud mayor que* $\sqrt{3}\}$

$$P(A) = \frac{\text{área}(C_{1/2})}{\text{área}(C_1)} = \frac{\pi(1/2)^2}{\pi} = \frac{1}{4}.$$

Caso 2.- El espacio muestral es ahora el segmento (radio) [0,1] y sólo las cuerdas cuyos puntos medios están en [0,1/2] cumplen la condición. Tendremos

$$P(A) = \frac{\text{longitud}([0,1/2])}{\text{longitud}([0,1])} = \frac{1}{2}.$$

Caso 3.- El espacio muestral es ahora la circunferencia del círculo unidad. Si fijamos un extremo de la cuerda y elegimos al azar la posición del otro, sólo aquellas cuerdas que tengan este último extremo sobre el tercio de la circunferencia opuesto al extremo fijo cumplen la condición. Tendremos

$$P(A) = \frac{2\pi/3}{2\pi} = \frac{1}{3}.$$

Capítulo 2

Variables y vectores aleatorios

2.1. Introducción

Cuando nos hemos referido en el capítulo anterior a los distintos sucesos con los que hemos ilustrado los ejemplos, lo hemos hecho aludiendo a características numéricas ligadas al resultado del experimento. Así, nos referíamos a {*puntos que distan a lo sumo 1/2 del centro del círculo*}, a {*la suma de las caras del dado es 8*} o a {*número de caras que muestran un número par de puntos*}. Pero los ejemplos podrían ser otros muchos e involucrar más de una característica numérica simultáneamente:

- número de llamadas que llegan a una centralita telefónica en un intervalo de tiempo,
- altura y peso de un individuo,
- suma y valor absoluto de la diferencia de las caras que muestran dos dados al ser lanzados.

En resumen, nuestro interés al examinar el resultado de un experimento aleatorio no es tanto el espacio de probabilidad resultante, como la o las características numéricas asociadas, lo que supone cambiar nuestro objetivo de Ω a $\mathcal{R}$ o $\mathcal{R}^k$. Hay dos razones que justifican este cambio:

1. el espacio de probabilidad es un espacio abstracto, mientras que $\mathcal{R}$ o $\mathcal{R}^k$ son espacios bien conocidos en los que resulta mucho más cómodo trabajar,
2. fijar nuestra atención en las características numéricas asociadas a cada resultado implica un proceso de abstracción que, al extraer los rasgos esenciales del espacio muestral, permite construir un modelo probabilístico aplicable a todos los espacios muestrales que comparten dichos rasgos.

Puesto que se trata de características numéricas ligadas a un experimento aleatorio, son ellas mismas *cantidades aleatorias*. Esto supone que para su estudio y conocimiento no bastará con saber que valores toman, habrá que conocer además la probabilidad con que lo hacen. De todo ello nos vamos a ocupar a continuación.

2.2. Variable aleatoria

La única forma que conocemos de trasladar información de un espacio a otro es mediante una aplicación. En nuestro caso la aplicación habrá de trasladar el concepto de suceso, lo que exige una mínima infraestructura en el espacio receptor de la información semejante a la σ-álgebra

que contiene a los sucesos. Como nos vamos a ocupar ahora del caso unidimensional, una sola característica numérica asociada a los puntos del espacio muestral, nuestro espacio imagen es $\mathcal{R}$. En $\mathcal{R}$, los intervalos son el lugar habitual de trabajo y por tanto lo más conveniente será exigir a esta infraestructura que los contenga. Existe en $\mathcal{R}$ la llamada σ-álgebra de Borel, β, que tiene la propiedad de ser la menor de las que contienen a los intervalos, lo que la hace la más adecuada para convertir a $\mathcal{R}$ en espacio probabilizable: $(\mathcal{R}, \beta)$. Estamos ahora en condiciones de definir la variable aleatoria.

Definición 2.1 (Variable aleatoria) *Consideremos los dos espacios probabilizables* $(\Omega, \mathcal{A})$ *y* $(\mathcal{R}, \beta)$. *Una* variable aleatoria *es un aplicación,* $X : \Omega \longrightarrow \mathcal{R}$, *que verifica*

$$X^{-1}(B) \in \mathcal{A}, \quad \forall B \in \beta. \tag{2.1}$$

En el contexto más general de la Teoría de la Medida, una aplicación que verifica (2.1) se dice que es una *aplicación medible*. De acuerdo con ello, una variable aleatoria no es más que una aplicación medible entre Ω y $\mathcal{R}$.

Cuando hacemos intervenir una variable aleatoria en nuestro proceso es porque ya estamos en presencia de un espacio de probabilidad, $(\Omega, \mathcal{A}, P)$. La variable aleatoria traslada la información probabilística relevante de Ω a $\mathcal{R}$ mediante una *probabilidad inducida* que se conoce como *ley de probabilidad de* X o *distribución de probabilidad de* X.

El concepto de σ-álgebra inducida

Dada la definición de variable aleatoria, es muy sencillo comprobar el siguiente resultado.

Lema 2.1 *La familia de sucesos*

$$\sigma(X) = \{X^{-1}(B),\ \forall B \in \beta\} = \{X^{-1}(\beta)\},$$

es una σ*-álgebra y además se verifica* $\sigma(X) \subset \mathcal{A}$

A la $\sigma(X)$ se la denomina σ*-álgebra inducida por* X.

2.2.1. Probabilidad inducida

X induce sobre $(\mathcal{R}, \beta)$ una probabilidad, P_X, de la siguiente forma,

$$P_X(A) = P(X^{-1}(A)),\ \forall A \in \beta.$$

Es fácil comprobar que P_X es una probabilidad sobre la σ-álgebra de Borel, de manera que $(\mathcal{R}, \beta, P_X)$ es un espacio de probabilidad al que podemos aplicar todo cuanto se dijo en el capítulo anterior. Observemos que P_X hereda las características que tenía P, pero lo hace a través de X. ¿Qué quiere esto decir? Un ejemplo nos ayudará a comprender este matiz.

Ejemplo 2.1 *Sobre el espacio de probabilidad resultante de lanzar dos dados, definimos las variables aletorias,* X=suma de las caras *e* Y=valor absoluto de la diferencia de las caras. *Aun cuando el espacio de probabilidad sobre el que ambas están definidas es el mismo,* P_X *y* P_Y *son distintas porque viene inducidas por variables distintas. En efecto,*

$$P_X(\{0\}) = P(X^{-1}(\{0\}) = P(\emptyset) = 0,$$

sin embargo,

$$P_Y(\{0\}) = P(Y^{-1}(\{0\}) = P(\{1,1\},\{2,2\},\{3,3\},\{4,4\},\{5,5\},\{6,6\}) = \frac{1}{6}.$$

La distribución de probabilidad de X, P_X, nos proporciona cuanta información necesitamos para conocer el comportamiento probabilístico de X, pero se trata de un objeto matemático complejo de incómodo manejo, al que no es ajena su condición de función de conjunto. Esta es la razón por la que recurrimos a funciones de punto para describir la aleatoriedad de X.

2.2.2. Función de distribución de probabilidad

A partir de la probabilidad inducida podemos definir sobre $\mathcal{R}$ la siguiente función,

$$F_X(x) = P_X((-\infty, x]) = P(X^{-1}\{(-\infty, x]\}) = P(X \leq x),\ \forall x \in \mathcal{R}. \tag{2.2}$$

Así definida esta función tiene las siguientes propiedades:

PF1) No negatividad.- Consecuencia inmediata de su definición.

PF2) Monotonía.- De la monotonía de la probabilidad se deduce fácilmente que $F_X(x_1) \leq F_X(x_2)$ si $x_1 \leq x_2$.

PF3) Continuidad por la derecha.- Consideremos una sucesión decreciente de números reales $x_n \downarrow x$. La correspondiente sucesión de intervalos verifica $\cap_n(-\infty, x_n] = (-\infty, x]$, y por la continuidad desde arriba de la probabilidad respecto del paso al límite tendremos $\lim_{x_n \downarrow x} F_X(x_n) = F_X(x)$.

Observemos por otra parte que $(-\infty, x] = \{x\} \cup \lim_{n\to+\infty}(-\infty, x - \frac{1}{n}]$, lo que al tomar probabilidades conduce a

$$F_X(x) = P(X = x) + \lim_{n\to+\infty} F_X\left(x - \frac{1}{n}\right) = P(X = x) + F(x-), \tag{2.3}$$

A partir de 2.3 se sigue que $F_X(x)$ *es continua en* x sí y solo sí $P(X = x) = 0$.

PF4) Valores límites.- Si $x_n \uparrow +\infty$ o $x_n \downarrow -\infty$ entonces $(-\infty, x_n] \uparrow \mathcal{R}$ y $(-\infty, x_n] \downarrow \emptyset$ y por tanto

$$F(+\infty) = \lim_{x_n \uparrow +\infty} F(x_n) = 1, \quad F(-\infty) = \lim_{x_n \downarrow -\infty} F(x_n) = 0.$$

A la función F_X se la conoce como *función de distribución de probabilidad de* X (en adelante simplemente función de distribución). En ocasiones se la denomina función de distribución acumulada, porque tal y como ha sido definida nos informa de la probabilidad acumulada por la variable X hasta el punto x. Nos permite obtener probabilidades del tipo $P(a < X \leq b)$ a partir de la expresión

$$P(a < X \leq b) = F_X(b) - F_X(a).$$

Si queremos que F_X sustituya con éxito a P_X en la descripción del comportamiento de X, debe proporcionar la misma información que ésta. En otras palabras, ambos conceptos deben ser equivalentes. Hasta ahora hemos visto que $P_X \Longrightarrow F_X$, en el sentido que la primera nos ha permitido obtener la segunda, pero, ¿es cierta la implicación contraria? La respuesta es afirmativa, porque la probabilidad no es más que un caso particular de medida de Lebesgue-Stieltjes y, como se demuestra en Teoría de la Medida, es posible recuperar P_X a partir de F_X mediante la definición sobre la familia de intervalos de la forma $]a, b]$ de la función

$$P'(]a, b]) = F_X(b) - F_X(a). \tag{2.4}$$

El teorema de extensión de Caratheodory permite extender P' sobre toda la σ-álgebra de Borel y demostrar que coincide con P_X. Así pues, $P_X \Longleftrightarrow F_X$.

En realidad el resultado va más allá de lo expuesto. Puede demostrarse una especie de teorema existencia según el cual, cualquier función que verifique las propiedades PF1) a PF4) define, mediante (2.4), una medida de Lebesgue-Stieltjes que es una probabilidad sobre el espacio $(\mathcal{R}, \beta)$, existiendo además una variable aleatoria X que la tiene por su distribución de probabilidad.

2.2.3. Variable aleatoria discreta. Función de cuantía

Existe una segunda función de punto que permite describir el comportamiento de X, pero para introducirla hemos de referirnos primero a las características del *soporte* de X, entendiendo por tal un conjunto $D_X \in \beta$ que verifica, $P_X(D_X) = P(X \in D_X) = 1$.

Cuando D_X es numerable, P_X es discreta y decimos que X es una variable aleatoria *discreta*. Como ya vimos en un ejemplo del capítulo anterior, $P_X(\{x_i\}) = P(X = x_i) > 0,\ \forall x_i \in D_X$, y siendo además $P(X \in D_X) = 1$, se deduce $P(X = x) = 0,\ \forall x \in D_X^c$. En este caso es fácil comprobar que la F_X asociada viene dada por

$$F_X(x) = \sum_{x_i \leq x} P(X = x_i). \tag{2.5}$$

De acuerdo con esto, si $x_{(i)}$ y $x_{(i+1)}$ son dos puntos consecutivos del soporte tendremos que $\forall x \in [x_{(i)}, x_{(i+1)}[$, $F_X(x) = F_X(x_{(i)})$. Como además $P_X(x) = 0,\ \forall x \in D_X^c$, la función será también continua. Por otra parte $P(X = x_i) > 0$, para $x_i \in D_X$, con lo que los únicos puntos de discontinuidad serán lo del soporte, discontinuidad de salto finito cuyo valor es $F_X(x_{(i)}) - F_X(x_{(i-1)}) = P(X = x_i)$. Se trata por tanto de una función escalonada, cuyos saltos se producen en los puntos de D_X.

A la variable aleatoria discreta podemos asociarle una nueva función puntual que nos será de gran utilidad. La definimos para cada $x \in \mathcal{R}$ mediante $f_X(x) = P_X(\{x\}) = P(X = x)$, lo que supone que

$$f_X(x) = \begin{cases} P(X = x), & \textit{si } x \in D_X \\ 0, & \textit{en el resto.} \end{cases}$$

Esta función es conocida como *función de cuantía* o de *probabilidad* de X y posee las dos propiedades siguientes:

Pfc1) Al tratarse de una probabilidad, $f_X(x) \geq 0,\ \forall x \in \mathcal{R}$,

Pfc2) Como $P(X \in D_X) = 1$,

$$\sum_{x_i \in D_X} f_X(x_i) = 1.$$

La relación entre f_X y F_X viene recogida en las dos expresiones que siguen, cuya obtención es evidente a partir de (2.3) y (2.5). La primera de ellas permite obtener F_X a partir de f_X,

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i).$$

La segunda proporciona f_X en función de F_X,

$$f_X(x) = F_X(x) - F_X(x-).$$

De ambas expresiones se deduce la equivalencia entre ambas funciones y si recordamos la equivalencia antes establecida podemos escribir,

$$P_X \Longleftrightarrow F_X \Longleftrightarrow f_X.$$

2.2.4. Algunos ejemplos de variables aleatorias discretas

Variable aleatoria Poisson

La distribución de *Poisson* de parámetro λ es una de las distribuciones de probabilidad discretas más conocida. Una variable con esta distribución se caracteriza porque su soporte es $D_X = \{0, 1, 2, 3, \ldots\}$ y su función de cuantía viene dada por

$$f_X(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & si\ x \in D_X \\ \\ 0, & en\ el\ resto, \end{cases}$$

que cumple las propiedades Pfc1) y Pfc2). La función de distribución tiene por expresión

$$F_X(x) = \sum_{n \leq x} \frac{e^{-\lambda}\lambda^n}{n!}.$$

Diremos que X es una *variable Poisson de parámetro* λ y lo denotaremos $X \sim Po(\lambda)$. Esta variable aparece ligada a experimentos en los que nos interesa la ocurrencia de un determinado suceso a lo largo de un intervalo finito de tiempo[1], verificándose las siguientes condiciones:

1. la probabilidad de que el suceso ocurra en un intervalo pequeño de tiempo es proporcional a la longitud del intervalo, siendo λ el factor de proporcionalidad,

2. la probabilidad de que el suceso ocurra en dos o más ocasiones en un intervalo pequeño de tiempo es prácticamente nula.

Fenómenos como el número de partículas que llegan a un contador Geiger procedentes de una fuente radiactiva, el número de llamadas que llegan a una centralita telefónica durante un intervalo de tiempo, las bombas caídas sobre la región de Londres durante la Segunda Guerra mundial y las bacterias que crecen en la superficie de un cultivo, entre otros, pueden ser descritos mediante una variable aleatoria Poisson.

Variable aleatoria Binomial

Decimos que X es una *variable Binomial* de parámetros n y p ($X \sim B(n, p)$) si $D_X = \{0, 1, 2, \ldots, n\}$ y

$$f_X(x) = \begin{cases} \dbinom{n}{x} p^x (1-p)^{n-x}, & si\ x \in D_X \\ \\ 0, & en\ el\ resto, \end{cases}$$

que se comprueba fácilmente que verifica Pfc1) y Pfc2).

Cuando llevamos a cabo un experimento aleatorio cuyos rasgos esenciales son:

1. se llevan a cabo n repeticiones independientes de una misma prueba en las mismas condiciones,

2. en cada repetición observamos la ocurrencia (*éxito*) o no (*fracaso*) de un mismo suceso, A, y

[1] En un planteamiento más general, el intervalo finito de tiempo puede ser sustituido por un subconjunto acotado de $\mathcal{R}^k$

3. la probabilidad de éxito es la misma en cada repetición, $P(A) = p$,

la variable que describe el número de éxitos alcanzado en las n repeticiones, es una Binomial de parámetros n y p.

Fenómenos aleatorios aparentemente tan diferentes como el número de hijos varones de un matrimonio con n hijos o el número de caras obtenidas al lanzar n veces una moneda correcta, son bien descritos mediante un variable Binomial. Este hecho, o el análogo que señalábamos en el ejemplo anterior, ponen de manifiesto el papel de modelo aleatorio que juega una variable aleatoria, al que aludíamos en la introducción. Esta es la razón por la que en muchas ocasiones se habla del *modelo Binomial* o del *modelo Poisson*.

Hagamos por último hincapié en un caso particular de variable aleatoria Binomial. Cuando $n = 1$ la variable $X \sim B(1, p)$ recibe el nombre de *variable Bernoulli* y se trata de una variable que solo toma los valores 0 y 1 con probabilidad distinta de cero. Es por tanto una variable dicotómica asociada a experimentos aleatorios en los que, realizada una sola prueba, nos interesamos en la ocurrencia de un suceso o su complementario. Este tipo de experimentos reciben el nombre de *pruebas Bernoulli*.

La distribución de Poisson como límite de la Binomial.- Consideremos la sucesión de variables aleatorias $X_n \sim B(n, p_n)$ en la que a medida que n aumenta, p_n disminuye de forma tal que $np_n \approx \lambda$. Más concretamente, $np_n \to \lambda$. Tendremos para la función de cuantía,

$$f_{X_n}(x) = \binom{n}{x} p_n^x (1-p_n)^{n-x} = \frac{n!}{x!(n-x)!} p_n^x (1-p_n)^{n-x},$$

y para n suficientemente grande,

$$\begin{aligned} f_{X_n}(x) &\approx \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\frac{n(n-1)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}. \end{aligned}$$

Al pasar al límite,

$$\frac{n(n-1)\cdots(n-x+1)}{n^x} \to 1, \quad \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}, \quad \left(1-\frac{\lambda}{n}\right)^{-x} \to 1,$$

y tendremos

$$\lim_{n\to+\infty} f_{X_n}(x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

La utilidad de este resultado reside en permitir la aproximación de la función de cuantía de una $B(n, p)$ mediante la función de cuantía de una $Po(\lambda = np)$ cuando n es grande y p pequeño.

Variable aleatoria Hipergeométrica. Relación con el modelo Binomial

Si tenemos una urna con N bolas, de las cuales r son rojas y el resto, $N - r$, son blancas y extraemos n de ellas *con reemplazamiento*, el número X de bolas rojas extraídas será una $B(n, p)$ con $p = r/N$.

¿Qué ocurre si llevamos a cabo las extracciones *sin reemplazamiento*? La variable X sigue ahora una *distribución Hipergeométrica* ($X \sim H(n, N, r)$) con soporte $D_X = \{0, 1, 2, \ldots, \text{mín}(n, r)\}$

y cuya función de cuantía se obtiene fácilmente a partir de la fórmula de Laplace

$$f_X(x) = \begin{cases} \dfrac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, & si\ x \in D_X \\ \\ 0, & en\ el\ resto, \end{cases}$$

que cumple de inmediato la condición Pfc1). Para comprobar Pfc2) debemos hacemos uso de una conocida propiedad de los números combinatorios,

$$\sum_{i=0}^{n} \binom{a}{i}\binom{b}{n-i} = \binom{a+b}{n}.$$

La diferencia entre los modelos Binomial e Hipergeométrico estriba en el tipo de extracción. Cuando ésta se lleva a cabo con reemplazamiento las sucesivas extracciones son independientes y la probabilidad de *éxito* se mantiene constante e igual a r/N, el modelo es Binomial. No ocurre así si las extracciones son sin reemplazamiento. No obstante, si n es muy pequeño respecto a N y r, la composición de la urna variará poco de extracción a extracción y existirá lo que podríamos denominar una *quasi-independencia* y la distribución Hipergeométrica deberá comportarse como una Binomial. En efecto,

$$\begin{aligned} f_X(x) &= \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \\ &= \frac{r!}{x!(r-x)!} \times \frac{(N-r)!}{(n-x)!(N-r-n+x)!} \times \frac{n!(N-n)!}{N!} \\ &= \binom{n}{x}\frac{r}{N} \times \frac{r-1}{N-1} \times \cdots \times \frac{r-x+1}{N-x+1} \times \frac{N-r}{N-x} \times \frac{N-r-1}{N-x-1} \times \\ &\quad \cdots \times \frac{N-r-n+x+1}{N-n+1} \\ &\approx \binom{n}{x} p^x (1-p)^{n-x}, \end{aligned}$$

con $p = r/N$.

Estimación del tamaño de una población animal a partir de datos de recaptura[2].- Queremos estimar la población de peces en un lago, para ello hemos capturado 1000 peces a los que, marcados mediante una mancha roja, hemos arrojado nuevamente al lago. Transcurrido un cierto tiempo, el necesario para que se mezclen con los restantes peces del lago, llevamos a cabo una nueva captura de otros 1000 peces entre los que hay 100 marcados. ¿Qué podemos decir acerca del total de peces en el lago?

El problema que planteamos en un problema típico de *estimación estadística* y vamos a dar una solución que, aunque particular para la situación descrita, está basada en una

[2]El ejemplo está sacado del libro de W. Feller (1968), *An Introduction to Probability Theory and Its Application*, Vol. I, 3rd. Edition, un libro clásico cuya lectura y consulta recomendamos vivamente

metodología de aplicación general en los problemas de estimación. Observemos en primer lugar que el número de peces marcados en la segunda captura (recaptura) es una variable aleatoria Hipergeométrica, $X \sim H(1000, N, 1000)$, siempre bajo el supuesto de que ambas capturas constituyen sendas muestras aleatorias de la población total de peces del lago (en la práctica semejante suposición excluye situaciones en las que las capturas se efectúan en el mismo lugar y en un corto periodo de tiempo). Suponemos también que el número de peces en el lago, N, no cambia entre las dos capturas.

Generalizemos el problema admitiendo tamaños arbitrarios para ambas muestras:

$$\begin{aligned} N &= \text{población de peces en el lago (desconocida)} \\ r &= \text{número de peces en la } 1\underline{\text{a}} \text{ captura} \\ n &= \text{número de peces en la } 2\underline{\text{a}} \text{ captura} \\ x &= \text{número de peces en con mancha roja en la } 2\underline{\text{a}} \text{ captura} \\ p_x(N) &= \text{probabilidad de } x \text{ peces con mancha roja en la } 2\underline{\text{a}} \text{ captura} \end{aligned}$$

Con esta formulación sabemos que

$$p_x(N) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}.$$

En la práctica, r, n y x son conocidos por observación, como en el ejemplo que planteamos, mientras que N es desconocido pero fijo y en modo alguno depende del azar. Al menos una cosa conocemos de N y es que $N \geq r + n - x$, que es el total de peces capturados entre ambas capturas. En nuestro ejemplo, $N \geq 1000 + 1000 - 100 = 1900$. ¿Qué ocurre si aceptamos $N = 1900$? Aunque se trata de un valor teóricamente posible, si calculamos $p_{100}(1900)$,

$$p_{100}(1900) = \frac{\binom{1000}{100}\binom{900}{900}}{\binom{1900}{1000}} \approx 10^{-430},$$

(podemos valernos de la fórmula de Stirling, $n! \approx \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$, para aproximar las factoriales), habremos de aceptar que ha ocurrido un suceso, $X = 100$, con una probabilidad extraordinariamente pequeña. Resulta difícil de admitir una hipótesis que exige casi un milagro para que el suceso observado tenga lugar. Otro tanto nos ocurre si suponemos que N es muy grande, por ejemplo $N = 10^6$. También ahora $p_{100}(10^6)$ es muy pequeña.

Una respuesta adecuada puede ser la de buscar el valor de N que maximiza $p_x(N)$. Dicho valor, que designamos mediante $\hat{N}$, recibe el nombre de *estimación máximo-verosímil* de N. Para encontrarlo, observemos que

$$\frac{p_x(N)}{p_x(N-1)} = \frac{(N-r)(N-n)}{(N-r-n+x)N} = \frac{N^2 - Nr - Nn + rn}{N^2 - Nr - Nn + Nx},$$

de donde se deduce

$$\begin{aligned} p_x(N) > p_x(N-1), &\qquad \text{si } Nx < rn, \\ p_x(N) < p_x(N-1), &\qquad \text{si } Nx > rn. \end{aligned}$$

Así pues, a medida que aumenta N la función $p_x(N)$ crece primero para decrecer después, alcanzando su máximo en $N = [rn/x]$, la parte entera de rn/x. En nuestro ejemplo, $\hat{N} = 10000$.

Variable aleatoria Binomial Negativa

Consideremos n pruebas Bernoulli independientes con la misma probabilidad de éxito, p, en cada una de ellas. Nos interesamos en la ocurrencia del r-ésimo éxito. La variable que describe el mínimo número de pruebas adicionales necesarias para alcanzar los r éxitos, es una *variable aleatoria Binomial Negativa*, $X \sim BN(r,p)$, con soporte numerable $D_X = \{0,1,2,\ldots\}$ y con función de cuantía

$$f_X(x) = \begin{cases} \binom{r+x-1}{x} p^r (1-p)^x, & si\ x \geq 0 \\ \\ 0, & en\ el\ resto. \end{cases}$$

El nombre de Binomial negativa se justifica a partir de la expresión alternativa que admite la función de cuantía,

$$f_X(x) = \begin{cases} \binom{-r}{x} p^r (-(1-p))^x, & si\ x \geq 0 \\ \\ 0, & en\ el\ resto, \end{cases}$$

obtenida al tener en cuenta que

$$\begin{aligned} \binom{-r}{x} &= \frac{(-r)(-r-1)\cdots(-r-x+1)}{x!} \\ &= \frac{(-1)^x r(r+1)\cdots(r+x-1)}{x!} \\ &= \frac{(-1)^x (r-1)! r(r+1)\cdots(r+x-1)}{(r-1)!x!} \\ &= (-1)^x \binom{r+x-1}{x}. \end{aligned}$$

La condición Pfc1) se cumple de inmediato, en cuanto a Pfc2) recordemos el desarrollo en serie de potencias de la función $f(x) = (1-x)^{-n}$,

$$\frac{1}{(1-x)^n} = \sum_{i \geq 0} \binom{n+i-1}{i} x^i, \quad |x| < 1.$$

En nuestro caso

$$f_X(x) = \sum_{x \geq 0} \binom{r+x-1}{x} p^r (1-p)^x = p^r \sum_{x \geq 0} \binom{r+x-1}{x} (1-p)^x = p^r \frac{1}{(1-(1-p))^r} = 1.$$

Un caso especial con nombre propio es el de $r = 1$. La variable aleatoria $X \sim BN(1,p)$ recibe el nombre de *variable aleatoria Geométrica* y su función de cuantía se reduce a

$$f_X(x) = \begin{cases} p(1-p)^x, & si\ x \geq 0 \\ \\ 0, & en\ el\ resto. \end{cases}$$

El problema de las cajas de cerillas de Banach.- En un acto académico celebrado en honor de Banach, H. Steinhaus contó una anécdota acerca del hábito de fumar que aquél

tenía. La anécdota se refería a la costumbre de Banach de llevar una caja de cerillas en cada uno de los bolsillos de su chaqueta, de manera que cuando necesitaba una cerilla elegía al azar uno de los bolsillos. El interés de la anécdota residía en calcular las probabilidades asociadas al número de cerillas que habría en una caja cuando, por primera vez, encontrara vacía la otra.

Si cada caja contiene N cerillas, en el momento de encontrar una vacía la otra puede contener $0, 1, 2, \ldots, N$ cerillas. Designemos por $A_r = \{$*el bolsillo no vacío contiene* r *cerillas*$\}$. Supongamos que la caja vacía es la del bolsillo izquierdo, para que ello ocurra $N-r$ *fracasos* (elecciones del bolsillo derecho) deben haber precedido al $N+1$-ésimo *éxito* (elección del bolsillo derecho). En términos de una variable aleatoria $X \sim BN(N+1, 1/2)$ se trata de obtener $P(X = N-r)$. El mismo argumento puede aplicarse si la caja vacía es la del bolsillo derecho. Así pues,

$$p_r = P(A_r) = 2P(X = N-r) = 2\binom{2N-r}{N-r}\left(\frac{1}{2}\right)^{N+1}\left(\frac{1}{2}\right)^{N-r} = \binom{2N-r}{N-r}2^{-2N+r}.$$

Por ejemplo, para $N = 50$ y $r = 4$, $p_r = 0{,}074790$; para $r = 29$, $p_r = 0{,}000232$.

Observación 2.1 *Es también habitual presentar la variable Binomial negativa como el número total de pruebas necesarias para alcanzar el r-ésimo éxito. En este caso, el soporte de la variable es $D_X = \{r, r+1, r+2, \ldots\}$ y su función de cuantía tiene la expresión*

$$f_X(x) = \begin{cases} \binom{x-1}{x-r}p^r(1-p)^{x-r}, & si\ x \geq r \\ \\ 0, & en\ el\ resto. \end{cases}$$

2.2.5. Variable aleatoria continua. Función de densidad de probabilidad

Para introducir el otro tipo de variables aleatorias que merecerán nuestra atención, las variables aleatorias continuas, hemos de recordar primero el concepto de *continuidad absoluta*. Como ya dijimos, la variable aleatoria nos permite probabilizar el espacio $(\mathcal{R}, \beta)$ mediante la probabilidad inducida P_X. Sobre este espacio, una medida muy habitual es la *medida de Lebesgue*, λ. La continuidad absoluta establece una relación interesante y fructífera entre ambas medidas.

Definición 2.2 (Continuidad absoluta para P_X) *Decimos que P_X es* absolutamente continua *respecto de la medida de Lebesgue, λ, si existe una función no negativa, f, tal que $\forall B \in \beta$ se verifica*

$$P_X(B) = \int_B f\, d\lambda.$$

Como probabilidad inducida y función de distribución son equivalentes, la anterior definición tiene también su equivalente en términos de F_X.

Definición 2.3 (Continuidad absoluta para F_X) *Decimos que F_X es* absolutamente continua *si para cada ϵ existe un δ tal que para cualquier familia finita de intervalos disjuntos, $[a_i, b_i], i = 1, \ldots, k$*

$$\sum_{i=1}^{k} |F_X(b_i) - F_X(a_i)| < \epsilon \quad si \quad \sum_{i=1}^{k}(b_i - a_i) < \delta.$$

Observemos que con esta definición, la continuidad uniforme es un caso particular cuando la familia de intervalos contiene un único elemento. Ello supone que F_X es continua.

Puede demostrarse que ambas definiciones son equivalentes, lo que nos permite escribir

$$F_X(x) = P(X \leq x) = \int_{]-\infty,x]} f \, d\lambda.$$

Pero si, como ocurre con frecuencia y será siempre nuestro caso, la función f es integrable Riemann, la integral anterior se reduce a

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t) \, dt. \tag{2.6}$$

Con todo cuanto precede diremos que la variable aleatoria X es *continua* si F_X es absolutamente continua. A efectos prácticos ello supone que existe una función $f_X(x)$, conocida como *función de densidad de probabilidad (fdp)* de X, tal que $\forall B \in \beta$

$$P(X \in B) = \int_B f_X \, d\lambda,$$

que, supuesta la integrabilidad Riemann de f_X, puede escribirse

$$P(X \in]a, b]) = P(a < X \leq b) = \int_a^b f(x) \, dx.$$

Se derivan para f_X dos interesantes propiedades que la caracterizan:

Pfdp1) $f_X(x)$ es no negativa, y

Pfdp2) como $P(X \in \mathcal{R}) = 1$,

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

Como consecuencia de esta definición, entre la función de distribución y la de densidad de probabilidad se establecen las siguientes relaciones

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

y si x es un punto de continuidad de f_X,

$$f_X(x) = F_X^{'}(x).$$

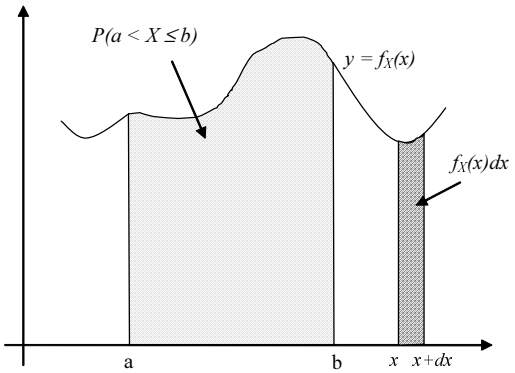
Esta última igualdad merece matizarse. En efecto, puesto que el origen de la densidad está en la continuidad absoluta de P_X respecto de la medida de Lebesgue, cualquier función que difiera de f_X en un conjunto con medida de Lebesgue nula será también una densidad. En otras palabras, la densidad de probabilidad no es única. Por ello sería más riguroso decir que $F_X^{'}(x)$ es una de las posibles densidades.

Al igual que ocurría en el caso discreto, las dos relaciones anteriores implican la equivalencia entre ambas funciones y podemos escribir,

$$P_X \Longleftrightarrow F_X \Longleftrightarrow f_X.$$

Significado físico de la fdp

La continuidad de F_X implica, recordemos (2.3), que en las variables aleatorias continuas $P(X = x) = 0,\ \forall x \in \mathcal{R}$. Este es un resultado que siempre sorprende y para cuya comprensión es conveniente interpretar físicamente la función de densidad de probabilidad.



La fdp es sencillamente eso, una densidad lineal de probabilidad que nos indica la cantidad de probabilidad por elemento infinitesimal de longitud. Es decir, $f_X(x)\,dx \approx P(X \in]x, x+dx])$. Ello explica que, para elementos con longitud cero, sea nula la correspondiente probabilidad. En este contexto, la probabilidad obtenida a través de la integral de Riemann pertinente se asimila a *un área*, la encerrada por f_X entre los límites de integración.

2.2.6. Algunos ejemplos de variables aleatorias continuas

Variable aleatoria Uniforme

La variable X diremos que tiene una *distribución uniforme en el intervalo* [a,b], $X \sim U(a, b)$, si su fdp es de la forma

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & si\ x \in [a, b] \\ \\ 0, & en\ el\ resto. \end{cases}$$

La función de distribución que obtendremos integrando f_X vale

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

Surge esta variable cuando elegimos al azar un punto en el intervalo [a,b] y describimos con X su abscisa.

Variable aleatoria Normal

Diremos que X es una *variable aleatoria Normal* de parámetros μ y σ^2, $X \sim N(\mu, \sigma^2)$, si tiene por densidad,

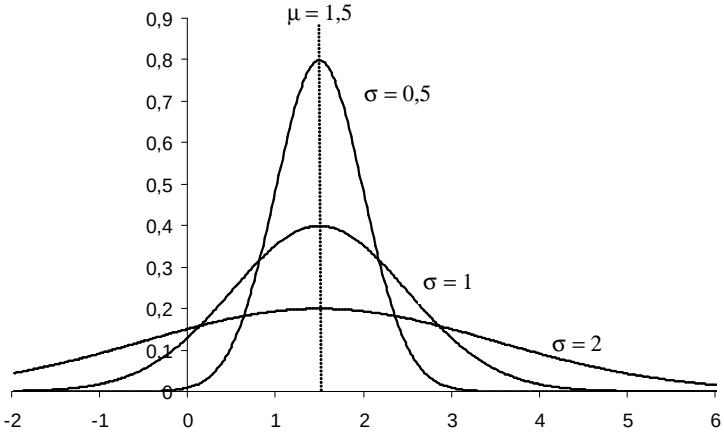
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty. \tag{2.7}$$

En tanto que densidad, f_X debe satisfacer las propiedades Pfdp1) y Pfdp2). La primera se deriva de inmediato de (2.7). Para comprobar la segunda,

$$\begin{aligned}
I^2 &= \left[\int_{-\infty}^{+\infty} f_X(x)dx\right]^2 \\
&= \int_{-\infty}^{+\infty} f_X(x)dx \cdot \int_{-\infty}^{+\infty} f_X(y)dy \\
&= \frac{1}{2\pi}\cdot\frac{1}{\sigma}\int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \cdot \frac{1}{\sigma}\int_{-\infty}^{+\infty} e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy \qquad (2.8)\\
&= \frac{1}{2\pi}\cdot\frac{1}{\sigma}\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\sigma dz \cdot \frac{1}{\sigma}\int_{-\infty}^{+\infty} e^{-\frac{v^2}{2}}\sigma dv \qquad (2.9)\\
&= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{z^2+v^2}{2}}dzdv \qquad (2.10)\\
&= \frac{1}{2\pi}\int_0^{2\pi}\left[\int_0^{+\infty} e^{-\frac{r^2}{2}}rdr\right]d\theta = 1, \qquad (2.11)
\end{aligned}$$

donde el paso de (2.8) a (2.9) se lleva a cabo mediante los cambios $z = (x-\mu)/\sigma$ y $v = (v-\mu)/\sigma$, y el de (2.10) a (2.11) mediante el cambio a polares $z = r\sin\theta$ y $v = r\cos\theta$.

La gráfica de f_X tiene forma de campana y es conocida como la campana de Gauss por ser Gauss quien la introdujo cuando estudiaba los errores en el cálculo de las órbitas de los planetas. En honor suyo, la distribución Normal es también conocida como distribución Gaussiana. Del significado de los parámetros nos ocuparemos más adelante, pero de (2.7) deducimos que $\mu \in \mathcal{R}$ y $\sigma > 0$. Además, el eje de simetría de f_X es la recta $x = \mu$ y el vértice de la campana (máximo de f_x) está en el punto de coordenadas $(\mu, 1/\sigma\sqrt{2\pi})$.



La figura ilustra el papel que los parámetros juegan en la forma y posición de la gráfica de la función de densidad de una $N(\mu, \sigma^2)$. A medida que σ disminuye se produce un mayor apun-

tamiento en la campana porque el máximo aumenta y porque, recordemos, el área encerrada bajo la curva es siempre la unidad.

Una característica de la densidad de la Normal es que carece de primitiva, por lo que su función de distribución no tiene expresión explícita y sus valores están tabulados o se calculan por integración numérica. Esto representa un serio inconveniente si recordamos que $P(a < X \leq b) = F_X(b) - F_X(a)$, puesto que nos obliga a disponer de una tabla distinta para cada par de valores de los parámetros μ y σ.

En realidad ello no es necesario, además de que sería imposible dada la variabilidad de ambos parámetros, porque podemos recurrir a la que se conoce como variable aleatoria *Normal tipificada*, $Z \sim N(0,1)$, cuya densidad es

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < +\infty.$$

En efecto, si para $X \sim N(\mu, \sigma^2)$ queremos calcular

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

efectuando el cambio $z = (t-\mu)/\sigma$ tendremos

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

donde $\Phi(z)$ es la función de distribución de la $N(0,1)$.

Hay que señalar que el mayor interés de la distribución Normal estriba en el hecho de servir de modelo probabilístico para una gran parte de los fenómenos aleatorios que surgen en el campo de las ciencias experimentales y naturales.

El lema que sigue nos asegura que cualquier transformación lineal de un variable Normal es otra Normal.

Lema 2.2 *Sea $X \sim N(\mu, \sigma^2)$ y definamos $Y = aX + b$, entonces $Y \sim N(a\mu + b, a^2\sigma^2)$.*

Demostración.- Supongamos $a > 0$, la función de distribución de Y viene dada por

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

Su función de densidad es

$$f_Y(y) = F_Y'(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-(a\mu+b)}{a\sigma}\right)^2\right\}.$$

Si $a < 0$ entonces

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right),$$

y la densidad será

$$f_Y(y) = F_Y'(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{|a|\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-(a\mu+b)}{a\sigma}\right)^2\right\}.$$

En ambos casos se deduce que $Y \sim N(a\mu + b, a^2\sigma^2)$.

Variable aleatoria Exponencial

Diremos que la variable aleatoria X tiene una *distribución Exponencial de parámetro* λ, $X \sim Exp(\lambda)$, si su función de densidad es de la forma

$$f_X(x) = \begin{cases} 0, & si\ x \leq 0 \\ \lambda e^{-\lambda x}, & si\ x > 0,\ \lambda > 0. \end{cases}$$

La función de distribución de X vendrá dada por

$$F_X(x) = \begin{cases} 0, & si\ x \leq 0 \\ \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}, & si\ x > 0. \end{cases}$$

La distribución exponencial surge en problemas relacionados con tiempos de espera y está relacionada con la distribución de Poisson de igual parámetro. En efecto, si consideramos un proceso de desintegración radiactiva con λ desintegraciones por unidad de tiempo, el número de desintegraciones que se producen en el intervalo [0,t] es $N_t \sim Po(\lambda t)$, y el tiempo que transcurre ente dos desintegraciones consecutivas es $X \sim Exp(\lambda)$.

La falta de memoria de la variable aleatoria Exponencial.- La variable aleatoria Exponencial tiene una curiosa e interesante propiedad conocida como *falta de memoria*. Consiste en la siguiente igualdad,

$$P(X > x + t | X > t) = P(X > x), \quad \forall x, t \geq 0.$$

En efecto,

$$P(X > x + t | X > t) = \\ = \frac{P\left(\{X > x + t\} \cap \{X > t\}\right)}{P(X > t)} = \frac{P(X > x + t)}{P(X > t)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = P(X > x).$$

Variable aleatoria Gamma

Diremos que la variable aleatoria X tiene una *distribución Gamma de parámetros* α *y* β, $X \sim Ga(\alpha, \beta)$, si su función de densidad es de la forma

$$f_X(x) = \begin{cases} 0, & si\ x \leq 0 \\ \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & si\ x > 0,\ \alpha > 0,\ \beta > 0, \end{cases}$$

donde $\Gamma(\alpha)$ es el valor de la función Gamma en α, es decir

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy, \quad \alpha > 0.$$

Para comprobar que Pfdp2) se satisface, la Pfdp1) es de comprobación inmediata, bastará hacer el cambio $y = x/\beta$ en la correspondiente integral

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty y^{\alpha-1} e^{-y} \beta^\alpha dy = \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) = 1.$$

Los valores de la función de distribución $F_X(x)$ aparecen tabulados, con tablas para las diferentes combinaciones de los parámetros α y β.

Obsérvese que la distribución Exponencial de parámetro λ es un caso particular de la Gamma. En concreto $Exp(\lambda) = Gamma(1, 1/\lambda)$.

Observación 2.2 *Nos será de utilidad más tarde recordar alguna característica adicional de la función Gamma. En particular la obtención de sus valores cuando* $\alpha = n$ *o* $\alpha = n + \frac{1}{2}$, n *natural. Es fácil comprobar, mediante sucesivas integraciones por partes, que*

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) = (\alpha - 1)(\alpha - 2)\Gamma(\alpha - 2),$$

lo que para $\alpha = n$ *da lugar a*

$$\Gamma(n) = (n - 1)(n - 2)\ldots 2\Gamma(1).$$

Pero

$$\Gamma(1) = \int_0^\infty e^{-x}dx = 1 \quad y \quad \Gamma(n) = (n - 1)!.$$

Para el caso en que $\alpha = n + \frac{1}{2}$ *deberemos calcular* $\Gamma(\frac{1}{2})$,

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-x}x^{-1/2}dx = \left[y = \frac{t^2}{2}\right] = \sqrt{2}\int_0^\infty e^{-t^2/2}dt = \frac{\sqrt{2}\sqrt{2\pi}}{2} = \sqrt{\pi}. \tag{2.12}$$

La última integral en (2.12), dividida por $\sqrt{2\pi}$, *es la mitad del área que cubre la* fdp *de la* $N(0, 1)$.

Variable aleatoria Ji-cuadrado (χ^2)

Una *variable aleatoria Ji-cuadrado de parámetro* r, $X \sim \chi^2_r$, es una variable aleatoria Gamma con $\alpha = r/2$ y $\beta = 2$ (r entero no negativo). Su función de densidad tiene la expresión

$$f_X(x) = \begin{cases} 0, & si\ x \leq 0 \\ \dfrac{1}{\Gamma(r/2)2^{r/2}}x^{r/2-1}e^{-x/2}, & si\ x > 0,\ r \geq 0. \end{cases}$$

El parámetro r es también conocido como el *número de grados de libertad* de la distribución.

Variable aleatoria Beta

Diremos que la variable aleatoria X tiene una *distribución Beta de parámetros* α *y* β, $X \sim Be(\alpha, \beta)$, si su función de densidad es de la forma

$$f_X(x) = \begin{cases} \dfrac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1 - x)^{\beta-1}, & si\ 0 < x < 1,\ \alpha > 0,\ \beta > 0, \\ 0, & en\ el\ resto. \end{cases}$$

Para comprobar que Pfdp2) se satisface observemos que

$$\Gamma(\alpha)\Gamma(\beta) = \left(\int_0^\infty x^{\alpha-1}e^{-x}dx\right)\left(\int_0^\infty y^{\beta-1}e^{-y}dy\right) = \int_0^\infty\int_0^\infty x^{\alpha-1}y^{\beta-1}e^{-(x+y)}dxdy.$$

Haciendo el cambio $u = x/(x+y)$ tendremos

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty \int_0^1 \frac{u^{\alpha-1}}{(1-u)^{\alpha-1}} y^{\alpha-1} y^{\beta-1} e^{-y/(1-u)} du dy \\
&= \\
&= \int_0^\infty \int_0^1 \frac{u^{\alpha-1}}{(1-u)^{\alpha-1}} y^{\alpha+\beta-1} e^{-y/(1-u)} du dy \qquad (2.13) \\
&= \\
&= \int_0^\infty \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} v^{\alpha+\beta-1} e^{-v} du dv \qquad (2.14) \\
&= \\
&= \int_0^\infty v^{\alpha+\beta-1} e^{-v} dv \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du \\
&= \\
&= \Gamma(\alpha+\beta) \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du,
\end{aligned}$$

donde el paso de (2.13) a (2.14) se lleva a cabo mediante el cambio $v = y/(1-u)$.

En definitiva,

$$\int_0^1 f_X(x)dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)} = 1.$$

Como en la caso de la distribución Gamma, también ahora los valores de la función de distribución $F_X(x)$ aparecen tabulados para las diferentes combinaciones de los parámetros α y β.

Observación 2.3 *La función de densidad de* $Be(1,1)$*, teniendo en cuenta que* $\Gamma(1) = \Gamma(2) = 1$*, tiene por expresión,*

$$f_X(x) = \begin{cases} 1, & si\ 0 < x < 1, \\ \\ 0, & \text{en el resto}, \end{cases}$$

que corresponde a la densidad de una variable aleatoria Uniforme en [0,1].

2.3. Vector aleatorio

Cuando estudiamos simultáneamente k características numéricas ligadas al resultado del experimento, por ejemplo la *altura* y el *peso* de las personas, nos movemos en $\mathcal{R}^k$ como espacio imágen de nuestra aplicación. La σ-álgebra de sucesos con la que dotamos a $\mathcal{R}^k$ para hacerlo probabilizable es la correspondiente σ-álgebra de Borel β^k, que tiene la propiedad de ser la menor que contiene a los *rectángulos* $(a,b] = \prod_{i=1}^k (a_i, b_i]$, con $a = (a_1, \ldots, a_k)$, $b = (b_1, \ldots, b_k)$ y $-\infty \leq a_i \leq b_i < +\infty$. De entre ellos merecen especial mención aquellos que tiene el extremo inferior en $-\infty$, $S_x = \prod_{i=1}^k (-\infty, x_i]$, a los que denominaremos *región suroeste de* x, por que sus puntos están situados al suroeste de x.

Definición 2.4 (Vector aleatorio) *Un* vector aleatorio*,* $X = (X_1, X_2, \ldots, X_k)$*, es una aplicación de* Ω *en* $\mathcal{R}^k$*, que verifica*

$$X^{-1}(B) \in \mathcal{A}, \quad \forall B \in \beta^k.$$

La presencia de una probabilidad sobre el espacio $(\Omega, \mathcal{A})$ permite al vector inducir una probabilidad sobre $(\mathcal{R}^k, \beta^k)$.

2.3.1. Probabilidad inducida

X induce sobre $(\mathcal{R}^k, \beta^k)$ una probabilidad, P_X, de la siguiente forma,

$$P_X(B) = P(X^{-1}(B)),\ \forall B \in \beta^k.$$

Es sencillo comprobar que verifica los tres axiomas que definen una probabilidad, por lo que la terna $(\mathcal{R}^k, \beta^k, P_X)$ constituye un espacio de probabilidad con las características de $(\Omega, \mathcal{A}, P)$ heredadas a través de X.

2.3.2. Funciones de distribución conjunta y marginales

La función de distribución asociada a P_X se define para cada punto $x = (x_1, \ldots, x_k)$ de $\mathcal{R}^k$ mediante

$$F_X(x) = F_X(x_1, \ldots, x_k) = P_X(S_x) = P(X \in S_x) = P\left(\bigcap_{i=1}^{k}\{X_i \leq x_i\}\right). \tag{2.15}$$

De la definición se derivan las siguientes propiedades:

PFC1) No negatividad.- Consecuencia inmediata de ser una probabilidad.

PFC2) Monotonía en cada componente.- Si $x \leq y$, es decir, $x_i \leq y_i,\ i = 1, \ldots, k$, $S_x \subset S_y$ y $F_X(x) \leq F_X(y)$.

PFC3) Continuidad conjunta por la derecha.- Si $x^{(n)} \downarrow x$, entonces $F_X(x^{(n)}) \downarrow F_X(x)$,

PFC4) Valores límites.- Al tender a $\pm\infty$ las componentes del punto, se tiene

$$\lim_{\forall x_i \uparrow +\infty} F_X(x_1, \ldots, x_k) = 1,$$

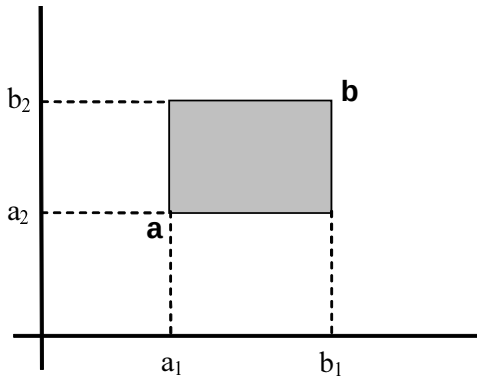
o bien,

$$\lim_{\exists x_i \downarrow -\infty} F(x_1, \ldots, x_k) = 0.$$

que no son más que la versión multidimensional de las propiedades ya conocidas para el caso unidimensional. Existe ahora una quinta propiedad sin la cual no sería posible recorrer el camino inverso, obtener P_X a partir de F_X, y establecer la deseada y conveniente equivalencia entre ambos conceptos.

PFC5) Supongamos que $k = 2$ y consideremos el rectángulo $(a, b] = (a_1, b_1] \times (a_2, b_2]$ tal como lo muestra la figura. Indudablemente $P_X(]a, b]) \geq 0$, pero teniendo en cuenta (2.15) podemos escribir,

$$P_X(]a, b]) = F_X(b_1, b_2) - F_X(b_1, a_2) - F_X(a_1, b_2) + F_X(a_1, a_2) \geq 0.$$



Este resultado es cierto, obviamente, para cualquier k y necesitamos introducir el operador diferencia a fin de obtener una representación sencilla del resultado en el caso general. Para $a_i \leq b_i,\ \Delta_{a_i,b_i}$ representa el operador diferencia que actua de la forma

$$\begin{aligned}\Delta_{a_i,b_i} F_X(x_1,\ldots,x_k) &= \\ &= F_X(x_1,\ldots,x_{i-1},b_i,\ldots,x_k) - F_X(x_1,\ldots,x_{i-1},a_i,\ldots,x_k).\end{aligned}$$

Sucesivas aplicaciones del mismo conducen a

$$\Delta_{a,b} F_X(x) = \Delta_{a_1,b_1},\ldots,\Delta_{a_k,b_k} F_X(x_1,\ldots,x_k) = P_X((a,b]), \tag{2.16}$$

lo que permite generalizar el anterior resultado mediante la expresión,

$$\Delta_{a,b} F_X(x) \geq 0, \quad \forall a,b \in \mathcal{R}^k, \text{ con } a_i \leq b_i.$$

La equivalencia entre P_X y F_X se establece a través de (2.16). En efecto, es posible recuperar P_X a partir de

$$P_X((a,b]) = \Delta_{a,b} F_X(x),$$

lo que nos autoriza a escribir $P_X \Longleftrightarrow F_X$. El comentario final del párrafo 2.2.2 es también aplicable ahora.

Funciones de distribución marginales

Si el vector es aleatorio cabe pensar que sus componentes también lo serán. La siguiente proposición establece una primera relación entre el vector y sus componentes.

Proposición 2.1 *$X = (X_1,\ldots,X_k)$ es un vector aleatorio* sí y solo sí *cada una de sus componentes es una variable aleatoria.*

Demostración.- Recordemos que la condición de variable aleatoria le viene a una aplicación por el hecho de transformar las antimágenes de conjuntos de Borel en sucesos:

$$X^{-1}(B) \in \mathcal{A},\ B \in \beta. \tag{2.17}$$

Existe un resultado que permite caracterizar esta propiedad utilizando una familia menor de conjuntos, evitándonos así la prueba para cualquier elemento de β. Basta, según dicha caracterización, comprobar la propiedad en una familia que engendre a la σ-álgebra. Pues bien, los intervalos de la forma $]a,b]$, en $\mathcal{R}$, y los conjuntos suroeste, S_x, en $\mathcal{R}^k$, engendran β y β^k, respectivamente. Ahora ya podemos abordar la demostración.

Supongamos que las componentes de $X = (X_1, X_2, \ldots, X_k)$ son variables aleatorias. Para $S_x = \prod_{i=1}^{k}]-\infty, x_i]$ podemos escribir

$$\{X \in S_x\} = \bigcap_{i=1}^{k} \{X_i \leq x_i\}. \tag{2.18}$$

Si cada componente es aleatoria, $\{X_i \leq x_i\} \in \mathcal{A},\ \forall i$, y X será un vector aleatorio.

Para demostrar el inverso observemos que

$$\{X_j \leq x_j\} = \lim_{x_i \uparrow +\infty,\ i \neq j} \{X \leq S_x\},$$

donde para conseguir la numerabilidad los x_i han de tender a $+\infty$ a través de una sucesión, lo que puede conseguirse haciendo $x_i = n,\ i \neq j$. En consecuencia, X_j es medible, pues al tratarse

de una sucesión monótona creciente de conjuntos, su límite es la unión de todos ellos que es una operación estable en $\mathcal{A}$. ♠

Si las componentes del vector son variables aleatorias tendrán asociadas sus correspondientes probabilidades inducidas y funciones de distribución. La nomenclatura hasta ahora utilizada necesita ser adaptada, lo que haremos añadiendo los adjetivos *conjunta* y *marginal*, respectivamente. Puesto que P_X y F_X describen el comportamiento conjunto de las componentes de X, nos referiremos a ellas como *distribución conjunta* y *función de distribución conjunta* del vector X, respectivamente. Cuando, en el mismo contexto, necesitemos referirnos a la distribución de alguna componente lo haremos aludiendo a la *distribución marginal* o a la *función de distribución marginal* de X_i.

La pregunta que surge de inmediato es, *¿qué relación existe entre la distribución conjunta y las marginales?* Estamos en condiciones de dar respuesta en una dirección: cómo obtener la distribución marginal de cada componente a partir de la conjunta. Para ello, basta tener en cuenta que

$$\lim_{x_j \overset{j\neq i}{\to} \infty} \bigcap_{j=1}^{k} \{X_j \leq x_j\} = \{X_i \leq x_i\},$$

y al tomar probabilidades obtendremos

$$F_{X_i}(x_i) = \lim_{x_j \overset{j\neq i}{\to} \infty} F_X(x_1, \ldots, x_k). \tag{2.19}$$

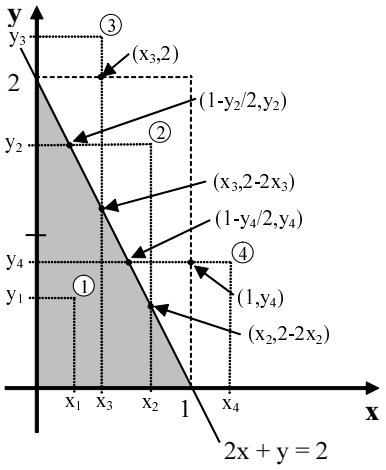
El concepto de *marginalidad* puede aplicarse a cualquier subvector del vector original. Así, para $l \leq k$, si $X^l = (X_{i_1}, \ldots, X_{i_l})$ es un subvector de X, podemos hablar de la *distribución conjunta marginal de* X^l, para cuya obtención a partir de la conjunta procederemos de forma análoga a como acabamos de hacer para una componente. Si en la relación (2.18) fijamos $x_{i_1}, \ldots, x_{i_l}$ y hacemos tender a ∞ el resto de las componentes, obtendremos

$$\{X^l \in S_{x^l}\} = \lim_{x_i \overset{i\neq i_1,\ldots,i_l}{\longrightarrow} \infty} \{X \in S_x\}.$$

Relación que nos permite obtener la función de distribución marginal conjunta de $X^l = (X_{i_1}, \ldots, X_{i_l})$ sin más que tomar probabilidades,

$$F_{X^l}(x_{i_1}, \ldots, x_{i_l}) = \lim_{x_i \overset{i\neq i_1,\ldots,i_l}{\longrightarrow} \infty} F_X(x_1, \ldots, x_k).$$

Ejemplo 2.2 *Elegimos un punto al azar sobre el triángulo T de vértices* (0,0), (1,0), (0,2).



Aplicando (2.20) obtenemos

Para encontrar la función de distribución conjunta del vector de sus componentes, (X,Y), observemos la figura y las distintas posiciones del punto. Como la masa de probabilidad está uniformemente repartida sobre el triángulo puesto que la elección del punto es al azar, tendremos que

$$P((X,Y)\in A)=\frac{\|A\cap T\|}{\|T\|},$$

donde $\|B\|$ es el área de B. Aplicado a la función de distribución dará lugar a

$$F_{XY}(x,y)=P((X,Y)\in S_{xy})=\|S_{xy}\cap T\|, \quad (2.20)$$

puesto que el área del triángulo vale 1.

$$F_{XY}(x,y)=\begin{cases} 0, & si\ x\le 0\ o\ y\le 0;\\ xy, & si\ (x,y)\ es\ del\ tipo\ \textbf{\textit{1}}\ ;\\ xy-(x+y/2-1)^2, & si\ (x,y)\ es\ del\ tipo\ \textbf{\textit{2}}\ ;\\ 2x-x^2, & si\ (x,y)\ es\ del\ tipo\ \textbf{\textit{3}}\ ;\\ y-y^2/4, & si\ (x,y)\ es\ del\ tipo\ \textbf{\textit{4}}\ ;\\ 1, & si\ x\ge 1\ e\ y\ge 2; \end{cases}$$

*Observemos que las expresiones de $F_{XY}(x,y)$ correspondientes a puntos del tipo **3** y **4** dependen solamente de x e y, respectivamente. Si recordamos la obtención de la función de distribución marginal veremos que se corresponden con $F_X(x)$ y $F_Y(y)$, respectivamente.*

2.3.3. Vector aleatorio discreto. Función de cuantía conjunta

Si el soporte, D_X, del vector es numerable, lo que supone que también lo son los de cada una de sus componentes, diremos que X es un *vector aleatorio discreto*. Como en el caso unidimensional, una tercera función puede asociarse al vector y nos permite conocer su comportamiento aleatorio. Se trata de la *función de cuantía o probabilidad conjunta* y su valor, en cada punto de $\mathcal{R}^k$, viene dado por

$$f_X(x_1,\ldots,x_k)=\begin{cases} P(X_i=x_i,\ i=1,\ldots,k), & si\ x=(x_1,\ldots,x_k)\in D_X\\ 0, & en\ el\ resto. \end{cases}$$

La función de cuantía conjunta posee las siguientes propiedades:

Pfcc1) Al tratarse de una probabilidad, $f_X(x)\ge 0,\ \forall x\in\mathcal{R}^k$,

Pfcc2) Como $P(X\in D_X)=1$,

$$\sum_{x_1}\cdots\sum_{x_k} f_X(x_1,x_2,\ldots,x_k)=1,\ (x_1,x_2,\ldots,x_k)\in D_X.$$

Entre F_X y f_X se establecen relaciones similares a las del caso unidimensional:

$$F_X(x)=\sum_{y\le x,\ y\in D_X} f_X(y_1,y_2,\ldots,y_k),$$

y

$$f_X(x_1, x_2, \ldots, x_k) = F_X(x_1, x_2, \ldots, x_k) - F_X(x_1-, x_2-, \ldots, x_k-).$$

De ambas expresiones se deduce la equivalencia entre ambas funciones y también la de éstas con P_X,

$$P_X \Longleftrightarrow F_X \Longleftrightarrow f_X.$$

Funciones de cuantía marginales

Si el vector aleatorio X es discreto también lo serán cada una de sus componentes. Si por D_i designamos el soporte de $X_i,\ i = 1, \ldots, k$, se verifica,

$$\{X_i = x_i\} = \bigcup_{x_j \in D_j,\ j \neq i} \left(\bigcap_{j=1}^{k} \{X_j = x_j\} \right),$$

siendo disjuntos los elementos que intervienen en la unión. Al tomar probabilidades tendremos

$$f_{X_i}(x_i) = \sum_{x_j \in D_j,\ j \neq i} f_X(x_1, \ldots, x_k),$$

que permite obtener la función de cuantía marginal de la X_i a partir de la conjunta. La marginal conjunta de cualquier subvector aleatorio se obtendría de manera análoga, extendiendo la suma sobre todas las componentes del subvector complementario,

$$f_{X^l}(x_{i_1}, \ldots, x_{i_l}) = \sum_{x_j \in D_j,\ j \neq i_1, \ldots, i_l} f_X(x_1, \ldots, x_k).$$

Ejemplo 2.3 *Supongamos un experimento consistente en lanzar 4 veces una moneda correcta. Sea X el numero de caras en los 3 primeros lanzamientos y sea Y el número de cruces en los 3 últimos lanzamientos. Se trata de un vector discreto puesto que cada componente lo es. En concreto $D_X = \{0, 1, 2, 3\}$ y $D_Y = \{0, 1, 2, 3\}$.*

La función de cuantía conjunta se recoge en la tabla siguiente, para cualquier otro valor no recogido en la tabla $f_{XY}(x, y) = 0$.

		X			
Y	**0**	**1**	**2**	**3**	$\boldsymbol{f_Y(y)}$
0	*0*	*0*	*1/16*	*1/16*	*2/16*
1	*0*	*2/16*	*3/16*	*1/16*	*6/16*
2	*1/16*	*3/16*	*2/16*	*0*	*6/16*
3	*1/16*	*1/16*	*0*	*0*	*2/16*
$\boldsymbol{f_X(x)}$	*2/16*	*6/16*	*6/16*	*2/16*	*1*

En los márgenes de la tabla parecen las funciones de cuantía marginales de X e Y, obtenidas al sumar a lo largo de la correspondiente fila o columna.

2.3.4. Algunos ejemplos de vectores aleatorios discretos

Vector aleatorio Multinomial

La versión k-dimensional de la distribución Binomial es el llamado vector aleatorio *Multinomial*. Surge este vector en el contexto de un experimento aleatorio en el que nos interesamos en la ocurrencia de alguno de los k sucesos, $A_1, A_2, \ldots, A_k$, que constituyen una partición de

Ω. Si $P(A_i) = p_i$ y repetimos n veces el experimento de manera que las repeticiones son independientes, el vector $X = (X_1, \ldots, X_k)$, con X_i =*número de ocurrencias de* A_i, decimos que tiene una *distribución Multinomial*, $X \sim M(n; p_1, p_2, \ldots, p_k)$. La función de cuantía conjunta viene dada por,

$$f_X(n_1, \ldots, n_k) = \begin{cases} \dfrac{n!}{n_1! n_2! \ldots n_k!} \displaystyle\prod_{i=1}^{k} p_i^{n_i}, & si\ 0 \leq n_i \leq n,\ i = 1, \ldots, k,\ \sum n_i = n \\ \\ 0, & en\ el\ resto, \end{cases}$$

que verifica Pfcc1) y Pfcc2), porque es no negativa y al sumarla para todos los posibles $n_1, n_2, \ldots, n_k$ obtenemos el desarrollo del polinomio $(p_1 + p_2 + \ldots + p_k)^n$, de suma 1 porque los A_i constituían una partición de Ω.

Para obtener la marginal de X_i observemos que

$$\frac{n!}{n_1! n_2! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i} = \binom{n}{n_i} p_i^{n_i} \frac{(n - n_i)!}{n_1! \ldots n_{i-1}! n_{i+1}! \ldots n_k!} \prod_{j \neq i} p_j^{n_j},$$

y al sumar para el resto de componentes,

$$\begin{aligned} f_{X_i}(n_i) &= \binom{n}{n_i} p_i^{n_i} \sum \frac{(n - n_i)!}{n_1! \ldots n_{i-1}! n_{i+1}! \ldots n_k!} \prod_{j \neq i} p_j^{n_j}, \\ &= \binom{n}{n_i} p_i^{n_i} (p_1 + \ldots + p_{i-1} + p_{i+1} + \ldots + p_k)^{n - n_i} \\ &= \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i}, \end{aligned}$$

llegamos a la conclusión que $X_i \sim B(n, p_i)$, como era de esperar, pues al fijar X_i sólo nos interesamos por la ocurrencia de A_i y el experimento que llevamos a cabo puede ser descrito mediante un modelo Binomial.

Vector aleatorio Binomial Negativo bivariante

Supongamos que A_1, A_2 y A_3 constituyen una partición del espacio muestral Ω, de manera que $P(A_i) = p_i$, $i = 1, 2, 3$ con $p_1 + p_2 + p_3 = 1$. Realizamos un experimento cuyo resultado es, necesariamente, A_1, A_2 o A_3. Repetimos el experimento de manera independiente en cada ocasión hasta que el suceso A_3 haya ocurrido r veces. Ello significa que A_1 habrá ocurrido x veces y A_2 lo habrá hecho en y ocasiones, todas ellas previas a la r-ésima ocurrencia de A_3. La probabilidad de un suceso de estas características vendrá dada por

$$\binom{x + y + r - 1}{x} \binom{y + r - 1}{y} p_1^x p_2^y (1 - p_1 - p_2)^r = \frac{(x + y + r - 1)!}{x! y! (r - 1)!} p_1^x p_2^y (1 - p_1 - p_2)^r.$$

Podemos ahora definir un vector aleatorio *Binomial Negativo bivariante* $(X, Y) \sim BN(r; p_1, p_2)$ como aquél que tiene por función de cuantía,

$$f_{XY}(x, y) = \begin{cases} \dfrac{(x + y + r - 1)!}{x! y! (r - 1)!} p_1^x p_2^y (1 - p_1 - p_2)^r, & (x, y) \in D_{xy} = [0, 1, 2, \ldots]^2, \\ \\ 0, & en\ el\ resto. \end{cases}$$

Se trata de una función de cuantía por cuanto $f_{XY}(x,y) \geq 0, \forall (x,y) \in \mathcal{R}^2$ y, teniendo en cuenta que

$$(1-p_1+p_2)^{-r} = \sum_{D_{xy}} \binom{x+y+r-1}{x} \binom{y+r-1}{y} p_1^x p_2^y,$$

también verifica

$$\sum_{D_{xy}} f_{XY}(x,y) = (1-p_1+p_2)^r (1-p_1+p_2)^{-r} = 1.$$

La marginal de cualquiera de las componentes, por ejemplo Y, la obtendremos a partir de $f_Y(y) = \sum_{D_x} f_{XY}(x,y)$, con $D_x = \{0,1,2,\ldots\}$,

$$\begin{aligned} f_Y(y) &= \sum_{D_x} \binom{x+y+r-1}{x} \binom{y+r-1}{y} p_1^x p_2^y (1-p_1-p_2)^r \\ &= \binom{y+r-1}{y} p_2^y (1-p_1-p_2)^r \sum_{D_x} \binom{x+y+r-1}{x} p_1^x, \end{aligned}$$

pero recordemos que $(1-x)^{-n} = \sum_{j\geq 0} \binom{n+j-1}{j} x^j$, por tanto,

$$\begin{aligned} f_Y(y) &= \binom{y+r-1}{y} p_2^y (1-p_1-p_2)^r (1-p_1)^{-(y+r)} \\ &= \binom{y+r-1}{y} \left(\frac{p_2}{1-p_1}\right)^y \left(\frac{1-p_1-p_2}{1-p_1}\right)^r \\ &= \binom{y+r-1}{y} p^y (1-p)^r. \end{aligned}$$

Luego $Y \sim BN(r,p)$ con $p = p_2/(1-p_1)$.

2.3.5. Vector aleatorio continuo. Función de densidad de probabilidad conjunta

Si la probabilidad inducida por el vector aleatorio es absolutamente continua respecto de λ_k, medida de Lebesgue en $\mathcal{R}^k$, decimos que X es un *vector aleatorio continuo*. Existe entonces un función, f_X sobre $\mathcal{R}^k$, conocida como *función de densidad de probabilidad conjunta* de X, tal que $\forall B \in \beta^k$

$$P(X \in B) = \int_B f_x \, d\lambda_k.$$

Si, como ocurre habitualmente, f_X es integrable Riemann en $\mathcal{R}^k$, podremos escribir

$$P(X \in (a,b]) = P(a_i < X_i \leq b_i,\ i = 1,\ldots,k) = \int_{a_k}^{b_k} \ldots \int_{a_1}^{b_1} f_X(x_1,\ldots,x_k)\, dx_1 \ldots dx_k.$$

Al igual que ocurría en el caso unidimensional, esta función tiene dos propiedades que la caracterizan,

Pfdpc1) $f_X(x)$ es no negativa, y

Pfdcp2) como $P(X \in \mathcal{R}^k) = 1$,

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f_X(x_1, \dots, x_k) dx_1 \dots dx_k = 1.$$

Como consecuencia de esta definición, entre la función de distribución conjunta y la de densidad de probabilidad conjunta se establecen las siguientes relaciones:

$$F_X(x_1, \dots, x_k) = P_X(S_x) = \int_{-\infty}^{x_k} \dots \int_{-\infty}^{x_1} f(t_1, \dots, t_k)\, dt_1 \dots dt_k, \tag{2.21}$$

y si $x \in \mathcal{R}^k$ es un punto de continuidad de f_X,

$$f_X(x_1, \dots, x_k) = \frac{\partial^k F_X(x_1, \dots, x_k)}{\partial x_1, \dots, \partial x_k}. \tag{2.22}$$

Aunque esta última relación ha ser matizada en los mismos términos en los que lo fue su versión unidimensional, a saber, la igualdad es cierta excepto en conjuntos de medida λ_k nula. En cualquier caso, las anteriores relaciones expresan la equivalencia de ambas funciones y podemos escribir,

$$P_X \Longleftrightarrow F_X \Longleftrightarrow f_X.$$

Funciones de densidad marginales

Para obtener la densidad marginal de X_i a partir de la función de la densidad conjunta tengamos en cuenta (2.19) y (2.21) y que integración y paso al límite pueden permutarse por ser la densidad conjunta integrable,

$$\begin{aligned} F_{X_i}(x_i) &= \lim_{x_j \overset{j \neq i}{\to} \infty} F_X(x_1, \dots, x_k) \\ &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{x_i} \dots \int_{-\infty}^{+\infty} f(t_1, \dots, t_i, \dots, t_k)\, dt_1 \dots dt_i \dots dt_k. \end{aligned}$$

Pero la derivada de F_{X_i} es una de las densidades de X_i y como las condiciones de la densidad conjunta permiten también intercambiar derivación e integración, tendremos finalmente

$$\begin{aligned} &f_{X_i}(x_i) = \\ &\quad = \int_{\mathcal{R}^{k-1}} f(t_1, \dots, x_i, \dots, t_k)\, dt_1 \dots dt_{i-1} dt_{i+1} \dots dt_k. \end{aligned} \tag{2.23}$$

Para el caso de un subvector, X^l, la densidad conjunta se obtiene de forma análoga,

$$\begin{aligned} &f_{X^l}(x_{i_1}, \dots, x_{i_l}) = \\ &\quad = \int_{\mathcal{R}^{k-l}} f_X(t_1, \dots, t_k) \prod_{j \neq i_1, \dots, i_l} dt_j. \end{aligned}$$

Ejemplo 2.4 *La función de densidad conjunta del vector aleatorio bidimensional* (X, Y) *viene dada por*

$$f_{XY}(x, y) = \begin{cases} 8xy, & si\ 0 \leq y \leq x \leq 1 \\ 0, & en\ el\ resto. \end{cases}$$

Si queremos obtener las marginales de cada componente, tendremos para X

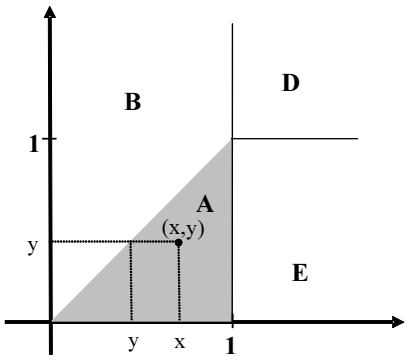
$$f_X(x) = \int_0^x f_{XY}(x,v)dv = \int_0^x 8xvdv = 4x^3, \quad 0 \le x \le 1,$$

y cero en el resto. Para Y*,*

$$f_Y(y) = \int_y^1 f_{XY}(u,y)du = \int_y^1 8uydu = 4y(1-y^2), \quad 0 \le y \le 1,$$

y cero en el resto.

Obtengamos ahora la función de distribución conjunta, $F_{XY}(x,y)$*. Observemos para ello el gráfico, la función de densidad es distinta de 0 en la región* A *por lo que* $F_{XY}(x,y) = 0$ *si* $x \le 0$ *o* $y \le 0$*.*



Si $(x,y) \in A$*,*

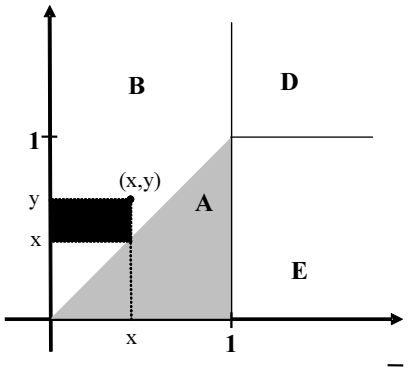
$$F_{XY}(x,y) = \int\int_{S_{xy} \cap A} f_{XY}(u,v)dudv,$$

pero $S_{xy} \cap A = \{(u,v); 0 \le v \le u \le y\} \cup \{(u,v); y \le u \le x,\ 0 \le v \le y\}$*, por tanto*

$$F_{XY}(x,y) = \int_0^y \left[\int_0^u 8uvdv\right] du + \int_y^x \left[\int_0^y 8uvdv\right] du = y^2(2x^2 - y^2). \tag{2.24}$$

Si $(x,y) \in B$*, como* $f_{XY}(x,y) = 0$ *en* B*, el rectángulo superior de la figura (en negro) no acumula probabilidad y por tanto*

$$F_{XY}(x,y) = P\left(S_{xy} \cap A\right) = P\left(S_{xx} \cap A\right).$$



Así pues,

$$F_{XY}(x,y) = \int_0^x \left[\int_0^u 8uvdv \right] du = x^4. \tag{2.25}$$

Observemos que (2.25) puede obtenerse a partir de (2.24) haciendo $y = x$. En efecto, de acuerdo con (2.19), (2.25) no es más que $F_X(x)$, la función de distribución marginal de X.

Si $(x,y) \in E$, un razonamiento similar conduce a

$$F_{XY}(x,y) = y^2(2-y^2),$$

que no es más que la función de distribución marginal de Y, que podríamos haber obtenido haciendo $x = 1$ en (2.24).

Por último, si $(x,y) \in D$, $F_{XY}(x,y) = 1$. Para su obtención basta hacer $x = 1$ e $y = 1$ en (2.24).

Resumiendo

$$F_{XY}(x,y) = \begin{cases} 0, & si\ x \leq 0\ o\ y \leq 0; \\ y^2(2x^2 - y^2), & si\ (x,y) \in A; \\ x^4, & si\ (x,y) \in B; \\ y^2(2-y^2), & si\ (x,y) \in E; \\ 1, & si\ (x,y) \in D. \end{cases}$$

Ejemplo 2.5 *La función de densidad conjunta del vector (X_1, X_2, X_3) es*

$$f(x_1,x_2,x_3) = \begin{cases} \dfrac{48x_1x_2x_3}{(1+x_1^2+x_2^2+x_3^2)^4}, & si\ x_1,x_2,x_3 \geq 0 \\ 0, & en\ el\ resto. \end{cases}$$

Obtengamos en primer lugar las densidades marginales de cada componente. Dada la simetría de la densidad conjunta bastará con obtener una cualquiera de ellas.

$$\begin{aligned} f_1(x_1) &= \int_0^\infty \int_0^\infty \frac{48x_1x_2x_3}{(1+x_1^2+x_2^2+x_3^2)^4} dx_2dx_3 \\ &= 48x_1 \int_0^\infty x_2 \left[\int_0^\infty \frac{x_3}{(1+x_1^2+x_2^2+x_3^2)^4} dx_3 \right] dx_2 \\ &= \int_0^\infty \int_0^\infty \frac{8x_1x_2}{(1+x_1^2+x_2^2)^3} dx_2 \qquad (2.26) \\ &= \frac{2x_1}{(1+x_1^2)^2}. \end{aligned}$$

Luego

$$f_i(x_i) = \begin{cases} \dfrac{2x_i}{(1+x_i^2)^2}, & si\ x_i \geq 0 \\ 0, & en\ el\ resto. \end{cases} \tag{2.27}$$

Por otra parte, en el transcurso de la obtención de $f_i(x_i)$ el integrando de (2.26) es la densidad marginal conjunta de (X_1, X_2), que por la simetría antes mencionada es la misma para cualquier pareja de componentes. Es decir, para $i, j = 1, 2, 3$

$$f_{ij}(x_i,x_j) = \begin{cases} \dfrac{8x_ix_j}{(1+x_i^2+x_j^2)^3}, & si\ x_i,x_j \geq 0 \\ 0, & en\ el\ resto. \end{cases}$$

Para obtener la función de distribución conjunta recordemos que

$$F(x_1,x_2,x_3)=P(X_i\leq x_i, i=1,2,3)=P\left(\bigcap_{i=1}^{3}\{X_i\leq x_i\}\right)=\int_0^{x_1}\int_0^{x_2}\int_0^{x_3} f(u,v,z)dudvdz.$$

pero en este caso será más sencillo recurrir a esta otra expresión,

$$F(x_1,x_2,x_3)=1-P\left(\left[\bigcap_{i=1}^{3}\{X_i\leq x_i\}\right]^c\right)=1-P\left(\bigcup_{i=1}^{3}A_i\right),$$

con $A_i=\{X_i>x_i\}$. Si aplicamos la fórmula de inclusión-exclusión (1.1),

$$F(x_1,x_2,x_3)=1-\left[\sum_{i=1}^{3}P(A_i)-\sum_{1\leq i<j\leq 3}P(A_i\cap A_j)+P(A_1\cap A_2\cap A_3)\right]. \tag{2.28}$$

La obtención de las probabilidades que aparecen en (2.28) involucran a las densidades antes calculadas. Así, para $P(A_i)$

$$P(A_i)=P(X_i>x_i)=\int_{x_i}^{\infty}\frac{2u}{(1+u^2)^2}du=\frac{1}{1+x_i^2}.$$

Para $P(A_i\cap A_j)$,

$$\begin{aligned}P(A_i\cap A_j) &= P(X_i>x_i, X_j>x_j)\\ &= \int_{x_i}^{\infty}\int_{x_j}^{\infty}\frac{8uv}{(1+u^2+v^2)^3}dudv\\ &= \frac{1}{1+x_i^2+x_j^2}.\end{aligned}$$

Finalmente

$$P(A_1\cap A_2\cap A_3)=\int_{x_1}^{\infty}\int_{x_2}^{\infty}\int_{x_3}^{\infty}ds\frac{48uvz}{(1+u^2+v^2+z^2)^4}dudvdz=\frac{1}{1+x_1^2+x_2^2+x_3^2}.$$

Sustituyendo en (2.28) tendremos,

$$F(x_1,x_2,x_3)=1-\sum_{i=1}^{3}\frac{1}{1+x_i^2}+\sum_{1\leq i<j\leq 3}\frac{1}{1+x_i^2+x_j^2}-\frac{1}{1+x_1^2+x_2^2+x_3^2}.$$

2.3.6. Algunos ejemplos de vectores aleatorios continuos

Vector aleatorio Uniforme en el círculo unidad

Al elegir un punto al azar en, C_1, círculo unidad, podemos definir sobre el correspondiente espacio de probabilidad un vector aleatorio de las coordenadas del punto, (X,Y). La elección al azar implica una probabilidad uniforme sobre C_1, lo que se traduce en un densidad conjunta constante sobre todo el círculo, pero como por otra parte $\int\int_{C_1} f(x,y)\,dxdy=1$, la densidad conjunta vendrá dada por

$$f_{XY}(x,y)=\begin{cases}\dfrac{1}{\pi}, & si\ (x,y)\in C_1\\ 0, & en\ el\ resto.\end{cases}$$

Para obtener la densidad marginal de X,

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy = \frac{1}{\pi}\int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} dy = \frac{2}{\pi}\sqrt{1-x^2},$$

por lo que

$$f_X(x) = \begin{cases} \dfrac{2}{\pi}\sqrt{1-x^2}, & si\ |x| \le 1 \\ \\ 0, & en\ el\ resto, \end{cases}$$

La marginal de Y, por simetría, tiene la misma expresión.

Si en lugar de llevar a cabo la elección en C_1, elegimos el punto en el cuadrado unidad, $Q = [0,1] \times [0,1]$, la densidad conjunta es constante e igual a 1 en el cuadrado y las marginales son ambas $U(0,1)$.

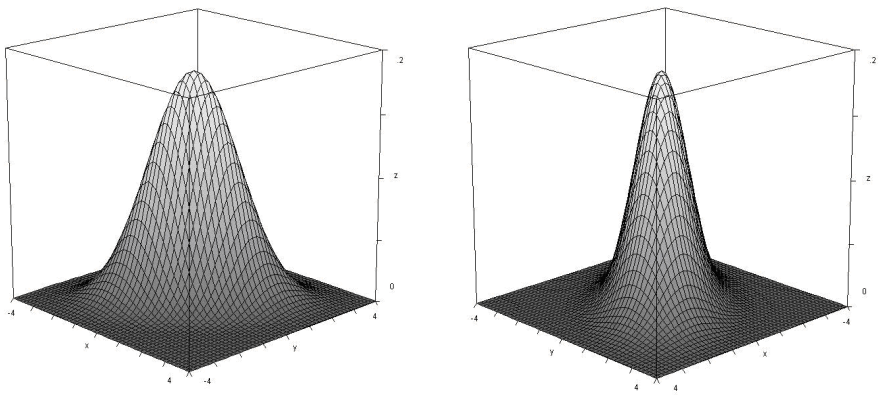
Vector aleatorio Normal bivariante

El vector aleatorio bidimensional (X,Y) tiene una distribución Normal bivariante de parámetros $\mu_X \in \mathcal{R}$, $\mu_y \in \mathcal{R}$, $\sigma_x > 0$, $\sigma_y > 0$ y ρ, $|\rho| < 1$, si su función de densidad conjunta es de la forma,

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{q(x,y)}{2}}, \quad (x,y) \in \mathcal{R}^2,$$

donde

$$q(x,y) = \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}.$$



La figura nos muestras sendas gráficas de la normal bivariante con parámetros $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ y $\rho = 0,5$. La gráfica está centrada en (μ_x, μ_y) (parámetros de posición) y su forma depende de σ_x, σ_y y ρ (parámetros de forma). Para ver el efecto de este último la gráfica de la derecha ha sido rotada $-90°$.

Para ver que $f_{XY}(x,y)$ es una densidad es inmediato comprobar que verifica la primera condición, en cuanto a la segunda, $\int_{\mathcal{R}^2} f_{XY}(x,y)dxdy = 1$, observemos que

$$\begin{aligned}(1-\rho^2)q(x,y) &= \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \\ &= \left[\left(\frac{x-\mu_x}{\sigma_x}\right) - \rho\left(\frac{y-\mu_y}{\sigma_y}\right)\right]^2 + (1-\rho^2)\left(\frac{y-\mu_y}{\sigma_y}\right)^2, \qquad (2.29)\end{aligned}$$

pero el primer sumando de (2.29) puede escribirse

$$\begin{aligned}\left(\frac{x-\mu_x}{\sigma_x}\right) - \rho\left(\frac{y-\mu_y}{\sigma_y}\right) &= \left(\frac{x-\mu_x}{\sigma_x}\right) - \frac{\rho\sigma_x}{\sigma_x}\left(\frac{y-\mu_y}{\sigma_y}\right) \\ &= \frac{1}{\sigma_x}\left[x - \left(\mu_x + \rho\sigma_x\frac{y-\mu_y}{\sigma_y}\right)\right] \\ &= \frac{1}{\sigma_x}(x-b),\end{aligned}$$

con $b = \mu_x + \rho\sigma_x\frac{y-\mu_y}{\sigma_y}$. Sustituyendo en (2.29)

$$(1-\rho^2)q(x,y) = \left(\frac{x-b}{\sigma_x}\right)^2 + (1-\rho^2)\left(\frac{y-\mu_y}{\sigma_y}\right)^2$$

y de aquí

$$\begin{aligned}&\int_{\mathcal{R}^2} f_{XY}(x,y)dxdy = \\ &\quad = \int_{-\infty}^{+\infty} \frac{1}{\sigma_y\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)} \left[\int_{-\infty}^{+\infty} \frac{1}{\sigma_x\sqrt{2\pi(1-\rho^2)}} e^{\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x-b}{\sigma_x}\right)^2\right)} dx\right] dy = 1, \qquad (2.30)\end{aligned}$$

porque el integrando de la integral interior es la función de densidad de una $N(b, \sigma_x^2(1-\rho^2))$ e integra la unidad. La integral resultante vale también la unidad por tratarse de la densidad de una $N(\mu_y, \sigma_y^2)$, que es precisamente la densidad marginal de Y (basta recordar la expresión (2.23) que permite obtener la densidad marginal a partir de la conjunta). Por simetría $X \sim N(\mu_x, \sigma_x^2)$.

Esta distribución puede extenderse a n dimensiones, hablaremos entonces de *Normal multivariante*. La expresión de su densidad la daremos en el próximo capítulo y utilizaremos una notación matricial que la haga más sencilla y compacta.

2.4. Independencia de variables aleatorias

La independencia entre dos sucesos A y B supone que ninguno de ellos aporta información de interés acerca del otro. Pretendemos ahora trasladar el concepto a la relación entre variables aleatorias, pero siendo un concepto originalmente definido para sucesos, la traslación deberá hacerse por medio de sucesos ligados a las variables. Para ello necesitamos recurrir al concepto de σ-álgebra inducida por una variable aleatoria que definimos en la página 18.

Definición 2.5 (Variables aleatorias independientes) *Decimos que las variables* $X_i,\ i = 1,\ldots,k$ *son* independientes *si las* σ*-álgebras que inducen lo son.*

Si recordamos lo que significaba la independencia de familias de sucesos, la definición implica que al tomar un $A_i \in \sigma(X_i),\ i = 1,\ldots,k$,

$$P(A_{j_1} \cap \ldots \cap A_{j_n}) = \prod_{l=1}^{n} P(A_{j_l}),\quad \forall (j_1,\ldots,j_n) \subset (1,\ldots,k).$$

Teniendo en cuenta cómo han sido inducidas las $\sigma(X_i)$, admite una expresión alternativa en términos de las distintas variables,

$$P(A_{j_1} \cap \ldots \cap A_{j_n}) = P(X_{j_l} \in B_{j_l}, l = 1,\ldots,n) = \prod_{l=1}^{n} P(X_{j_l} \in B_{j_l}),\ B_{j_l} \in \beta, \tag{2.31}$$

donde $A_{j_l} = X^{-1}(B_{j_l})$.

Comprobar la independencia de variables aleatorias mediante (2.31) es prácticamente imposible. Necesitamos una caracterización más de más sencilla comprobación y para encontrarla nos apoyaremos en una propiedad de las σ-álgebras engendradas.

Definición 2.6 (σ-álgebra engendrada) *Decimos que la* σ*-álgebra* $\mathcal{B}$ *está* engendrada *por la familia de conjuntos* $\mathcal{C}$*, si es la menor* σ*-álgebra que la contiene. Lo que denotaremos mediante* $\sigma(\mathcal{C}) = \mathcal{B}$*.*

La σ-álgebra de Borel, β, está engendrada, entre otras, por la familia de intervalos de la forma $\{]-\infty, x],\ x \in \mathcal{R}\}$. Como consecuencia de ello, si X es una variable aleatoria y $\mathcal{C}$ la familia

$$\mathcal{C} = X^{-1}\{]-\infty, x],\ x \in \mathcal{R}\},$$

se puede demostrar que $\sigma(X) = \sigma(\mathcal{C})$.

Podemos ahora enunciar una nueva caracterización de las σ-álgebras independientes.

Teorema 2.1 *Las* σ*-álgebras de sucesos engendradas por familias independientes y estables para la intersección, son también independientes.*

La demostración del teorema es compleja y está fuera del alcance y pretensión de estas notas.

Estamos ahora en condiciones de enunciar el teorema que ha de permitirnos comprobar con facilidad la independencia entre variables aleatorias.

Teorema 2.2 (Teorema de factorización) *Sea* $X = (X_1,\ldots,X_k)$ *un vector aleatorio cuyas funciones de distribución y densidad o cuantía conjuntas son, respectivamente,* $F_X(x_1,\ldots,x_k)$ *y* $f_X(x_1,\ldots,x_k)$*. Sean* $F_j(x_j)$ *y* $f_j(x_j)$*,* $j = 1,\ldots,k$*, las respectivas marginales. Las variables aleatorias* $X_1,\ldots,X_k$ *son independientes* sí y solo sí *se verifica alguna de las siguientes condiciones equivalentes:*

1. $F_X(x_1,\ldots,x_k) = \prod_{j=1}^{k} F_j(x_j),\ \forall (x_1,\ldots,x_k) \in \mathcal{R}^k$
2. $f_X(x_1,\ldots,x_k) = \prod_{j=1}^{k} f_j(x_j),\ \forall (x_1,\ldots,x_k) \in \mathcal{R}^k$

Demostración

1. Veamos las dos implicaciones:

   Directo.- Si las variables son independientes, $\forall (x_1, \ldots, x_k) \in \mathcal{R}^k$ definamos los conjuntos $B_j =]-\infty, x_j]$

   $$F_X(x_1, \ldots, x_k) = P(\cap_{j=1}^k X_j^{-1}(B_j)) = \prod_{j=1}^k P(X_j^{-1}(B_j)) = \prod_{j=1}^k F_j(x_j).$$

   Inverso.- Si la función de distribución conjunta se puede factorizar, las familias $\{X_i \leq x_i,\ x_i \in \mathcal{R}\}$ son independientes, pero estas familias engendran las respectivas $\sigma(X_i)$ y son estables para la intersección, el teorema 2.1 hace el resto.

2. Veamos las dos implicaciones distinguiendo el caso continuo del discreto.

   a) Caso discreto

   Directo.- Para $(x_1, \ldots, x_k) \in D_X$, soporte del vector,

   $$f_X(x_1, \ldots, x_k) = P(\cap_{j=1}^k (X_j = x_j)) = \prod_{j=1}^k P(X_j = x_j) = \prod_{j=1}^k f_j(x_j).$$

   Si el punto no está en D_X, la igualdad es trivialmente cierta porque ambos miembros son nulos.

   Inverso.- Sean $B_j =]-\infty, x_j],\ x_j \in \mathcal{R}$.

   $$\begin{aligned} F_X(x_1, \ldots, x_k) &= \\ &= P((X_1, \ldots, X_k) \in B_1 \times \ldots \times B_k) = \\ &= \sum_{(t_1, \ldots, t_k) \in B_1 \times \ldots \times B_k} f_X(t_1, \ldots, t_k) \\ &= \sum_{(t_1, \ldots, t_k) \in B_1 \times \ldots \times B_k} \prod_{j=1}^k f_j(t_j) \\ &= \prod_{j=1}^k \left[\sum_{t_j \leq x_j} f_j(t_j) \right] = \prod_{j=1}^k F_j(x_j), \end{aligned}$$

   y por 1) las variables son independientes, quedando demostrada también la equivalencia de ambas condiciones.

   a) Caso continuo

   Directo.- La independencia permite la factorización de la función de distribución conjunta,

   $$F_X(x_1, \ldots, x_k) = \prod_{j=1}^k F_j(x_j),\ \forall (x_1, \ldots, x_k) \in \mathcal{R}^k.$$

   Bastará derivar para obtener la relación deseada. Conviene recordar que $(x_1, \ldots, x_k)$ ha de ser un punto de continuidad de la función de densidad conjunta, pero como el número de puntos de discontinuidad para un vector continuo tiene medida nula, esta exigencia no afecta al resultado.

Inverso.- Al expresar F_X en función de f_X,

$$\begin{aligned} F_X(x_1,\ldots,x_k) &= \\ &= \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_k} f_X(t_1,\ldots,t_k)\,dt_1\ldots dt_k \\ &= \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_k}\prod_{j=1}^{k} f_j(t_j)\,dt_1\ldots dt_k \\ &= \prod_{j=1}^{k}\left[\int_{-\infty}^{x_j} f_j(t_j)\right] = \prod_{j=1}^{k} F_j(x_j). \end{aligned}$$ ♠

Observación 2.4 *Hemos visto anteriormente que a partir de la distribución conjunta del vector es posible conocer la distribución de cada una de sus componentes. El teorema de factorización implica que a partir de las marginales podemos reconstruir la distribución conjunta, si bien es cierto que no siempre pues se exige la independencia de las variables. La recuperación en cualquier circunstancia requiere de la noción de distribución condicionada.*

Ejemplo 2.6 *En la sección 2.3.6 estudiábamos el vector aleatorio determinado por las coordenadas de un punto elegido al azar en el círculo unidad. La densidad conjunta venía dada por*

$$f_{XY}(x,y) = \begin{cases} \dfrac{1}{\pi}, & si\ (x,y)\in C_1 \\ \\ 0, & en\ el\ resto. \end{cases}$$

Por simetría, las marginales de X e Y son idénticas y tienen la forma,

$$f_X(x) = \begin{cases} \dfrac{2}{\pi}\sqrt{1-x^2}, & si\ |x|\le 1 \\ \\ 0, & en\ el\ resto. \end{cases}$$

De inmediato se comprueba que $f_{XY}(x,y) \neq f_X(x)f_Y(y)$ y ambas variables no son independientes.

2.4.1. La aguja de Buffon

Ya hicimos mención al problema al hablar de probabilidades geométricas en el capítulo anterior (ver página 15). Veamos una forma de resolverlo mediante un vector bidimensional con componentes independientes.

Problema de la aguja de Buffon.- *Sobre una trama de líneas paralelas equidistantes entre sí* d *unidades, lanzamos al azar una aguja de longitud* l *unidades, con* l<d. *¿Cuál es la probabilidad de que la aguja corte alguna de las paralelas?*

Solución.- Si denotamos por X, la distancia del centro de la aguja a la paralela más próxima, y por Θ, el ángulo que la aguja forma con dicha paralela, la posición de la aguja sobre el entramado de paralelas queda unívocamente determinada con ambas variables. El lanzamiento al azar cabe interpretarlo en el sentido que ambas variables tienen distribuciones uniformes y son independientes. En concreto, $X \sim U(0,d)$ y $\Theta \sim U(0,\pi)$ (por razones de simetría, la

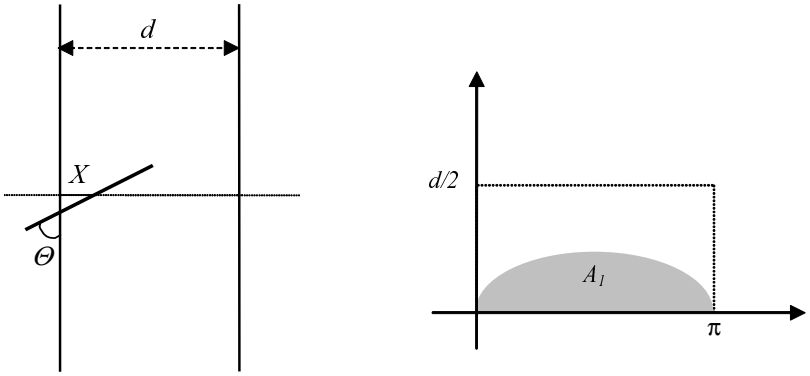
elección de un ángulo en $]\pi, 2\pi]$ duplicaría las posiciones) y su distribución conjunta tendrá por densidad:

$$f_{X\Theta}(x,\theta) = \begin{cases} \dfrac{1}{\pi d}, & \text{si } (x,\theta) \in [0,d] \times [0,\pi], \\ 0 & \text{en el resto.} \end{cases}$$

Como puede apreciarse en la figura, para cada θ fijo la aguja cortará a la paralela de su izquierda o de su derecha si,

$$0 \le x \le \frac{l}{2}\sin\theta \quad \longrightarrow \quad \text{corta a la izquierda}$$

$$0 \le d - x \le \frac{l}{2}\sin\theta \quad \longrightarrow \quad \text{corta a la derecha}$$



Si definimos los sucesos

$$A_1 = \{(x,\theta);\ 0 \le \theta \le \pi,\ 0 \le x \le \tfrac{l}{2}\sin\theta\}$$

$$A_2 = \{(x,\theta);\ 0 \le \theta \le \pi,\ d - \tfrac{l}{2}\sin\theta \le x \le d\},$$

la probabilidad de corte vendrá dada por

$$\begin{aligned} P(\text{Corte}) &= P((X,\Theta) \in A_1 \cup A_2) \\ &= \int_0^{\pi}\int_0^{\frac{l}{2}\sin\theta} \frac{1}{\pi d}\,dx d\theta + \int_0^{\pi}\int_{d-\frac{l}{2}\sin\theta}^{d} \frac{1}{\pi d}\,dx d\theta \\ &= \frac{1}{\pi d}\int_0^{\pi}\left[\frac{l}{2}\sin\theta + \frac{l}{2}\sin\theta\right] d\theta \\ &= \frac{l}{\pi d}\int_0^{\pi}\sin\theta\,d\theta \\ &= \frac{2l}{\pi d} \end{aligned} \tag{2.32}$$

2.5. Distribuciones condicionadas

2.5.1. Caso discreto

Consideremos un vector aleatorio bidimensional (X, Y), con soportes para cada una de sus componentes D_x y D_y, respectivamente, y con función de cuantía conjunta $f_{XY}(x, y)$.

Definición 2.7 *La* función de cuantía condicionada *de Y dado $\{X = x\}$, $x \in D_x$, se define mediante,*

$$f_{Y|X}(y|x) = P(Y = y|X = x) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

La función de distribución condicionada *de Y dado $\{X = x\}$, $x \in D_x$, se define mediante,*

$$F_{Y|X} = P(Y \leq y|X = x) = \frac{\sum_{v \leq y,\ v \in D_y} f_{XY}(x, v)}{f_X(x)} = \sum_{v \leq y,\ v \in D_y} f_{Y|X}(v|x).$$

La función $f_{Y|X}(y|x)$ es efectivamente una función de cuantía por cuanto cumple con las dos consabidas condiciones,

1. es no negativa por tratarse de una probabilidad condicionada, y
2. suma la unidad sobre D_y,

$$\sum_{y \in D_y} f_{Y|X}(y|x) = \frac{\sum_{y \in D_y} f_{XY}(x, y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$$

El concepto de distribución condicional se extiende con facilidad al caso k-dimensional. Si $X = (X_1, \ldots, X_k)$ es un vector aleatorio k-dimensional y $X^l = (X_{i_1}, \ldots, X_{i_l})$, $l \leq k$ y $X^{k-l} = (X_{j_1}, \ldots, X_{j_{k-l}})$ son subvectores de dimensiones complementarias, con soportes D_{x^l} y $D_{x^{k-l}}$, respectivamente, la *función de cuantía condicionada* de X^l dado $X^{k-l} = (x_{j_1}, \ldots, x_{j_{k-l}})$, $(x_{j_1}, \ldots, x_{j_{k-l}}) \in D_{x^{k-l}}$, se define mediante,

$$f_{X^l|X^{k-l}}(x_{i_1}, \ldots, x_{i_l}|x_{j_1}, \ldots, x_{j_{k-l}}) = \frac{f_X(x_1, \ldots, x_k)}{f_{X^{k-l}}(x_{j_1}, \ldots, x_{j_{k-l}})},$$

donde el argumento del numerador, $(x_1, \ldots, x_k)$, está formado por las componentes $(x_{i_1}, \ldots, x_{i_l})$ y $(x_{j_1}, \ldots, x_{j_{k-l}})$ adecuadamente ordenadas.

Ejemplo 2.7 *Consideremos dos variables aleatorias independientes X e Y, con distribución de Poisson de parámetros μ y λ, respectivamente. Queremos encontrar al distribución de la variable condicionada $X|X + Y = r$.*

Recordemos que

$$f_{X|X+Y}(k|r) = P(X = k|X + Y = r) = \frac{P(X = k, Y = r - k)}{P(X + Y = r)} = \frac{f_{XY}(k, r - k)}{f_{X+Y}(r)}. \tag{2.33}$$

La distribución conjunta del vector (X, Y) es conocida por tratarse de variables independientes,

$$f_{XY}(k, r - k) = \frac{\mu^k}{k!} e^{-\mu} \frac{\lambda^{r-k}}{r - k!} e^{-\lambda}. \tag{2.34}$$

La distribución de la variable $X+Y$ se obtiene de la forma

$$\begin{aligned} f_{X+Y}(r) &= P\left(\bigcup_{k=0}^{r}\{X=k, Y=r-k\}\right) \\ &= \sum_{k=0}^{r} f_{XY}(k, r-k) \\ &= \sum_{k=0}^{r} \frac{\mu^k \lambda^{r-k}}{k!r-k!} e^{-(\mu+\lambda)} \\ &= \frac{e^{-(\mu+\lambda)}}{r!} \sum_{k=0}^{r} \frac{r!}{k!r-k!} \mu^k \lambda^{r-k} \\ &= \frac{(\mu+\lambda)^r}{r!} e^{-(\mu+\lambda)}. \end{aligned} \tag{2.35}$$

Lo que nos dice que $X+Y \sim Po(\mu+\lambda)$.

Sustituyendo (2.34) y (2.35) en (2.33),

$$\begin{aligned} f_{X|X+Y}(k|r) &= \frac{\frac{\mu^k}{k!} e^{-\mu} \frac{\lambda^{r-k}}{r-k!} e^{-\lambda}}{\frac{(\mu+\lambda)^r}{r!} e^{-(\mu+\lambda)}} \\ &= \frac{r!}{k!(r-k)!} \frac{\mu^k \lambda^{r-k}}{(\mu+\lambda)^r} \\ &= \binom{r}{k} \left(\frac{\mu}{\mu+\lambda}\right)^k \left(1-\frac{\mu}{\mu+\lambda}\right)^{r-k}, \end{aligned}$$

concluimos que $X|X+Y=r \sim B(r, \mu/(\mu+\lambda))$.

Ejemplo 2.8 *Cuando se inicia una partida de cartas* la mano *se suele decidir a la carta más alta. Supongamos una partida entre dos jugadores que, antes de iniciarla, extraen al azar sendas cartas para decidir cuál de los dos repartirá inicialmente las cartas. Si por X designamos la mayor de las dos cartas y por Y la menor, vamos encontrar la distribución conjunta del vector (X,Y), las marginales y las condicionadas.*

La baraja española consta de 4 palos con 12 cartas cada uno de ellos, numeradas del 1 (As) al 12 (Rey). Como el As se considera siempre la carta más alta, podemos asignarle el número 13 y suponer las cartas numeradas del 2 al 13. No pueden haber empates, con lo que el problema es equivalente al de extraer al azar, sin reemplazamiento, dos bolas de una urna que contiene 12 bolas numeradas del 2 al 13. Observemos que el orden no cuenta en la extracción, sólo cuál de ellas es la mayor y cuál la menor, por lo que las posibles extracciones son $\binom{12}{2} = 66$. En definitiva,

$$f_{XY}(x,y) = \begin{cases} \dfrac{1}{66}, & 2 \le y < x \le 13; \\ 0, & \text{en el resto.} \end{cases}$$

La función de cuantía marginal de X vale,

$$f_X(x) = \sum_{2\le y<x} \frac{1}{66} = \frac{x-2}{66}, \quad x \in D_X = \{3, \ldots, 13\},$$

y $f_X(x) = 0$ *si* $x \in D_X^c$. *Para* Y,

$$f_Y(y) = \sum_{y<x<\leq 13} \frac{1}{66} = \frac{13-y}{66}, \quad y \in D_Y = \{2, \ldots, 12\},$$

y $f_Y(y) = 0$ *si* $y \in D_Y^c$.

En cuanto a las condicionadas,

$$f_{X|Y}(x|y) = \frac{1/66}{(13-y)/66} = \frac{1}{13-y}, \quad y < x \leq 13,$$

que es la distribución uniforme sobre el conjunto $D_{X|Y} = \{y+1, y+2, \ldots, 13\}$.

$$f_{Y|X}(y|x) = \frac{1/66}{(x-2)/66} = \frac{1}{x-2}, \quad 2 \leq y < x,$$

que es la distribución uniforme sobre el conjunto $D_{Y|X} = \{2, 3, \ldots, x-1\}$.

Distribuciones condicionadas en la Multinomial

Si el vector $X = (X_1, \ldots, X_k) \sim M(n; p_1, \ldots, p_k)$ sabemos que la marginal del subvector $X^l = (X_1, \ldots, X_l)$ es una $M(n; p_1, \ldots, p_l, (1-p^*))$, $p^* = p_1 + \cdots + p_l$ (en definitiva la partición de sucesos que genera la multinomial queda reducida a $A_1, \ldots, A_l, A^*$, con $A^* = \left(\cup_{i=1}^l A_i\right)^c$). La distribución condicionada de $X^{k-l} = (X_{l+1}, \ldots, X_k)$ dado $X^l = (n_1, \ldots, n_l)$, viene dada por

$$\begin{aligned} f_{X^{k-l}|X^l}(n_{l+1}, \ldots, n_k | n_1, \ldots, n_l) &= \frac{\dfrac{n!}{n_1! n_2! \ldots n_k!} \displaystyle\prod_{i=1}^k p_i^{n_i}}{\dfrac{n!}{n_1! \ldots n_l! (n-n^*)!} (1-p^*)^{(n-n^*)} \displaystyle\prod_{i=1}^l p_i^{n_i}} \\ &= \frac{(n-n^*)!}{n_{l+1}! \ldots n_k!} \prod_{i=l+1}^k \left(\frac{p_i}{1-p^*}\right)^{n_i} \end{aligned}$$

con $n^* = n_1 + \cdots + n_l$ y $\sum_{i=l+1}^k n_i = n - n^*$. Se trata, en definitiva, de una $M(n-n^*; \frac{p_{l+1}}{1-p^*}, \ldots, \frac{p_k}{1-p^*})$.

Distribuciones condicionadas en la Binomial Negativa bivariante

La función de cuantía condicionada de Y dado $X = x$, cuando $(X, Y) \sim BN(r; p_1, p_2)$, vale

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{\binom{x+y+r-1}{y}\binom{x+r-1}{x} p_1^x p_2^y (1-p_1-p_2)^r}{\binom{x+r-1}{x}\left(\frac{p_1}{1-p_2}\right)^x \left(\frac{1-p_1-p_2}{1-p_2}\right)^r} \\ &= \binom{x+y+r-1}{y} p_2^y (1-p_2)^{x+r}, \\ & \qquad\qquad x = 0, 1, \ldots \quad y = 0, 1, \ldots \end{aligned}$$

Luego $Y|X = x \sim BN(x+r, p_2)$.

2.5.2. Caso continuo

Si tratamos de trasladar al caso continuo el desarrollo anterior nos encontramos, de entrada, con una dificultad aparentemente insalvable. En efecto, si tenemos un vector bidimensional (X, Y) en la expresión

$$P(Y \leq y | X = x) = \frac{P\left(\{Y \leq y\} \cap \{X = x\}\right)}{P(X = x)}$$

el denominador $P(X = x)$ es nulo. Puede parecer que el concepto de distribución condicionada carezca de sentido en el contexto de variables continuas.

Pensemos, no obstante, en la elección de un punto al azar en C_1, círculo unidad. Fijemos la abscisa del punto en $X = x$, $|x| < 1$, y consideremos cómo se distribuirá la ordenada Y sobre la correspondiente cuerda. Estamos hablando de la distribución condicionada de $Y|X = x$, que no sólo tiene sentido, si no que intuimos que será uniforme sobre la cuerda. ¿Cómo comprobar nuestra intuición? Aceptemos en principio para el caso continuo la expresión que hemos encontrado para el caso discreto,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

Recordando las expresiones de las densidades conjuntas y las marginales obtenidas en la sección 2.3.6 y el ejemplo 2.6, tendremos

$$f_{Y|X}(y|x) = \begin{cases} \dfrac{1/\pi}{2\sqrt{1-x^2}/\pi}, & si\ |y| \leq \sqrt{1-x^2} \\ \\ 0, & en\ el\ resto, \end{cases}$$

que confirma nuestra intuición, $Y|X = x \sim U(-\sqrt{1-x^2}, +\sqrt{1-x^2})$. Parece lógico pensar que la densidad condicionada sea, efectivamente, la que hemos supuesto.

Una obtención rigurosa de las expresiones de $f_{Y|X}(y|x)$ y $F_{Y|X}(y|x)$ está fuera del alcance de esta introducción, pero una aproximación válida consiste en obtener $F_{Y|X}(y|x) = P(Y \leq y|X = x)$ como límite de $P(Y \leq y|x - \varepsilon < X \leq x + \varepsilon)$ cuando $\varepsilon \downarrow 0$ y siempre que $f_X(x) > 0$. Veámoslo.

$$\begin{aligned} F_{Y|X}(y|x) &= \lim_{\varepsilon \downarrow 0} P(Y \leq y | x - \varepsilon < X \leq x + \varepsilon) \\ &= \lim_{\varepsilon \downarrow 0} \frac{P(Y \leq y, x - \varepsilon < X \leq x + \varepsilon)}{P(x - \varepsilon < X \leq x + \varepsilon)} \\ &= \lim_{\varepsilon \downarrow 0} \frac{\int_{-\infty}^{y} \left[\int_{x-\varepsilon}^{x+\varepsilon} f_{XY}(u, v) du\right] dv}{\int_{x-\varepsilon}^{x+\varepsilon} f_X(u) du}. \end{aligned}$$

Dividiendo numerador y denominador por 2ε, pasando al límite y teniendo en cuenta la relación (2.22) que liga a las funciones de densidad y de distribución en los puntos de continuidad de aquellas,

$$F_{Y|X}(y|x) = \frac{\int_{-\infty}^{y} f_{XY}(x, v) dv}{f_X(x)} = \int_{-\infty}^{y} \frac{f_{XY}(x, v)}{f_X(x)} dv, \quad f_X(x) > 0.$$

Al derivar en la expresión anterior respecto de v obtendremos una de las posibles densidades condicionadas,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad f_X(x) > 0,$$

justamente la que hemos utilizado anteriormente.

Ambas expresiones se generalizan fácilmente para el caso de un vector X, k-dimensional, y subvectores X^l y X^{k-l} de dimensiones complementarias l y $k-l$, respectivamente.

$$F_{X^l|X^{k-l}}(x_{i_1},\ldots,x_{i_l}|x_{j_1},\ldots,x_{j_{k-l}}) = \frac{\int_{-\infty}^{x_{i_1}}\cdots\int_{-\infty}^{x_{i_l}} f_X(x_1,\ldots,x_k)dx_{i_1}\cdots dx_{i_l}}{f_{X^{k-l}}(x_{j_1},\ldots,x_{j_{k-l}})},$$

y

$$f_{X^l|X^{k-l}}(x_{i_1},\ldots,x_{i_l}|x_{j_1},\ldots,x_{j_{k-l}}) = \frac{f_X(x_1,\ldots,x_k)}{f_{X^{k-l}}(x_{j_1},\ldots,x_{j_{k-l}})},$$

con $f_{X^{k-l}}(x_{j_1},\ldots,x_{j_{k-l}}) > 0$, y donde el argumento de ambos numeradores, $(x_1,\ldots,x_k)$, está formado por las componentes $(x_{i_1},\ldots,x_{i_l})$ y $(x_{j_1},\ldots,x_{j_{k-l}})$ adecuadamente ordenadas.

Ejemplo 2.9 *Elegimos al azar X en [0,1] y a continuación Y, también al azar, en $[0,X^2]$. Es decir*

$$f_X(x) = \begin{cases} 1, & x\in[0,1]; \\ 0, & \textit{en el resto.} \end{cases} \quad \text{y} \quad f_{Y|X}(y|x) = \begin{cases} 1/x^2, & y\in[0,x^2]; \\ 0, & \textit{en el resto.} \end{cases}$$

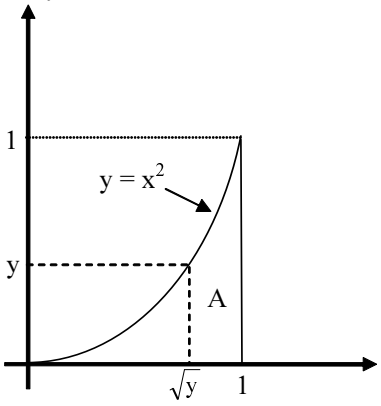
La densidad conjunta de (X,Y) vale

$$f_{XY}(x,y) = f_X(x)f_{Y|X}(y|x) = \begin{cases} \dfrac{1}{x^2}, & x\in[0,1],\ y\in[0,x^2]; \\ 0, & \textit{en el resto.} \end{cases}$$

La densidad marginal de Y es

$$f_Y(y) = \int_{\sqrt{y}}^1 \frac{1}{x^2}dx = \frac{1}{\sqrt{y}} - 1, \quad y\in[0,1],$$

y vale 0 fuera del intervalo.



Cabe preguntarse si la elección de X e Y que hemos hecho se corresponde con la elección al azar de un punto en el recinto A de la figura, determinado por la parábola $y=x^2$ entre $x=0$ y $x=1$. La respuesta es negativa, puesto que la densidad conjunta vendría dada en este caso por

$$f^*_{XY}(x,y) = \begin{cases} \dfrac{1}{\textit{área de } A} = 3, & (x,y)\in A \\ 0, & \textit{en el resto.} \end{cases}$$

y evidentemente, $f_{XY}(x,y) \neq f^*_{XY}(x,y)$.

Puede comprobarse que en este caso

$$f^*_X(x) = \begin{cases} 3x^2, & x\in[0,1]; \\ 0, & \textit{en el resto.} \end{cases} \quad \text{y} \quad f^*_{Y|X}(y|x) = \begin{cases} 1/x^2, & y\in[0,x^2]; \\ 0, & \textit{en el resto.} \end{cases}$$

Es decir, elegida la componente X la elección de Y continua siendo al azar en el intervalo $[0,X^2]$, pero a diferencia de cómo elegíamos X inicialmente, ahora ha de elegirse con la densidad $f^*_X(x)$.

Distribuciones condicionadas en la Normal bivariante

Si (X,Y) es un vector Normal bivariante,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}}}{\frac{1}{\sigma_x\sqrt{2\pi}}e^{\left(-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right)}} \\ &= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}}e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{y-\mu_y}{\sigma_y}\right)-\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\right\}^2} \\ &= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}}e^{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left\{y-\left(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right\}^2}. \end{aligned}$$

Es decir, $Y|X=x \sim N\left(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), \sigma_y^2(1-\rho^2)\right)$.

2.6. Función de una o varias variables aleatorias

2.6.1. Caso univariante

Si g es una función medible, $Y=g(X)$ es una variable aleatoria porque,

$$Y:\Omega \xrightarrow{X} \mathcal{R} \xrightarrow{g} \mathcal{R},$$

e $Y^{-1}(B)=X^{-1}[g^{-1}(B)] \in \mathcal{A}$.

Tiene sentido hablar de la distribución de probabilidad asociada a Y, que como ya hemos visto podrá ser conocida mediante cualquiera de las tres funciones: P_Y, F_Y o f_Y. Lo inmediato es preguntarse por la relación entre las distribuciones de probabilidad de ambas variables. Es aparentemente sencillo, al menos en teoría, obtener F_Y en función de F_X. En efecto,

$$F_Y(y)=P(Y\leq y)=P(g(X)\leq y)=P(X\in g^{-1}\{]-\infty,y]\}). \tag{2.36}$$

Si la variable X es discreta se obtiene la siguiente relación entre las funciones de cuantía f_Y y f_X,

$$f_Y(y)=P(Y=y)=P(g(X)=y)=\sum_{\{g^{-1}(y)\cap D_X\}} f_X(x). \tag{2.37}$$

Pero la obtención de $g^{-1}\{]-\infty,y]\}$ o $g^{-1}(y)$ no siempre es sencilla. Veamos ejemplos en los que (2.36) puede ser utilizada directamente.

Ejemplo 2.10 *Sea $X\sim U(-1,1)$ y definamos $Y=X^2$. Para obtener F_Y, sea $y\in[0,1]$,*

$$F_Y(y)=P(Y\leq y)=P(X^2\leq y)=P(-\sqrt{y}\leq X\leq\sqrt{y})=F_X(\sqrt{y})-F_X(-\sqrt{y})=\sqrt{y}.$$

Entonces,

$$F_Y(y)=\begin{cases} 0, & si\ y<0; \\ \sqrt{y}, & si\ 0\leq y\leq 1; \\ 1, & si\ y>1. \end{cases}$$

Ejemplo 2.11 *Si X es una variable discreta con soporte D_X, definamos Y mediante*

$$Y = signo(X) = \begin{cases} \frac{X}{|X|}, & si\ X \neq 0 \\ 0, & si\ X = 0. \end{cases}$$

Con esta definición, $D_Y = \{-1, 0, 1\}$, y su función de cuantía viene dada por

$$f_Y(y) = \begin{cases} \sum_{x<0} f_X(x), & si\ y = -1 \\ f_X(0), & si\ y = 0 \\ \sum_{x>0} f_X(x), & si\ y = 1 \end{cases}$$

Cuando la variable aleatoria es discreta (2.36) y (2.37) son la únicas expresiones que tenemos para obtener la distribución de probabilidad de Y. El caso continuo ofrece, bajo ciertas condiciones, otra alternativa.

Teorema 2.3 *Sea X una variable aleatoria continua y sea g monótona, diferenciable con $g'(x) \neq 0,\ \forall x$. Entonces $Y = g(X)$ es una variable aleatoria con función de densidad,*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|, & si\ y \in g(\{D_X\}) \\ 0, & en\ el\ resto. \end{cases}$$

Demostración.- Como g es medible por ser continua, Y será una variable aleatoria. Supongamos ahora que g es monótona creciente. Tendremos, para $y \in g(\{D_X\})$,

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

Derivando respecto de y obtendremos una función de densidad para Y,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(g^{-1}(y))}{dg^{-1}(y)} \cdot \frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|.$$

En caso de monotonía decreciente para g,

$$F_Y(y) = P(Y \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

El resto se obtiene análogamente. ♠

Ejemplo 2.12 *Consideremos la variable aleatoria X cuya densidad viene dada por*

$$f_X(x) = \begin{cases} 0, & si\ x < 0, \\ \frac{1}{2}, & si\ 0 \leq x \leq 1, \\ \frac{1}{2x^2}, & si\ x > 1, \end{cases}$$

Definimos una nueva variable mediante la transformación $Y = 1/X$. La transformación cumple con las condiciones del teorema, $x = g^{-1}(y) = 1/y$ y $\frac{dg^{-1}(y)}{dy} = -\frac{1}{y^2}$, por tanto la densidad de Y vendrá dada por

$$f_Y(y) = \begin{cases} 0, & si\ y < 0, \\ \frac{1}{2} \cdot \frac{1}{y^2}, & si\ 1 \leq y < \infty, \\ \frac{1}{2(1/y)^2} \cdot \frac{1}{y^2}, & si\ 0 < y < 1, \end{cases}$$

que adecuadamente ordenado da lugar a la misma densidad que poseía X.

El teorema anterior admite la siguiente generalización.

Teorema 2.4 *Sea D el dominio de g y supongamos que admite una partición finita $D = \cup_{i=1}^{n} D_i$, de manera que $g_i = g_{|D_i}$, restricción de g a D_i, es estrictamente monótona, diferenciable y con derivada no nula. La densidad de $Y = g(X)$ tiene la expresión,*

$$f_Y(y) = \sum_{i:y\in g_i(D_i)} f_X(g_i^{-1}(y)) \left| \frac{dg_i^{-1}(y)}{dy} \right|. \tag{2.38}$$

Demostración.- Si $D_i = (a_i, b_i)$, $i = 1, \ldots, n$, fijemos y en el rango de g. Tendremos

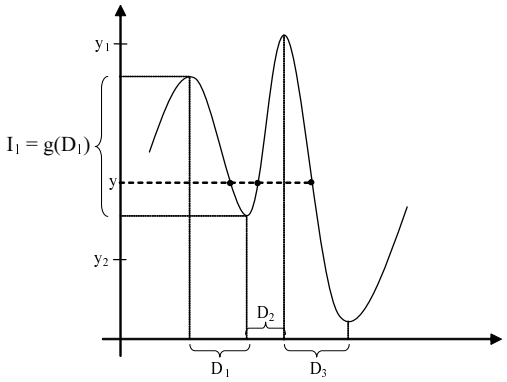
$$\begin{aligned} F_Y(y) &= P(Y \le y) = P\left(g(X) \le y, X \in \cup_{i=1}^{n} D_i\right) \\ &= \sum_{i=1}^{n} P\left(g(X) \le y, X \in D_i\right) = \sum_{i=1}^{n} P\left(g_i(X) \le y\right). \end{aligned} \tag{2.39}$$

Si $y \in g_i(D_i)$ y g_i es creciente,

$$P\left(g_i(X) \le y\right) = P\left(a_i \le X \le g_i^{-1}(y)\right) = F_X(g_i^{-1}(y)) - F_X(a_i).$$

Si g_i es decreciente,

$$P\left(g_i(X) \le y\right) = P\left(g_i^{-1}(y) \le X \le b_i\right) = F_X(b_i) - F_X(g_i^{-1}(y)).$$

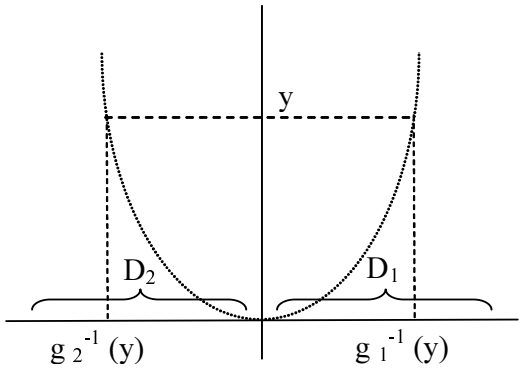


Si $y \notin g_i(D_i)$, como $I_i = g_i(D_i)$ es un intervalo, es fácil comprobar (véase la figura) que

$$\begin{aligned} P\left(g_i(X) \le y\right) &= 0 \quad si \quad \text{ínf}\, I_i > y \\ P\left(g_i(X) \le y\right) &= 1 \quad si \quad \sup I_i < y. \end{aligned}$$

En cualquiera de los dos casos, al ser constante, su contribución a f_Y es nula. En definitiva, si sustituimos en (2.39) y derivamos respecto de las $g_i^{-1}(y)$ obtendremos (2.38). ♠

Ejemplo 2.13 *Si queremos obtener la densidad de una variable aleatoria definida mediante la transformación $Y = X^2$ a partir de $X \sim N(0,1)$, observamos en la figura que $D = D_1 \cup D_2$, de manera que la restricción de g sobre cada D_i es una biyección que cumple las condiciones del teorema.*



Tenemos además que $g_1^{-1}(y) = \sqrt{y}$ *y* $g_2^{-1}(y) = -\sqrt{y}$*. Aplicando (2.38) se obtiene*

$$f_Y(y) = \begin{cases} 0, & \text{si } y < 0, \\ \dfrac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{\frac{-y}{2}}, & \text{si } y \geq 0. \end{cases}$$

Se trata de la densidad de una χ_1^2*.*

Hay dos transformaciones especialmente interesantes porque permiten obtener variables aleatorias con distribuciones preestablecidas. La primera conduce siempre a una $U(0,1)$ y la otra, conocida como *transformación integral de probabilidad*, proporciona la distribución que deseemos. Necesitamos previamente la siguiente definición.

Definición 2.8 (Inversa de una función de distribución) *Sea* F *una función en* $\mathcal{R}$ *que verifica las propiedades PF1) a PF4) de la página 19, es decir, se trata de una función de distribución de probabilidad. La* inversa de F *es la función definida mediante*

$$F^{-1}(x) = \text{ínf}\{t : \ F(t) \geq x\}.$$

Observemos que F^{-1} existe siempre, aun cuando F no sea continua ni estrictamente creciente. Como contrapartida, F^{-1} no es una inversa puntual de F, pero goza de algunas propiedades interesantes de fácil comprobación.

Proposición 2.2 *Sea* F^{-1} *la* inversa *de* F*. Entonces,*

a) *para cada* x *y* t*,* $F^{-1}(x) \leq t \longleftrightarrow x \leq F(t)$*,*

b) F^{-1} *es creciente y continua por la izquierda, y*

c) *si* F *es continua, entonces* $F(F^{-1}(x)) = x, \ \forall x \in [0,1]$*.*

Podemos ya definir las dos transformaciones antes mencionadas.

Proposición 2.3 (Transformada integral de probabilidad) *Sea* $U \sim U(0,1)$*,* F *una función de distribución de probabilidad y definimos* $X = F^{-1}(U)$*. Entonces,* $F_X = F$*.*

Demostración.- Como F^{-1} es monótona, X es una variable aleatoria. Por a) en la proposición anterior, $\forall t \in \mathcal{R}$,

$$F_X(t) = P(X \leq t) = P(F^{-1}(U) \leq t) = P(U \leq F(t)) = F(t). \qquad \spadesuit$$

Este resultado es la base de muchos procedimientos de simulación aleatoria porque permite obtener valores de cualquier variable aleatoria a partir de valores de una Uniforme, los valores de la Uniforme son a su vez generados con facilidad por los ordenadores. A fuer de ser rigurosos, debemos precisar que los ordenadores no generan exactamente valores de una Uniforme, lo que generan son valores pseudoaleatorios que gozan de propiedades semejantes a los de una Uniforme.

Proposición 2.4 (Obtención de una U(0,1)) *Si F_X es continua, $U = F_X(X) \sim U(0,1)$.*

Demostración.- Hagamos $F = F_X$. Para $x \in [0,1]$, por la proposición 2.2 a), $P(U \geq x) = P(F(X) \geq x) = P(X \geq F^{-1}(x))$. La continuidad de F y la proposición 2.2 c) hacen el resto,

$$P(U \geq x) = P(X \geq F^{-1}(x)) = 1 - F(F^{-1}(x)) = 1 - x. \qquad \spadesuit$$

2.6.2. Caso multivariante

Para $X = (X_1, \ldots, X_k)$, vector aleatorio k-dimensional, abordaremos el problema solamente para el caso continuo. La obtención de la densidad de la nueva variable o vector resultante en función de $f_X(x_1, \ldots, x_k)$ plantea dificultades en el caso más general, pero bajo ciertas condiciones, equivalentes a las impuestas para el caso univariante, es posible disponer de una expresión relativamente sencilla.

Teorema 2.5 *Sea $X = (X_1, \ldots, X_k)$ es un vector aleatorio continuo con soporte D_X y sea $g = (g_1, \ldots, g_k) : \mathcal{R}^k \to \mathcal{R}^k$ una función vectorial que verifica:*

1. *g es uno a uno sobre D_X,*
2. *el Jacobiano de g, $J = \frac{\partial(g_1,\ldots,g_k)}{\partial(x_1,\ldots,x_k)}$, es distinto de cero $\forall x \in D_X$, y*
3. *existe $h = (h_1, \ldots, h_k)$ inversa de g.*

Entonces, $Y = g(X)$ es un vector aleatorio continuo cuya densidad conjunta, para $y = (y_1, \ldots, y_k) \in g(D_X)$, viene dada por

$$f_Y(y_1, \ldots, y_k) = f_X\left(h_1(y_1, \ldots, y_k), \ldots, h_k(y_1, \ldots, y_k)\right) |J^{-1}|, \tag{2.40}$$

donde $J^{-1} = \frac{\partial(h_1,\ldots,h_k)}{\partial(y_1,\ldots,y_k)}$ es el Jacobiano de h.

Este teorema no es más que el teorema del cambio de variable en una integral múltiple y su demostración rigurosa, de gran dificultad técnica, puede encontrarse en cualquier libro de Análisis Matemático. Un argumento heurístico que justifique (2.40) puede ser el siguiente. Para cada y,

$$\begin{aligned} f_Y(y_1, \ldots, y_k) dy_1 \cdots dy_k &\approx P\left\{ Y \in \prod_{i=1}^{k} (y_i, y_i + dy_i) \right\} \\ &= P\left\{ X \in h\left(\prod_{i=1}^{k} (y_i, y_i + dy_i) \right) \right\} \\ &= f_X(h(y)) \times vol\left[h\left(\prod_{i=1}^{k} (y_i, y_i + dy_i) \right) \right], \end{aligned}$$

Pero $vol\left[h\left(\prod_{i=1}^{k}(y_i, y_i + dy_i)\right)\right]$ es precisamente $|J^{-1}| dy_1, \ldots, dy_k$.

Veamos el interés del resultado a través de los siguientes ejemplos.

Ejemplo 2.14 (Continuación del ejemplo 2.6) *En la sección 2.3.6 estudiábamos el vector aleatorio determinado por las coordenadas de un punto elegido al azar en el círculo unidad. La densidad conjunta venía dada por*

$$f_{XY}(x,y) = \begin{cases} \dfrac{1}{\pi}, & si\ (x,y) \in C_1 \\ \\ 0, & en\ el\ resto. \end{cases}$$

Consideremos ahora las coordenadas polares del punto, $R = \sqrt{X^2+Y^2}$ *y* $\Theta = \arctan Y/X$. *Para obtener su densidad conjunta, necesitamos las transformaciones inversas,* $X = R\cos\Theta$ *e* $Y = R\sin\Theta$. *El correspondiente jacobiano vale* $J_1 = R$ *y la densidad conjunta,*

$$f_{R\Theta}(r,\theta) = \begin{cases} \dfrac{r}{\pi}, & si\ (r,\theta) \in [0,1]\times[0,2\pi] \\ \\ 0, & en\ el\ resto. \end{cases}$$

Con facilidad se obtienen las marginales correspondientes, que resultan ser

$$f_R(r) = \begin{cases} 2r, & si\ r \in [0,1] \\ \\ 0, & en\ el\ resto, \end{cases}$$

y para Θ,

$$f_\Theta(\theta) = \begin{cases} \dfrac{1}{2\pi}, & si\ \theta \in [0,2\pi] \\ \\ 0, & en\ el\ resto. \end{cases}$$

Como $f_{R\Theta}(r,\theta) = f_R(r)f_\Theta(\theta),\ \forall(r,\theta)$, R *y* Θ son independientes.

Ejemplo 2.15 (Suma y producto de dos variables aleatorias) *Sea* $X = (X_1, X_2)$ *un vector aleatorio bidimensional con densidad conjunta* $f_X(x_1,x_2)$. *Definimos* $U = X_1 + X_2$ *y queremos obtener su densidad. Para poder utilizar el resultado anterior la transformación debe de ser también bidimensional, cosa que conseguimos si definimos una nueva variable* $V = X_1$. *Con* $Y = (U,V)$ *podemos aplicar el teorema, siendo la inversa* $X_1 = V$ *y* $X_2 = U - V$, *cuyo Jacobiano es* $J^{-1} = -1$. *Tendremos pues,*

$$f_Y(u,v) = f_X(v, u-v),$$

y para obtener la densidad marginal de la suma, U,

$$f_U(u) = \int_{-\infty}^{+\infty} f_X(v, u-v)\, dv. \tag{2.41}$$

Para obtener la densidad de $W = X_1X_2$, *definimos* $T = X_1$ *y actuamos como antes. Con* $Y = (T,W)$ *y transformaciones inversas* $X_1 = T$ *y* $X_2 = W/T$, *el Jacobiano es* $J^{-1} = 1/T$ *y la densidad conjunta de* Y,

$$f_Y(t,w) = \frac{1}{|t|} f_X\left(t, \frac{w}{t}\right).$$

La marginal del producto se obtiene integrando respecto de la otra componente,

$$f_W(w) = \int_{-\infty}^{+\infty} \frac{1}{|t|} f_X\left(t, \frac{w}{t}\right) dt. \tag{2.42}$$

Hubieramos podido también proceder utilizando la transformación bidimensional $Y = (Y_1, Y_2)$, con $Y_1 = X_1 + X_2$ e $Y_2 = X_1 X_2$, lo que en teoría nos hubiera hecho ganar tiempo; pero sólo en teoría, porque en la práctica las inversas hubieran sido más complicadas de manejar que las anteriores.

Ejemplo 2.16 (Distribución del máximo y el mínimo) *Si $X_1, X_2, \ldots, X_n$ son n variables aleatorias independientes, vamos a obtener la distribución de probabilidad del máximo y el mínimo de todas ellas. Definimos*

$$X_M = \text{máx}\{X_1, X_2, \ldots, X_n\}, \qquad X_m = \text{mín}\{X_1, X_2, \ldots, X_n\}.$$

Distribución del máximo.- *Lo más sencillo es obtener la función de distribución de X_M, puesto que*

$$\{X_M \leq x\} \iff \bigcap_{i=1}^{n} \{X_i \leq x\},$$

y de aquí

$$F_{X_M}(x) = P\left(\cap_{i=1}^{n} \{X_i \leq x\}\right) = \prod_{i=1}^{n} F_i(x).$$

La función de densidad puede obtenerse derivando la expresión anterior,

$$f_{X_M}(x) = \sum_{i=1}^{n} \left[f_i(x) \prod_{j \neq i} F_j(x) \right].$$

Un caso especial es aquel en el que las variables tiene todas la misma distribución de probabilidad. Si F y f son, respectivamente, las funciones de distribución y densidad comunes a todas ellas,

$$F_{X_M}(x) = [F(x)]^n$$

y

$$f_{X_M}(x) = n\,[F(x)]^{n-1} f(x).$$

Distribución del mínimo.- *Observemos ahora que*

$$\{X_m > x\} \iff \bigcap_{i=1}^{n} \{X_i > x\},$$

y de aquí

$$F_{X_M}(x) = 1 - P(X_m > x) = 1 - P\left(\cap_{i=1}^{n} \{X_i > x\}\right) = 1 - \prod_{i=1}^{n} \left[1 - F_i(x)\right],$$

y

$$f_{X_M}(x) = \sum_{i=1}^{n} \left[f_i(x) \prod_{j \neq i} (1 - F_j(x)) \right].$$

Si todas las variables comparten una distribución común con F y f como funciones de distribución y densidad, respectivamente,

$$F_{X_M}(x) = 1 - [1 - F(x)]^n$$

y

$$f_{X_M}(x) = n\,[1 - F(x)]^{n-1} f(x).$$

Distribución conjunta del máximo y el mínimo.- *Para obtener al distribución conjunta de X_M y X_m observemos que,*

$$\{X_M \leq x\} = (\{X_M \leq x\} \cap \{X_m \leq y\}) \cup (\{X_M \leq x\} \cap \{X_m > y\}) .$$

Al tomar probabilidades hemos de tener en cuenta que

$$P(\{X_M \leq x\} \cap \{X_m > y\}) = \begin{cases} 0, & si\ x \leq y; \\ P\left(\cap_{i=1}^n \{y < X_i \leq x\}\right), & si\ x > y. \end{cases}$$

En definitiva

$$F_{X_M X_m}(x,y) = \begin{cases} \prod_{i=1}^n F_i(x) - \prod_{i=1}^n [F_i(x) - F_i(y)], & si\ x > y; \\ \prod_{i=1}^n F_i(x), & si\ x \leq y, \end{cases}$$

resultado que nos indica que X_M y X_m no son independientes. La función de densidad conjunta la obtendremos mediante (2.22),

$$f_{X_M X_m}(x,y) = \begin{cases} \sum_{i \neq j} \left[f_i(x) f_j(y) \prod_{k \neq i,j} [F_k(x) - F_k(y)] \right], & si\ x > y; \\ 0, & si\ x \leq y. \end{cases}$$

Cuando las variables tiene la misma distribución con F y f como funciones de distribución y densidad comunes, respectivamente,

$$F_{X_M X_m}(x,y) = \begin{cases} [F(x)]^n - [F(x) - F(y)]^n , & si\ x > y; \\ [F(x)]^n , & si\ x \leq y, \end{cases}$$

y

$$f_{X_M X_m}(x,y) = \begin{cases} n(n-1) f(x) f(y) [F(x) - F(y)]^{n-2}, & si\ x > y; \\ 0, & si\ x \leq y. \end{cases}$$

Ejemplo 2.17 (Suma de Gammas independientes) *Supongamos que X_1 y X_2 son variables aleatorias independientes Gamma con parámetros (α_i, β), $i = 1, 2$. La densidad de $U = X_1 + X_2$ la podemos obtener aplicando (2.41). Como las variables son independientes podemos factorizar la densidad conjunta y*

$$\begin{aligned} f_{X_1 X_2}(v, u-v) &= f_{X_1}(v) f_{X_2}(u-v) \\ &= \left[\frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} v^{\alpha_1 - 1} e^{-v/\beta} \right] \left[\frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} (u-v)^{\alpha_2 - 1} e^{-(u-v)/\beta} \right] \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1 + \alpha_2}} e^{-u/\beta} v^{\alpha_1 - 1} (u-v)^{\alpha_2 - 1} \end{aligned}$$

y de aquí, teniendo en cuenta que $0 \leq v \leq u$,

$$f_U(u) = \int_0^u f_X(v, u-v)\, dv = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1 + \alpha_2}} e^{-u/\beta} \int_0^u v^{\alpha_1 - 1} (u-v)^{\alpha_2 - 1} dv. \qquad (2.43)$$

Haciendo el cambio $z = v/u$, $0 \leq z \leq 1$, la integral del último miembro quedará de la forma,

$$\int_0^u v^{\alpha_1-1}(u-v)^{\alpha_2-1}dv = u^{\alpha_1+\alpha_2-1}\int_0^1 z^{\alpha_1-1}(1-z)^{\alpha_2-1}dz = u^{\alpha_1+\alpha_2-1}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

Sustituyendo en (2.43),

$$f_U(u) = \frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}}u^{\alpha_1+\alpha_2-1}e^{-u/\beta}.$$

Es decir, $U \sim Gamma(\alpha_1+\alpha_2, \beta)$.

Reiterando el proceso para un vector con k componentes independientes, $X_i \sim Gamma(\alpha_i, \beta)$, encontraríamos que la suma de sus componentes, $U = X_1 + X_2 + \cdots + X_k$, es una distribución $Gamma(\sum_{i=1}^k \alpha_i, \beta)$.

Dos consecuencias inmediatas de este resultado:

1. *La suma de k exponenciales independientes con parámetro común λ es una $Gamma(k, 1/\lambda)$.*
2. *Para $X_i \sim N(0,1)$, $i = 1, \ldots, k$, independientes, sabemos por el ejemplo 2.13 que cada $X_i^2 \sim \chi_1^2$, por tanto $Y = \sum_{i=1}^k X_i^2 \sim \chi_k^2$.*

Como corolario de la definición (2.5) se comprueba fácilmente que las transformaciones medibles de variables independientes también lo son.

Corolario 2.1 *Si g_i, $i = 1, \ldots, n$, son funciones medibles de $\mathcal{R}$ en $\mathcal{R}$, las variables aleatorias $Y_i = g_i(X_i)$, $i = 1, \ldots, n$ son independientes si las X_i lo son.*

Demostración.- El resultado se sigue de la relación,

$$\forall i \text{ y } \forall B \in \beta, \quad Y_i^{-1}(B) = X_i^{-1}[g_i^{-1}(B)] \in \sigma(X_i).$$

que implica $\sigma(Y_i) \subset \sigma(X_i)$, $\forall i$, lo que supone la independencia de las Y_i. ♠

Ejemplo 2.18 (Combinación lineal de Normales independientes) *Sean $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_1 2, \sigma_1 2^2)$ variables independientes. Por el Lema 2.2 sabemos que las variables $Y_i = a_i X_i \sim N(a_i\mu_i, a_i^2\sigma_i^2)$, $i = 1, 2$. Por otra parte, el anterior corolario nos asegura la independencia de Y_1 e Y_2 y la distribución conjunta de ambas variables tendrá por densidad el producto de sus respectivas densidades,*

$$f_{Y_1Y_2}(y_1, y_2) = \frac{1}{2\pi a_1\sigma_1 a_2\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{y_1 - a_1\mu_1}{a_1\sigma_1}\right)^2 + \left(\frac{y_2 - a_2\mu_2}{a_2\sigma_2}\right)^2\right]\right\}.$$

Si hacemos $U = Y_1 + Y_2$ y $V = Y_1$ y aplicamos (2.41),

$$f_U(u) = \int_{-\infty}^{+\infty} f_{Y_1Y_2}(v, u-v)\, dv, \tag{2.44}$$

con

$$f_{Y_1Y_2}(v, u-v) = \frac{1}{2\pi a_1\sigma_1 a_2\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{v - a_1\mu_1}{a_1\sigma_1}\right)^2 + \left(\frac{(u-v) - a_2\mu_2}{a_2\sigma_2}\right)^2\right]\right\}.$$

Simplifiquemos la forma cuadrática del exponente,

$$\begin{aligned} q(u,v) &= -\frac{1}{2}\left[\left(\frac{v-a_1\mu_1}{a_1\sigma_1}\right)^2+\left(\frac{(u-v)-a_2\mu_2}{a_2\sigma_2}\right)^2\right] \\ &= -\frac{1}{2}\left[\left(\frac{v-a_1\mu_1}{a_1\sigma_1}\right)^2+\left(\frac{u-(a_1\mu_1+a_2\mu_2)-(v-a_1\mu_1)}{a_2\sigma_2}\right)^2\right]. \end{aligned}$$

Haciendo $t = v - a_1\mu_1$ *y* $z = u-(a_1\mu_1+a_2\mu_2)$ *y operando, la última expresión puede escribirse*

$$q(u,v) = -\frac{1}{2}\cdot\frac{(a_1\sigma_1)^2+(a_2\sigma_2)^2}{(a_1\sigma_1)^2(a_2\sigma_2)^2}\left[\left(t-\frac{(a_1\sigma_1)^2(a_2\sigma_2)^2}{a_1\sigma_1)^2+(a_2\sigma_2)^2}\right)^2+\frac{z^2(a_1\sigma_1)^2(a_2\sigma_2)^2}{(a_1\sigma_1)^2+(a_2\sigma_2)^2}\right],$$

con lo que la exponencial quedará de la forma

$$\begin{aligned} \exp\{q(z,t)\} &= \exp\left\{-\frac{1}{2}\cdot\frac{(a_1\sigma_1)^2+(a_2\sigma_2)^2}{(a_1\sigma_1)^2(a_2\sigma_2)^2}\left(t-\frac{(a_1\sigma_1)^2(a_2\sigma_2)^2}{(a_1\sigma_1)^2+(a_2\sigma_2)^2}z\right)^2\right\}\times \\ &\quad \exp\left\{-\frac{1}{2}\left(\frac{z}{[(a_1\sigma_1)^2+(a_2\sigma_2)^2]^{1/2}}\right)^2\right\}. \end{aligned}$$

Sustituyendo en (2.44),

$$\begin{aligned} f_U(u) &= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{u-(a_1\mu_1+a_2\mu_2)}{[(a_1\sigma_1)^2+(a_2\sigma_2)^2]^{1/2}}\right)^2\right\}\times \\ &\quad \int_{-\infty}^{+\infty}\frac{1}{a_1\sigma_1a_2\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\cdot\frac{(a_1\sigma_1)^2+(a_2\sigma_2)^2}{(a_1\sigma_1)^2(a_2\sigma_2)^2}\left(t-\frac{(a_1\sigma_1)^2(a_2\sigma_2)^2}{(a_1\sigma_1)^2+(a_2\sigma_2)^2}z\right)^2\right\}dt. \\ &= \frac{1}{\sqrt{2\pi[(a_1\sigma_1)^2+(a_2\sigma_2)^2]}}\exp\left\{-\frac{1}{2}\left(\frac{u-(a_1\mu_1+a_2\mu_2)}{[(a_1\sigma_1)^2+(a_2\sigma_2)^2]^{1/2}}\right)^2\right\}\times \\ &\quad \int_{-\infty}^{+\infty}\frac{[(a_1\sigma_1)^2+(a_2\sigma_2)^2]^{1/2}}{a_1\sigma_1a_2\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\cdot\frac{(a_1\sigma_1)^2+(a_2\sigma_2)^2}{(a_1\sigma_1)^2(a_2\sigma_2)^2}\left(t-\frac{(a_1\sigma_1)^2(a_2\sigma_2)^2}{(a_1\sigma_1)^2+(a_2\sigma_2)^2}z\right)^2\right\}dt. \end{aligned}$$

Observemos que el integrando es la función de densidad de una variable aleatoria Normal con parámetros $\mu = \frac{z[(a_1\sigma_1)^2(a_2\sigma_2)^2]}{(a_1\sigma_1)^2+(a_2\sigma_2)^2}$ *y* $\sigma^2 = \frac{(a_1\sigma_1)^2+(a_2\sigma_2)^2}{(a_1\sigma_1)^2(a_2\sigma_2)^2}$*, por tanto la integral valdrá 1. En definitiva,*

$$f_U(u) = \frac{1}{\sqrt{2\pi[(a_1\sigma_1)^2+(a_2\sigma_2)^2]}}\exp\left\{-\frac{1}{2}\left(\frac{u-(a_1\mu_1+a_2\mu_2)}{[(a_1\sigma_1)^2+(a_2\sigma_2)^2]^{1/2}}\right)^2\right\},$$

lo que nos dice que $U = a_1X_1 + a_2X_2 \sim N(a_1\mu_1+a_2\mu_2, (a_1\sigma_1)^2+(a_2\sigma_2)^2)$.

El resultado puede extenderse a una combinación lineal finita de variables aleatoria Normales independientes. Así, si $X_i \sim N(\mu_i, \sigma_i^2)$*,*

$$X = \sum_{1=1}^{n} a_iX_i \sim N\left(\sum_{i=1}^{n}a_i\mu_i, \sum_{i=1}^{n}(a_i\sigma_i)^2\right).$$

Ejemplo 2.19 (La distribución t de Student) *Consideremos las $n+1$ variables aleatorias $X, X_1, X_2, \ldots, X_n$, independientes y todas ellas $N(0,1)$. Definimos las variable*

$$t = \frac{X}{Y}. \tag{2.45}$$

con $Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$.

Para obtener la distribución de t observemos que de acuerdo con el ejemplo anterior, $U = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$. Su densidad es (ver página 32)

$$f_U(u) = \begin{cases} 0, & si\ u \leq 0 \\ \dfrac{1}{\Gamma(n/2)2^{n/2}} u^{u/2-1} e^{-u/2}, & si\ u > 0. \end{cases}$$

La variable $Y = \sqrt{U/n}$ tendrá por densidad

$$f_Y(y) = \begin{cases} 0, & si\ y \leq 0 \\ \dfrac{n^{n/2}}{\Gamma(n/2)2^{n/2-1}} y^{n-1} e^{-y^2/2}, & si\ y > 0. \end{cases}$$

Por otra parte, por el Corolario 2.1 X e Y son independientes y su densidad conjunta vendrá dada por

$$f_{XY}(x,y) = \left(\frac{2}{\pi}\right)^{1/2} \frac{\left(\frac{n}{2}\right)^{n/2}}{\Gamma(n/2)} e^{-\frac{1}{2}(x^2+ny^2)} y^{n-1},$$

para $x \in \mathcal{R}$ e $y > 0$.

Si hacemos el doble cambio

$$t = \frac{X}{Y}, \qquad U = Y,$$

el Jacobiano de la transformación inversa vale $J = u$ y la densidad conjunta de t, U es

$$f_{XY}(x,y) = Ce^{-\frac{1}{2}(t^2u^2+nu^2)} u^n, \qquad t \in \mathcal{R}, \quad u > 0,$$

con $C = \left(\frac{2}{\pi}\right)^{1/2} \frac{\left(\frac{n}{2}\right)^{n/2}}{\Gamma(n/2)}$. La densidad marginal de t se obtiene de la integral

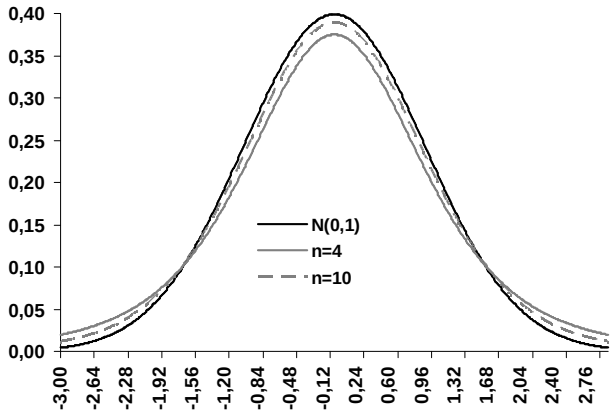
$$f_t(t) = \int_0^\infty Ce^{-\frac{1}{2}(t^2u^2+nu^2)} u^n du.$$

Haciendo el cambio $u^2 = v$

$$\begin{aligned} f_t(t) &= \int_0^\infty \frac{C}{2} e^{-v\left(\frac{1}{2}(t^2+n)\right)} v^{\frac{n-1}{2}} dv \\ &= \frac{C}{2}\left(\frac{t^2+n}{2}\right)^{-\frac{n-1}{2}} \int_0^\infty e^{-v\left(\frac{1}{2}(t^2+n)\right)} \left[v\left(\frac{1}{2}(t^2+n)\right)\right]^{\frac{n+1}{2}-1} dv \\ &= \frac{C}{2}\left(\frac{t^2+n}{2}\right)^{-\frac{n+1}{2}} \int_0^\infty e^{-z} z^{\frac{n+1}{2}-1} dz \\ &= \frac{C}{2}\left(\frac{t^2+n}{2}\right)^{-\frac{n+1}{2}} \Gamma\left(\frac{n+1}{2}\right). \end{aligned}$$

La distribución recibe el nombre de t de Student *con n grados de libertad. Student era el seudónimo de W. L. Gosset, estadístico inglés de siglo XIX, quién descubrió por primera vez esta distribución de manera empírica.*

El interés de la t de Student reside en que surge de forma natural al estudiar la distribución de algunas características ligadas a una muestra de tamaño n de una variable $N(\mu, \sigma^2)$. *Si representamos gráficamente la densidad para distintos valores de n comprobaremos que su forma es muy parecida a la de una* $N(0,1)$.



Se observa en la gráfica que a medida que n aumenta la curva de la t de Student se asemeja más a la de la Normal. No es casual este comportamiento, en efecto, la densidad de t puede escribirse de la forma,

$$f_t(t) = \frac{\left(\frac{n}{2}\right)^{n/2}}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{t^2+n}{2}\right)^{-\frac{n+1}{2}} = \frac{1}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n}{2}} \left(\frac{t^2+n}{2}\right)^{-\frac{1}{2}}.$$

Al pasar al límite cuando $n \to \infty$,

$$f_t(t) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n}{2}} \left(\frac{t^2+n}{2}\right)^{-\frac{1}{2}} \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-t^2/2},$$

que es la densidad de la $N(0,1)$.

Ejemplo 2.20 (La distribución F de Snedecor) *Sean* $(X_1, X_2, \ldots, X_m)$ *e* $(Y_1, Y_2, \ldots, Y_n)$ *sendos vectores aleatorios cuyas componentes son todas ellas variables aleatorias independientes* $N(0,1)$. *Queremos obtener la distribución de una nueva variable F definida mediante la relación*

$$F = \frac{\frac{1}{m}\sum_1^m X_i^2}{\frac{1}{n}\sum_1^n Y_j^2}.$$

Razonando de manera análoga a como lo hemos hecho en la obtención de la distribución de la t de Student, la densidad de F tiene la expresión,

$$f_F(x) \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{m/2-1}\left(1+\dfrac{mx}{n}\right)^{-(m+n)/2}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Esta distribución se conoce con el nombre de F de Snedecor con m y n grados de libertad *y surge al estudiar la distribución del cociente de las varianzas muestrales asociadas a sendas muestras de tamaño m y n de variables aleatorias $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente.*

De la F de Snedecor se deriva una nueva variable, la z de Fisher*, mediante la transformación*

$$z = \frac{1}{2} \ln F.$$

Capítulo 3

Esperanza

3.1. Introducción

En el capítulo precedente hemos visto que la descripción completa de una variable o de un vector aleatorio nos la proporciona cualquiera de las funciones allí estudiadas. Es cierto que unas son de manejo más sencillo que otras, pero todas son equivalentes para el cometido citado.

En ocasiones no necesitamos un conocimiento tan exhaustivo y nos basta con una idea general. Ciertas características numéricas ligadas a las variables o los vectores aleatorios pueden satisfacernos. Estas cantidades son muy importantes en Teoría de la Probabilidad y sus aplicaciones, y su obtención se lleva a cabo a partir de las correspondientes distribuciones de probabilidad.

Entre estas constantes, sin duda las que denominaremos *esperanza matemática* y *varianza* son las de uso más difundido. La primera juega el papel de centro de gravedad de la distribución y nos indica alrededor de qué valor se situa nuestra variable o vector. La segunda completa la información indicándonos cuan dispersos o agrupados se presentan los valores alrededor de aquella. Existen también otras constantes que proporcionan información acerca de la distribución de probabilidad, son los llamados *momentos*, de los cuales esperanza y varianza son casos particulares. Los momentos pueden llegar a aportarnos un conocimiento exhaustivo de la variable aleatoria.

La herramienta que nos permite acceder a todas estas cantidades es el concepto de *esperanza* del que nos ocupamos a continuación.

3.2. Esperanza de una variable aleatoria

Un mínimo rigor en la definición del concepto de esperanza nos exige aludir a lo largo de la exposición a algunos resultados de Teoria de la Medida e Integración.

Comencemos recordando que el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ es un caso particular de *espacio de medida* en el que ésta es una medida de probabilidad, cuya definición y propiedades conocemos del primer capítulo. En este nuevo contexto, la variable aleatoria X es simplemente una *función medible*, siendo la medibilidad una condición ya conocida por nosotros: $X^{-1}(B) \in \mathcal{A},\ \forall B \in \beta$. En el espacio de probabilidad podemos definir la integral respecto de P y diremos que X es integrable respecto de P si $\int_\Omega |X|\, dP < +\infty$.

Estamos ya en condiciones de dar la definición más general de esperanza de una variable aleatoria, si bien es cierto que buscaremos de inmediato una *traducción* que nos haga sencilla y accesible su obtención.

Definición 3.1 (Esperanza de una variable aleatoria) *Si X, variable aleatoria definida sobre el espacio de probabilidad $(\Omega, \mathcal{A}, P)$, es integrable en Ω respecto de P, diremos que existe su* esperanza *o* valor esperado*, cuyo valor es*

$$E(X) = \int_{\Omega} X \, d\mathcal{P}. \tag{3.1}$$

Si g es una función medible definida de $(\mathcal{R}, \beta)$ en $(\mathcal{R}, \beta)$, ya hemos visto anteriormente que $g(X)$ es una variable aleatoria, cuya esperanza vamos a suponer que existe. Un resultado de Integración, el teorema del cambio de variable, permite trasladar la integral a $\mathcal{R}$ y expresarla en términos de P_X.

$$E[g(X)] = \int_{\Omega} g(X) \, dP = \int_{\mathcal{R}} g \, dP_X. \tag{3.2}$$

Hemos dado un primer paso, todavía insuficiente, en la dirección de expresar la esperanza en un espacio que admita una expresión más manejable. Observemos que nada se ha dicho todavía acerca del tipo de variable con el que estamos trabajando, ya que (3.1) es absolutamente general. Si queremos finalizar el proceso de simplificación de (3.1) se hace imprescindible atender a las características de la distribución de X.

X **es discreta.-** Ello supone que D_X, soporte de P_X, es numerable y la integral se expresa en la forma

$$E[g(X)] = \sum_{x_i \in D_X} g(x_i) P_X(\{x_i\}) = \sum_{x_i \in D_X} g(x_i) P(X = x_i) = \sum_{x_i \in D_X} g(x_i) f_X(x_i). \tag{3.3}$$

X **es continua.-** Entonces P_X es absolutamente continua respecto de la medida de Lebesgue, λ, y si f_X es la densidad de probabilidad y además integrable Riemann, un resultado de Integración nos permite escribir (3.2) de la forma

$$E[g(X)] = \int_{\mathcal{R}} g \, dP_X = \int_{\mathcal{R}} g f_X \, d\lambda = \int_{-\infty}^{+\infty} g(x) f(x) dx. \tag{3.4}$$

Observación 3.1 *Es costumbre escribir también 3.2 de la forma*

$$E[g(X)] = \int_{\Omega} g(X) \, dP = \int_{\mathcal{R}} g \, dF_X,$$

que se justifica porque $\mathcal{P}_X$ es la medida de Lebesgue-Stieltjes engendrada por F_X.

3.2.1. Momentos de una variable aleatoria

Formas particulares de $g(X)$ dan lugar lo que denominamos *momentos de X*. En la tabla resumimos los distintos tipos de momentos y la correspondiente función que los origina, siempre que ésta sea integrable pues de lo contrario la esperanza no existe.

	de orden k	**absoluto de orden k**
Respecto del origen	X^k	$\lvert X\rvert^k$
Respecto de a	$(X-a)^k$	$\lvert X-a\rvert^k$
Factoriales	$X(X-1)\dots(X-k+1)$	$\lvert X(X-1)\dots(X-k+1)\rvert$

Tabla 1.- Forma de $g(X)$ para los distintos momentos de X

Respecto de la existencia de los momentos se verifica el siguiente resultado.

Proposición 3.1 *Si $E(X^k)$ existe, existen todos los momentos de orden inferior.*

La comprobación es inmediata a partir de la desigualdad $|X|^j \leq 1 + |X|^k,\ j \leq k$.

Ya hemos dicho en la introducción que el interés de los momentos de una variable aleatoria estriba en que son características numéricas que resumen su comportamiento probabilístico. Bajo ciertas condiciones el conocimiento de todos los momentos permite conocer completamente la distribución de probabilidad de la variable.

Especialmente relevante es el caso $k = 1$, cuyo correspondiente momento coincide con $E(X)$ y recibe también el nombre de *media*. Suele designarse mediante la letra griega μ (μ_X, si existe riesgo de confusión). Puesto que μ es una constante, en la tabla anterior podemos hacer $a = \mu$, obteniendo así una familia de momentos respecto de μ que tienen nombre propio: los *momentos centrales de orden k*, $E[(X - \mu)^k]$. De entre todos ellos cabe destacar la *varianza*,

$$Var(X) = \sigma_X^2 = E[(X - \mu)^2].$$

Propiedades de $E(X)$ y $V(X)$

Un primer grupo de propiedades no merecen demostración dada su sencillez. Conviene señalar que todas ellas derivan de las propiedades de la integral.

1. Propiedades de $E(X)$.

   PE1) La esperanza es un operador lineal,

   $$E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)].$$

   En particular,

   $$E(aX + b) = aE(X) + b.$$

   PE2) $P(a \leq X \leq b) = 1 \Longrightarrow a \leq E(X) \leq b$.

   PE3) $P(g(X) \leq h(X)) = 1 \Longrightarrow E[g(X)] \leq E[h(X)]$.

   PE4) $|E[g(X)]| \leq E[|g(X)|]$.

2. Propiedades de $V(X)$.

   PV1) $V(X) \geq 0$.

   PV2) $V(aX + b) = a^2 V(X)$.

   PV3) $V(X) = E(X^2) - [E(X)]^2$.

   PV4) $V(X)$ hace mínima $E\left[(X - a)^2\right]$.

   En efecto,

   $$\begin{aligned} E\left[(X-a)^2\right] &= E\left[(X - E(X) + E(X) - a)^2\right] \\ &= E\left[(X - E(X))^2\right] + E\left[(E(X) - a)^2\right] + 2E\left[(X - E(X)(E(X) - a)\right] \\ &= V(X) + (E(X) - a)^2. \end{aligned}$$

El siguiente resultado nos ofrece una forma alternativa de obtener la $E(X)$ cuando X es no negativa.

Proposición 3.2 *Si para $X \geq 0$, existe $E(X)$, entonces*

$$E(X) = \int_0^{+\infty} P(X > x)\, dx \ = \int_0^{+\infty} (1 - F_X(x))\, dx \tag{3.5}$$

Demostración.- Consideremos las dos situaciones posibles:

X **es una variable continua.-** Puesto que $E(X)$ existe, tendremos

$$E(X) = \int_0^{+\infty} x f_X(x)\, dx = \lim_{n\to+\infty} \int_0^n x f_X(x)\, dx.$$

Integrando por partes,

$$\int_0^n x f_X(x)\, dx = x F_X(x)|_0^n - \int_0^n F_X(x)\, dx = n F_X(n) - \int_0^n F_X(x)\, dx,$$

y sumando y restando n, tendremos

$$\begin{aligned} \int_0^n x f_X(x)\, dx &= -n + n F_X(n) + n - \int_0^n F_X(x)\, dx \\ &= -n(1 - F_X(n)) + \int_0^n (1 - F_X(x))\, dx. \end{aligned}$$

Pero,

$$n(1 - F_X(n)) = n \int_n^{+\infty} f_X(x))\, dx < \int_n^{+\infty} x f_X(x))\, dx \overset{n\uparrow+\infty}{\longrightarrow} 0$$

al ser $E(|X|) < +\infty$. En definitiva,

$$E(X) = \int_0^{+\infty} x f_X(x)\, dx = \lim_{n\to+\infty} \int_0^n (1 - F_X(x))\, dx = \int_0^{+\infty} (1 - F_X(x))\, dx.$$

X **es una variable discreta.-** Si D_X es el soporte de X, $E(X)$ viene dada por

$$E(X) = \sum_{x_j \in D_X} x_j f_X(x_j).$$

Si $I = \int_0^{+\infty} (1 - F_X(x))\, dx$,

$$I = \sum_{k\geq 1} \int_{(k-1)/n}^{k/n} P(X > x)\, dx,$$

y puesto que $P(X > x)$ es decreciente en x, para $(k-1)/n \leq x \leq k/n$ y $\forall n$, tendremos

$$P(X > k/n) \leq P(X > x) \leq P(X > (k-1)/n).$$

Integrando y sumando sobre k,

$$\frac{1}{n} \sum_{k\geq 1} P(X > k/n) \leq I \leq \frac{1}{n} \sum_{k\geq 1} P(X > (k-1)/n),\ \forall n. \tag{3.6}$$

Si llamamos L_n al primer miembro de la desigualdad,

$$L_n = \frac{1}{n} \sum_{k\geq 1} \sum_{j\geq k} P\left(\frac{j}{n} < X \leq \frac{j+1}{n}\right) = \frac{1}{n} \sum_{k\geq 1} (k-1) P\left(\frac{k-1}{n} < X \leq \frac{k}{n}\right).$$

Así,

$$\begin{aligned} L_n &= \sum_{k\geq 1} \frac{k}{n} P\left(\frac{k-1}{n} < X \leq \frac{k}{n}\right) - \sum_{k\geq 1} \frac{1}{n} P\left(\frac{k-1}{n} < X \leq \frac{k}{n}\right) \\ &= \sum_{k\geq 1} \frac{k}{n} \left[\sum_{\frac{k-1}{n} < x_j \leq \frac{k}{n}} f_X(x_j)\right] - \frac{1}{n} P\left(X > \frac{1}{n}\right) \\ &\geq \sum_{k\geq 1} \left[\sum_{\frac{k-1}{n} < x_j \leq \frac{k}{n}} x_j f_X(x_j)\right] - \frac{1}{n} = E(X) - \frac{1}{n},\ \forall n. \end{aligned}$$

Análogamente se comprueba que el tercer miembro de (3.6), U_n, está acotado por

$$U_n \leq E(X) + \frac{1}{n}.$$

Reuniendo todas las desigualdades se llega al resultado enunciado,

$$E(X) - \frac{1}{n} \leq \int_0^{+\infty} (1 - F_X(x))\, dx \leq E(X) + \frac{1}{n},\ \forall n.$$

♠

Observación 3.2 *Como $P(X > x)$ y $P(X \geq x)$ son funciones que difieren a lo sumo en un conjunto numerable de valores de x, son iguales casi por todas partes respecto de la medida de Lebesgue. Ello permite escribir (3.5) de la forma,*

$$E(X) = \int_0^{+\infty} P(X > x)\, dx \ = \int_0^{+\infty} P(X \geq x)\, dx.$$

3.2.2. Desigualdades

Si para $X \geq 0$ existe su esperanza, sea $\varepsilon > 0$ y escribamos (3.5) de la forma,

$$E(X) = \int_0^{+\infty} P(X \geq x)\, dx = \int_0^{\varepsilon} P(X \geq x)\, dx + \int_{\varepsilon}^{+\infty} P(X \geq x)\, dx.$$

Como la segunda integral es no negativa y la función $P(X \geq x)$ es decreciente,

$$E(X) \geq \int_0^{\varepsilon} P(X \geq x)\, dx \geq \int_0^{\varepsilon} P(X \geq \varepsilon)\, dx = \varepsilon P(X \geq \varepsilon),$$

y de aquí,

$$P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}. \tag{3.7}$$

Este resultado da lugar a dos conocidas desigualdades generales que proporcionan cotas superiores para la probabilidad de ciertos conjuntos. Estas desigualdades son válidas independientemente de cuál sea la distribución de probabilidad de la variable involucrada.

Desigualdad de Markov.- La primera de ellas se obtiene al sustituir en (3.7) X por $|X|^k$ y ε por ε^k,

$$P\left(|X| \geq \varepsilon\right) = P\left(|X|^k \geq \varepsilon^k\right) \leq \frac{1}{\varepsilon^k} E\left(|X|^k\right), \tag{3.8}$$

y es conocida como la *desigualdad de Markov*.

Desigualdad de Chebyshev.- Un caso especial de (3.8) se conoce como la *desigualdad de Chebyshev* y se obtiene para $k = 2$ y $X = X - E(X)$,

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} Var(X). \tag{3.9}$$

Un interesante resultado se deriva de esta última desigualdad.

Proposición 3.3 *Si $V(X) = 0$, entonces X es constante con probabilidad 1.*

Demostración.- Supongamos $E(X) = \mu$ y consideremos los conjuntos $A_n = \{|X - \mu| \geq 1/n\}$, aplicando (3.9)

$$P(A_n) = P\left(|X - \mu| \geq \frac{1}{n}\right) = 0, \ \forall n$$

y de aquí $P\left(\cup_n A_n\right) = 0$ y $P\left(\cap_n A_n^c\right) = 1$. Pero

$$\bigcap_{n\geq 1} A_n^c = \bigcap_{n\geq 1} \left\{|X - \mu| < \frac{1}{n}\right\} = \{X = \mu\},$$

luego $P(X = \mu) = 1$. ♠

Desigualdad de Jensen.- Si $g(X)$ es convexa sabemos que $\forall a, \exists \lambda_a$ tal que $g(x) \geq g(a) + \lambda_a(x - a), \forall x$. Si hacemos ahora $a = E(X)$,

$$g(X) \geq g\left(E(X)\right) + \lambda_a(X - E(X)),$$

y tomando esperanzas obtenemos la que se conoce como desigualdad de Jensen,

$$E(g(X)) \geq g(E(X)).$$

3.2.3. Momentos de algunas variables aleatorias conocidas

Binomial

Si $X \sim B(n, p)$,

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \\ &= \sum_{x=0}^{n} x \frac{n(n-1)\ldots(n-x+1)}{x!} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} \frac{(n-1)\ldots(n-x+1)}{(x-1)!} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-y-1} = np \end{aligned}$$

Para obtener $V(X)$, observemos que $E[(X(X-1)] = E(X^2) - E(X)$, y de aquí $V(X) = E[(X(X-1)] + E(X) - [E(X)]^2$. Aplicando un desarrollo análogo al anterior se obtiene $E[X(X-1)] = n(n-1)p^2$ y finalmente

$$V(X) = n(n-1)p^2 + np - n^2p^2 = np(1-p).$$

Poisson

Si $X \sim P(\lambda)$,

$$E(X) = \sum_{x\geq 0} x e^{-\lambda} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x-1\geq 0} \frac{\lambda^{x-1}}{(x-1)!} = \lambda.$$

Por otra parte,

$$E[X(X-1)] = \sum_{x\geq 0} x(x-1) e^{-\lambda} \frac{\lambda^x}{x!} = \lambda^2 e^{-\lambda} \sum_{x-2\geq 0} \frac{\lambda^{x-2}}{(x-2)!} = \lambda^2.$$

De aquí,

$$V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Uniforme

Si $X \sim U(0,1)$,

$$E(X) = \int_{-\infty}^{+\infty} x f_X \, dx = \int_0^1 x \, dx = \frac{1}{2}$$

Para obtener $V(X)$ utilizaremos la expresión alternativa, $V(X) = E(X^2) - [E(X)]^2$,

$$E(X^2) = \int_0^1 x^2 \, dx = \frac{1}{3},$$

y de aquí,

$$V(X) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.$$

Normal tipificada

Si $X \sim N(0,1)$, como su función de densidad es simétrica respecto del origen,

$$E(X^k) = \int_{-\infty}^{+\infty} x^k \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx = \begin{cases} 0, & si\ k = 2n+1 \\ m_{2n}, & si\ k = 2n. \end{cases}$$

Ello supone que $E(X) = 0$ y $V(X) = E(X^2)$. Para obtener los momentos de orden par,

$$m_{2n} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^{2n} e^{-\frac{x^2}{2}} \, dx = \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} x^{2n} e^{-\frac{x^2}{2}} \, dx.$$

Integrando por partes,

$$\int_0^{+\infty} x^{2n} e^{-\frac{x^2}{2}} \, dx =$$

$$-x^{2n-1} e^{-\frac{x^2}{2}} \Big|_0^{+\infty} + (2n-1) \int_0^{+\infty} x^{2n-2} e^{-\frac{x^2}{2}} \, dx = (2n-1) \int_0^{+\infty} x^{2n-2} e^{-\frac{x^2}{2}} \, dx,$$

lo que conduce a la fórmula de recurrencia $m_{2n} = (2n-1) m_{2n-2}$ y recurriendo sobre n,

$$m_{2n} = (2n-1)(2n-3)\cdots 1 =$$
$$= \frac{2n(2n-1)(2n-2)\cdots 2 \cdot 1}{2n(2n-2)\cdots 2} = \frac{(2n)!}{2^n n!}.$$

La varianza valdrá por tanto,

$$V(X) = E(X^2) = \frac{2!}{2 \cdot 1!} = 1.$$

Normal con parámetros μ y σ^2

Si $Z \sim N(0,1)$ es fácil comprobar que la variable definida mediante la expresión $X = \sigma Z + \mu$ es $N(\mu, \sigma^2)$. Teniendo en cuenta las propiedades de la esperanza y de la varianza,

$$E(X) = \sigma E(X) + \mu = \mu, \quad var(X) = \sigma^2 var(Z) = \sigma^2,$$

que son precisamente los parámetros de la distribución.

Cauchy

Hemos insistido a lo largo de la exposición que precede en que la $E(X)$ existe siempre que la variable X sea absolutamente integrable. Puede ocurrir que una variable aleatoria carezca, por este motivo, de momentos de cualquier orden. Es el caso de la distribución de Cauchy.

Decimos que X tiene una distribución de Cauchy si su función de de densidad viene dada por,

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \ -\infty < x < +\infty.$$

Observemos que

$$E(|X|) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{|x|}{1+x^2}\,dx = \frac{2}{\pi}\int_{0}^{+\infty} \frac{x}{1+x^2}\,dx = log(1+x^2)\big|_0^{+\infty} = +\infty.$$

Este resultado supone que no existe $E(X)$ ni ningún otro momento.

3.3. Esperanza de un vector aleatorio

Sea $X = (X_1, \ldots, X_k)$ un vector aletorio y sea g una función medible de $\mathcal{R}^k$ en $\mathcal{R}$. Si $|g(X)|$ es integrable respecto de la medida de probabilidad, se define la *esperanza de* $g(X)$ mediante

$$E(g(X)) = \int_\Omega g(X)\,dP.$$

Ya sabemos que la expresión final de la esperanza depende del carácter del vector aleatorio.

Vector aleatorio discreto.- Si D_X es el soporte del vector y f_X su función de cuantía conjunta, la esperanza se obtiene a partir de

$$E(g(X)) = \sum_{(x_1,\ldots,x_k)\in D_X} g(x_1,\ldots,x_k) f_X(x_1,\ldots,x_k).$$

Vector aleatorio continuo.- Si f_X es la función de densidad conjunta,

$$E(g(X)) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} g(x_1,\ldots,x_k) f_X(x_1,\ldots,x_k)\,dx_1 \ldots dx_k.$$

3.3.1. Momentos de un vector aleatorio

Como ya hemos visto en el caso de una variable aleatoria, determinadas formas de la función g dan lugar a los llamados *momentos* que se definen de forma análoga a como lo hicimos entonces. Las situaciones de mayor interés son ahora:

Momento conjunto.- El *momento conjunto de orden* $(n_1, \ldots, n_k)$ se obtiene, siempre que la esperanza exista, para

$$g(X_1, \ldots, X_k) = X_1^{n_1} \ldots X_k^{n_k},\ n_i \geq 0, \tag{3.10}$$

lo que da lugar a $E(X_1^{n_1} \ldots X_k^{n_k})$. Obsérvese que los momentos de orden k respecto del origen para cada componente pueden obtenerse como casos particulares de (3.10) haciendo $n_i = k$ y $n_j = 0,\ j \neq i$, pues entonces $E(X_1^{n_1} \ldots X_k^{n_k}) = E(X_i^k)$.

Momento conjunto central.- El *momento conjunto central de orden* $(n_1, \ldots, n_k)$ se obtienen, siempre que la esperanza exista, para

$$g(X_1, \ldots, X_k) = (X_1 - E(X_1))^{n_1} \ldots (X_k - E(X_k))^{n_k},\ n_i \geq 0,$$

Covarianza

De especial interés es el momento conjunto central obtenido para $n_i = 1$, $n_j = 1$ y $n_l = 0,\ l \neq (i, j)$. Recibe el nombre de *covarianza de* X_i *y* X_j y su expresión es,

$$cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))] = E(X_i X_j) - E(X_i)E(X_j).$$

La covarianza nos informa acerca del grado y tipo de dependencia existente entre ambas variables mediante su magnitud y signo, porque a diferencia de lo que ocurría con la varianza, la covarianza puede tener signo negativo.

Signo de la covarianza.- Para mejor comprender el significado del signo de la covarianza, consideremos dos sucesos A y B con probabilidades $P(A)$ y $P(B)$, respectivamente. Las correspondientes funciones características, $X = \mathbf{1}_A$ e $Y = \mathbf{1}_B$, son sendas variables aleatorias cuya covarianza vale

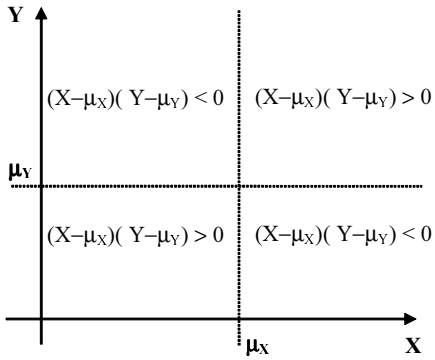
$$cov(X, Y) = E(XY) - E(X)E(Y) = P(A \cap B) - P(A)P(B),$$

puesto que $XY = \mathbf{1}_{A \cap B}$ y $E(X) = P(A)$ y $E(Y) = P(B)$. Observemos que

$$\begin{array}{lllll} cov(X,Y) > 0 & \text{si} & P(A \cap B) > P(A)P(B) & \Longrightarrow & P(A|B) > P(A) \text{ y } P(B|A) > P(B) \\ cov(X,Y) = 0 & \text{si} & P(A \cap B) = P(A)P(B) & \Longrightarrow & P(A|B) = P(A) \text{ y } P(B|A) = P(B) \\ cov(X,Y) < 0 & \text{si} & P(A \cap B) < P(A)P(B) & \Longrightarrow & P(A|B) < P(A) \text{ y } P(B|A) < P(B) \end{array}$$

Así pues, una covarianza positiva supone que el conocimiento previo de la ocurrencia de B aumenta la probabilidad de A, $P(A|B) > P(A)$, y análogamente para A. De aquí que hablemos de *dependencia positiva* entre los sucesos. En el caso contrario hablaremos de *dependencia negativa*. La covarianza vale cero cuando ambos sucesos son independientes.

Cuando las variables X e Y son variables aleatorias cualesquiera, ¿qué significado hemos de darle a la expresión *dependencia positiva* o *negativa*? En la gráfica hemos representado los cuatro cuadrantes en los que $\mu_X = E(X)$ y $\mu_Y = E(Y)$ dividen al plano, indicando el signo que el producto $(X - \mu_X)(Y - \mu_Y)$ tiene en cada uno de ellos.



Si los valores de X superiores (inferiores) a μ_X tienden a asociarse con los valores de Y superiores (inferiores) μ_Y, la covarianza será positiva y las variables tendrán una relación creciente. En caso contrario, superior con inferior o viceversa, la covarianza será negativa. En este contexto hemos de entender *tienden a asociarse* como *son más probables.*

En definitiva, una covarianza positiva (negativa) hemos de interpretarla como la existencia de una relación creciente (decreciente) entre las variables.

Magnitud de la covarianza.- La magnitud de la covarianza mide el grado de dependencia entre las variables. A mayor covarianza, medida en valor absoluto, mayor relación de dependencia entre las variables. Veamos un ejemplo.

Lanzamos tres monedas correctas y definimos el vector aleatorio (X, Y) con, X=$\{$*número de caras*$\}$ e Y=$\{$*número de cruces*$\}$. La tabla nos muestra la función de cuantía conjunta.

	$Y=0$	$Y=1$	$Y=2$	$Y=3$
$X=0$	0	0	0	1/8
$X=1$	0	0	3/8	0
$X=2$	0	3/8	0	0
$X=3$	1/8	0	0	0

A partir de ella calculamos la covarianza mediante

$$cov(X,Y) = E(XY) - E(X)E(Y) = \frac{12}{8} - \frac{3}{2} \times \frac{3}{2} = -\frac{3}{4}.$$

Si definimos ahora otro vector aleatorio (X, Z) con, X=$\{$*número de caras*$\}$ y Z=$\{$*número de cruces en los 2 primeros lanzamientos*$\}$, la función de cuantía conjunta es

	$Z=0$	$Z=1$	$Z=2$
$X=0$	0	0	1/8
$X=1$	0	2/8	1/8
$X=2$	1/8	2/8	0
$X=3$	1/8	0	0

La covarianza vale ahora

$$cov(X,Z) = E(XZ) - E(X)E(Z) = \frac{8}{8} - \frac{3}{2} \times 1 = -\frac{1}{2}.$$

Ambas covarianzas son negativas, indicando una relación decreciente entre ambos pares de variables. Así es, puesto que se trata del número de caras y cruces en un mismo conjunto de lanzamientos. A más caras, menos cruces. La primera covarianza es mayor en valor absoluto que la segunda, lo que supone un grado de dependencia mayor entre las componentes del primer vector. En efecto, existe una relación lineal entre el primer par de variables, $X + Y = 3$, que no existe para el segundo par, que sin duda están también relacionadas pero no linealmente.

Sin embargo, el hecho de que la covarianza, como ya ocurría con la varianza, sea sensible a los cambios de escala, hace que debamos ser precavidos a la hora de valorar el grado de dependencia entre dos variables aleatorias mediante la magnitud de la covarianza. Si las variables X e Y las sustituimos por $X' = aX$ e $Y' = bY$, fácilmente comprobaremos que

$$cov(X', Y') = a \times b \times cov(X, Y).$$

¿Significa ello que por el mero hecho de cambiar de escala la calidad de la relación entre X e Y cambia? Como la respuesta es obviamente no, es necesario introducir una nueva característica numérica ligada a las dos variables que mida su relación y sea invariante frente a los cambios de escala. Se trata del *coeficiente de correlación* que más adelante definiremos.

La covarianza nos permite también obtener la varianza asociada a la suma de variables aleatorias, como vemos a continuación.

Esperanza y varianza de una suma

La linealidad de la esperanza aplicada cuando $g(X) = X_1 + \cdots + X_n$, permite escribir

$$E(X_1 + \cdots + X_k) = \sum_{i=1}^{k} E(X_i).$$

Si las varianzas de las variables $X_1, \ldots, X_k$ existen, la varianza de $S = \sum_{i=1}^{k} a_i X_i$, donde los a_i son reales cualesquiera, existe y viene dada por

$$\begin{aligned} V(S) &= E[(S - E(S))^2] = \\ &= E\left[\left(\sum_{i=1}^{k} a_i (X_i - E(X_i))\right)^2\right] \\ &= \sum_{i=1}^{k} a_i^2 V(X_i) + 2a_i a_j \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} cov(X_i, X_j). \end{aligned} \tag{3.11}$$

Independencia y momentos de un vector aleatorio

Un resultado interesante acerca de los momentos conjuntos se recoge en la proposición que sigue.

Proposición 3.4 *Si las variables aleatorias $X_1, \ldots, X_k$ son independientes, entonces*

$$E(X_1^{n_1} \ldots X_k^{n_k}) = \prod_{i=1}^{k} E(X_i^{n_i}).$$

Demostración.- Si suponemos continuo el vector, la densidad conjunta puede factorizarse como producto de las marginales y

$$\begin{aligned}
&E(X_1^{n_1}\dots X_k^{n_k}) = \\
&\quad = \int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} x_1^{n_1}\dots x_k^{n_k} f_X(x_1,\dots,x_k)\,dx_1\dots dx_k = \\
&\quad = \int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} \left[\prod_{i=1}^{k} x_i^{n_i} f_i(x_i)\right] dx_1\dots dx_k = \\
&\quad = \prod_{i=1}^{k}\left[\int_{-\infty}^{+\infty} x_i^{n_i} f_i(x_i)\,dx_i\right] = \prod_{i=1}^{k} E(X_i^{n_i}).
\end{aligned}$$

El caso discreto se demuestra análogamente. ♠

Observación 3.3 *El anterior resultado admite una formulación más general. Si las funciones $g_i,\ i = 1,\dots,k$ son medibles, por el corolario (2.1) de la página 64 las $g_i(X_i)$ también son variables independientes y podemos escribir*

$$E\left[\prod_{i=1}^{k} g_i(X_i)\right] = \prod_{i=1}^{k} E[g_i(X_i)]. \tag{3.12}$$

Corolario 3.1 *Si las variables $X_1,\dots,X_k$ son independientes, entonces $cov(X_i,X_j) = 0,\ \forall i,j$.*

Corolario 3.2 *Si las variables $X_1,\dots,X_k$ son independientes y sus varianzas existen, la varianza de $S = \sum_{i=1}^{k} a_i X_i$, donde los a_i son reales cualesquiera, existe y viene dada por $V(S) = \sum_{i=1}^{k} a_i^2 V(X_i)$.*

Una aplicación de los anteriors resultados permite la obtención de la esperanza y la varianza de algunas conocidas variables de manera mucho más sencilla a como lo hicimos anteriormente. Veamos algunos ejemplos.

Ejemplo 3.1 (La Binomial y la Hipergeométrica como suma de Bernoullis) *Si recordamos las características de una Binomial fácilmente se comprueba que si $X \sim B(n,p)$ entonces $X = \sum_{i=1}^{n} X_i$ con $X_i \sim B(1,p)$ e independientes. Como $\forall i$,*

$$E(X_i) = p, \quad y \quad var(X_i) = p(1-p),$$

tendremos que

$$E(X) = \sum_{i=1}^{n} E(X_i) = nE(X_i) = np,$$

y

$$var(X) = \sum_{i=1}^{n} var(X_i) = np(1-p). \tag{3.13}$$

.

Si $X \sim H(n,N,r)$ también podemos escribir $X = \sum_{i=1}^{n} X_i$ con $X_i \sim B(1,r/N)$ pero a diferencia de la Binomial las variables X_i son ahora dependientes. Tendremos pues

$$E(X) = \sum_{i=1}^{n} E(X_i) = nE(X_i) = n\frac{r}{N},$$

y aplicando (3.11)

$$var(X) = \sum_{i=1}^{n} var(X_i) + 2\sum_{i=1}^{k}\sum_{j>i}^{k} cov(X_i, X_j) = n\ \frac{r}{N}\ \frac{N-r}{N} + n(n-1)cov(X_1, X_2), \quad (3.14)$$

puesto que todas las covarianzas son iguales.

Para obtener $cov(X_1, X_2)$,

$$\begin{aligned} cov(X_1, X_2) &= E(X_1X_2) - E(X_1)E(X_2) \\ &= P(X_1 = 1, X_2 = 1) - \left(\frac{r}{N}\right)^2 \\ &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) - \left(\frac{r}{N}\right)^2 \\ &= \frac{r-1}{N-1}\ \frac{r}{N} - \left(\frac{r}{N}\right)^2 \\ &= -\frac{r(N-r)}{N^2(N-1)}. \end{aligned}$$

Sustituyendo en (3.14)

$$var(X) = n\ \frac{r}{N}\ \frac{N-r}{N}\left(1 - \frac{n-1}{N-1}\right). \quad (3.15)$$

Es interesante comparar (3.15) con (3.13), para $p = r/N$. *Vemos que difieren en el último factor, factor que será muy próximo a la unidad si* $n \ll N$. *Es decir, si la* fracción de muestreo *(así se la denomina en Teoría de Muestras) es muy pequeña. Conviene recordar aquí lo que dijimos en la página 23 al deducir la relación entre ambas distribuciones.*

Ejemplo 3.2 (La Binomial Negativa como suma de Geométricas) *Si* $X \sim BN(r, p)$ *y recordamos su definición, podemos expresarla como* $X = \sum_{i=1}^{r} X_i$, *donde cada* $X_i \sim BN(1, p)$ *e independiente de las demás y representa las pruebas Bernoulli necesarias después del* $(i-1)$*-ésimo éxito para alcanzar el* i*-ésimo.*

Obtendremos primero la esperanza y la varianza de una variable Geométrica de parámetro p.

$$E(X_i) = \sum_{n\geq 0} np(1-p)^n = p\sum_{n\geq 1} n(1-p)^n = \frac{1-p}{p},\ \forall i.$$

Para calcular la varianza necesitamos conocer $E(X_i^2)$,

$$E(X_i^2) = \sum_{n\geq 0} n^2p(1-p)^n = p\sum_{n\geq 1} n^2(1-p)^n = \frac{1-p}{p^2}(2-p),\ \forall i.$$

y de aquí,

$$var(X) = \frac{1-p}{p^2}(2-p) - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}.$$

La esperanza y la varianza de la Binomial Negativa de parámetros r *y* p, *valdrán*

$$E(X) = \sum_{i=1}^{r} E(X_i) = \frac{r(1-p)}{p}, \qquad var(X) = \sum_{i=1}^{r} var(X_i) = \frac{r(1-p)}{p^2}.$$

3.3.2. Desigualdades

Teorema 3.1 (Desigualdad de Cauchy-Schwarz) *Sean X e Y variables aleatorias con varianzas finitas. Entonces $cov(X,Y)$ existe y se verifica*

$$[E(XY)]^2 \leq E(X^2)E(Y^2),$$

verificándose la igualdad si y solo si *existe un real α tal que $P(\alpha X + Y = 0) = 1$.*

Demostración.- Para cualesquiera números reales a y b se verifica

$$|ab| \leq \frac{a^2 + b^2}{2},$$

lo que significa que $E(XY) < \infty$ si $E(X^2) < \infty$ y $E(Y^2) < \infty$. Por otra parte, para cualquier real α, se tiene

$$E[(\alpha X + Y)^2] = \alpha^2 E(X^2) + 2\alpha E(XY) + E(Y^2) \geq 0,$$

lo que supone que la ecuación de segundo grado tiene a lo sumo una raíz y su discriminante será no positivo. Es decir,

$$[E(XY)]^2 \leq E(X^2)E(Y^2).$$

Si se diera la igualdad, la ecuación tendría una raíz doble, α_0, y $E[(\alpha_0 X + Y)^2] = 0$. Tratándose de una función no negativa, esto implica que $P(\alpha_0 X + Y = 0) = 1$. ♠

El coeficiente de correlación

El coeficiente de correlación entre dos componentes cualesquiera de un vector aleatorio, X y Y, se define como la covarianza de dichas variables tipificadas[1].

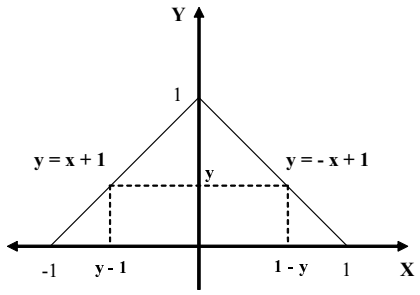
$$\rho_{XY} = cov(X_t, Y_t) = E\left[\left(\frac{X - E(X)}{\sigma_X}\right)\left(\frac{Y - E(Y)}{\sigma_Y}\right)\right] = \frac{cov(X,Y)}{\sigma_X \sigma_Y}.$$

De la desigualdad de Cauchy-Schwarz se desprende que $\rho_{XY}^2 \leq 1$ y en particular que $-1 \leq \rho_{XY} \leq 1$. Recordemos que cuando en Cauchy-Schwarz se daba la igualdad X e Y estaban relacionadas linealmente con probabilidad 1, $P(\alpha_0 X + Y = 0) = 1$, pero por otra parte dicha igualdad implica $\rho_{XY}^2 = 1$. El valor 0 es otro valor de interés para ρ_{XY}. Si $\rho_{XY} = 0$ decimos que las variables están *incorreladas* y además $cov(X,Y) = 0$. Hay entonces una ausencia total de relación lineal entre ambas variables. Podemos decir que el valor absoluto del coeficiente de correlación es una medida del grado de linealidad entre las variables medido, de menor a mayor, en una escala de 0 a 1.

Acabamos de ver que $\rho_{XY} = 0 \rightarrow cov(X,Y) = 0$. La implicación contraria también es cierta, lógicamente. Como la independencia entre dos variables implicaba que su covarianza era nula, deducimos que *independencia* $\rightarrow$ *incorrelación*, pero ahora la implicación contraria no es cierta como puede comprobarse en el siguiente ejemplo.

Ejemplo 3.3 *Consideremos el vector (X,Y) uniformemente distribuido en el recinto triangular con vértices en (-1,0), (1,0) y (0,1).*

[1] Una variable tipificada es la que resulta de transformar la original restándole la media y dividiendo por la desviación típica, $X_t = (X - \mu_X)/\sigma_X$. Como consecuencia de esta transformación $E(X_t) = 0$ y $var(X_t) = 1$.



Se verifica

$$E(XY) = \int_0^1 y \left[\int_{y-1}^{1-y} x dx \right] dy = 0$$

y

$$E(X) = \int_0^1 \left[\int_{y-1}^{1-y} x dx \right] dy = 0$$

y por tanto $cov(X, Y) = 0$ *pero* X *e* Y *no son independientes.*

3.3.3. Covarianza en algunos vectores aleatorios conocidos

Multinomial

Si $X \sim M(n; p_1, p_2, \ldots, p_k)$, sabemos que cada componente $X_i \sim B(n, p_i)$, $i = 1, \ldots, k$, y puede por tanto expresarse como suma de n Bernoullis de parámetro p_i. La covarianza entre dos cualesquiera puede expresarse como,

$$cov(X_i, X_j) = cov\left(\sum_{k=1}^{n} X_{ik}, \sum_{l=1}^{n} X_{jl} \right).$$

Se demuestra fácilmente que

$$cov\left(\sum_{k=1}^{n} X_{ik}, \sum_{l=1}^{n} X_{jl} \right) = \sum_{k=1}^{n} \sum_{l=1}^{n} cov(X_{ik}, X_{jl}).$$

Para calcular $cov(X_{ik}, X_{jl})$ recordemos que

$$X_{ik} = \begin{cases} 1, & \text{si en la prueba } k \text{ ocurre } A_i; \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

y

$$X_{jl} = \begin{cases} 1, & \text{si en la prueba } l \text{ ocurre } A_j; \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

En consecuencia, $cov(X_{ik}, X_{jl}) = 0$ si $k \neq l$ porque las pruebas de Bernoulli son independientes y

$$cov(X_{ik}, X_{jk}) = E(X_{ik}X_{jk}) - E(X_{ik})E(X_{jk}) = 0 - p_i p_j,$$

donde $E(X_{ik}X_{jk}) = 0$ porque en una misma prueba, la k-ésima, no puede ocurrir simultáneamente A_i y A_j. En definitiva,

$$cov(X_i, X_j) = \sum_{k=1}^{n} cov(X_{ik}, X_{jk}) = -np_i p_j.$$

El coeficiente de correlación entre ambas variables vale,

$$\rho_{ij} = -\frac{np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_j(1-p_j)}} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}.$$

El valor negativo de la covarianza y del coeficiente de correlación se explica por el hecho de que siendo el número total de pruebas fijo, n, a mayor número de ocurrencias de A_i, menor número de ocurrencias de A_j.

Normal bivariante

Si (X, Y) tiene una distribución Normal bivariante de parámetros μ_X, μ_y, σ_x, σ_y y ρ, ya vimos en la página 45 que $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$. Para simplificar la obtención de $cov(X, Y)$ podemos hacer uso de la invarianza por traslación de la covarianza (se trata de una propiedad de fácil demostración que ya comprobamos para la varianza). De acuerdo con ella, $cov(X, Y) = cov(X - \mu_x, Y - \mu_y)$ y podemos suponer que X e Y tienen media 0, lo que permite expresar la covarianza como $cov(X, Y) = E(XY)$. Procedamos a su cálculo.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{XY}(x,y)dxdy \\ &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xye^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x}{\sigma_x}\right)^2-2\rho\left(\frac{x}{\sigma_x}\right)\left(\frac{y}{\sigma_y}\right)+\left(\frac{y}{\sigma_y}\right)^2\right\}}dxdy \\ &= \int_{-\infty}^{+\infty}\frac{y}{\sigma_y\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2}\left[\int_{-\infty}^{+\infty}\frac{x}{\sigma_x\sqrt{2\pi(1-\rho^2)}}e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x}{\sigma_x}-\rho\frac{y}{\sigma_y}\right)^2}dx\right]dy. \end{aligned} \tag{3.16}$$

La integral interior en (3.16) es la esperanza de una $N(\rho\sigma_x y/\sigma_y, \sigma_x^2(1-\rho^2))$ y su valor será por tanto $\rho\sigma_x y/\sigma_y$. Sustituyendo en (3.16)

$$E(XY) = \frac{\rho\sigma_x}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left(\frac{y}{\sigma_y}\right)^2 e^{-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2}dy = \frac{\rho\sigma_x\sigma_y}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} z^2e^{-\frac{1}{2}z^2}dy = \rho\sigma_x\sigma_y.$$

El coeficiente de correlación entre ambas variables es

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_x\sigma_y} = \rho.$$

Todos los parámetros de la Normal bivariante adquieren ahora significado.

3.3.4. La distribución Normal multivariante

La expresión de la densidad de la Normal bivariante que dábamos en la página 45 admite una forma alternativa más compacta a partir de lo que se conoce como la matriz de covarianzas de la distribución, $\mathbf{\Sigma}$, una matriz 2×2 cuyos elementos son las varianzas y las covarianzas del vector (X_1, X_2),

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

La nueva expresión de la densidad es

$$f(x_1, x_2) = \frac{|\mathbf{\Sigma}|^{-\frac{1}{2}}}{2\pi}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad (x_1, x_2) \in \mathcal{R}^2,$$

donde $|\mathbf{\Sigma}|$ es el determinante de $\mathbf{\Sigma}$, cuyo valor es

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2(1-\rho_{12}^2),$$

$\mathbf{x}' = (x_1\ x_2)$ es el vector de variables y $\boldsymbol{\mu}' = (\mu_1\ \mu_2)$ es el vector de medias.

Si el vector tiene n componentes, $X_1, X_2, \ldots, X_n$, la extensión a n dimensiones de la expresión de la densidad es ahora inmediata con esta notación. La matriz de covarianzas es una matriz $n \times n$ con componentes

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1(n-1)} & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2(n-1)} & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{1(n-1)} & \sigma_{2(n-1)} & \cdots & \sigma_{n-1}^2 & \sigma_{(n-1)n} \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{(n-1)n} & \sigma_n^2 \end{pmatrix},$$

el vector de medias es $\boldsymbol{\mu}' = (\mu_1\ \mu_2\ \ldots\ \mu_n)$ y la densidad tiene por expresión

$$f(x_1, x_2, \ldots, x_n) = \frac{|\mathbf{\Sigma}|^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \qquad (x_1, x_2, \ldots, x_n) \in \mathcal{R}^n. \tag{3.17}$$

Cuando las componentes del vector son independientes, las covarianzas son todas nulas y $\mathbf{\Sigma}$ es una matriz diagonal cuyos elementos son las varianzas de cada componente, por tanto

$$|\mathbf{\Sigma}|^{-1} = \frac{1}{\sigma_1^2\sigma_2^2\cdots\sigma_n^2}.$$

Además, la forma cuadrática que aparece en el exponente de (3.17) se simplifica y la densidad adquiere la forma

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\prod_{i=1}^n (2\pi\sigma_i^2)^{\frac{1}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2} = \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}\right],$$

que no es más que el producto de las densidades de cada una de las componentes.

Añadamos por último que la matriz $\mathbf{\Sigma}$ es siempre definida positiva y lo es también la forma cuadrática que aparece en el exponente de (3.17).

3.3.5. Muestra aleatoria: media y varianza muestrales

En este apartado introduciremos el concepto de muestra aleatoria y algunas variables aleatorias con ella asociadas, particularmente la media y la varianza muestrales. Cuando la muestra proceda de una distribución Normal, las distribuciones de probabilidad relacionadas con dichas variables constituyen el eje sobre el que se basa la Inferencia Estadística clásica.

Definición 3.2 *Una muestra aleatoria de tamaño* n *es un vector aleatorio* $X_1, X_2, \ldots, X_n$*, cuyas componentes son* n *variables aleatorias independientes e idénticamente distribuidas* (i.i.d.).

A partir de la muestra aleatoria podemos definir la media y la varianza muestrales, $\overline{X}_n$ y S_n^2, mediante las expresiones

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{y} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

respectivamente.

Si las componentes de la muestra aleatoria poseen momentos de segundo orden, y su media y varianza comunes son μ y σ, respectivamente, se verifica que

$$E\left(\overline{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \tag{3.18}$$

y

$$var\left(\overline{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n}.$$

Por lo que respecta a la varianza muestral, observemos en primer lugar que

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X}_n)^2 \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i-\mu)^2 + \sum_{i=1}^{n}(\overline{X}_n-\mu)^2 - 2(\overline{X}_n-\mu)\sum_{i=1}^{n}(X_i-\mu)\right\} \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i-\mu)^2 + n(\overline{X}_n-\mu)^2 - 2(\overline{X}_n-\mu)n(\overline{X}_n-\mu)\right\} \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}_n-\mu)^2\right\}. \end{aligned}$$

Tomando ahora esperanzas,

$$\begin{aligned} E(S_n^2) &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - nE\left[(\overline{X}_n-\mu)^2\right]\right\} \\ &= \frac{1}{n-1}\left(n\sigma^2 - n\times var(\overline{X}_n)\right) \\ &= \sigma^2. \end{aligned} \tag{3.19}$$

La $var(S_n^2)$ existe si las componentes de la muestra aleatoria poseen momentos de cuarto orden, si es así, la expresión de $var(S_n^2)$ viene dada por

$$var(S_n^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

con

$$\mu_4 = E\left[(X_i-\mu)^4\right] \qquad \text{y} \qquad \sigma^4 = [\sigma^2]^2,$$

cuya deducción dejamos al cuidado del lector.

Observación 3.4 *La Inferencia Estadística proporciona una serie de técnicas que permiten conocer el comportamiento probabilístico de un fenómeno aleatorio a partir de la observación parcial del mismo. Por ejemplo, la estatura de una población de individuos tiene un comportamiento aleatorio que podemos deducir con un cierto margen de error (probabilístico) de la observación de una muestra de alturas de* n *individuos de esa población. Si postulamos la hipótesis de que dicha altura, X, se distribuye $N(\mu,\sigma^2)$, podemos estimar los valores de μ y σ^2 a partir de sus homólogos muestrales, $\overline{X}_n$ y S_n^2. Esta elección se basa, entre otras, en una propiedad conocida como* insesgadez*, que es la que representan las igualdades (3.18) y (3.19). A saber, que las esperanzas de los estimadores coinciden con el parámetro estimado.*

3.4. Esperanza condicionada

Sea (X, Y) un vector aleatorio definido sobre el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ y denotemos por $P_{X|Y=y}$ la distribución de probabilidad de X condicionada a $Y = y$. Si g es una función medible definida de $(\mathcal{R}, \beta)$ en $(\mathcal{R}, \beta)$, tal que $E(g(X))$ existe, la *esperanza condicionada* de $g(X)$ dado Y, $E[g(X)|Y]$, es una variable aleatoria que para $Y = y$ toma el valor

$$E[g(X)|y] = \int_{\Omega} g(X)\, dP_{X|Y=y}. \tag{3.20}$$

La forma de (3.20) depende de las características de la distribución del vector (X, Y).

(X, Y) **es discreto.-** Ello supone que el soporte de la distribución, D, es numerable y

$$E[g(X)|y] = \sum_{x \in D_y} g(x) P(X = x|Y = y) = \sum_{x \in D_y} g(x) f_{X|Y}(x|y),$$

donde $D_y = \{x; (x, y) \in D\}$ es la sección de D mediante y.

(X, Y) **es continuo.-** Entonces

$$E[g(X)|y] = \int_{\mathcal{R}} g(x) f_{X|Y}(x|y) dx.$$

Una definición similar puede darse para $E[h(Y)|X]$ siempre que $E[h(Y)]$ exista.

Ejemplo 3.4 *Sean X e Y variables aleatorias independientes ambas con distribución $B(n, p)$. La distribución de $X + Y$ se obtiene fácilmente a partir de*

$$\begin{aligned}
f_{X+Y}(m) &= P\left(\bigcup_{k=0}^{m} \{X = k, Y = m - k\}\right) \\
&= \sum_{k=0}^{m} P(X = k, Y = m - k) \\
&= \sum_{k=0}^{m} P(X = k) P(Y = m - k) \\
&= \sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k} \binom{n}{m-k} p^{m-k} (1-p)^{n-(m-k)} \\
&= p^m (1-p)^{2n-m} \sum_{k=0}^{m} \binom{n}{k} \binom{n}{m-k} \\
&= \binom{2n}{m} p^m (1-p)^{2n-m},
\end{aligned}$$

de donde $X + Y \sim B(2n, p)$.

La distribución condicionada de $Y|X+Y=m$ *es*

$$\begin{aligned} P(Y=k|X+Y=m) &= \frac{P(Y=k, X+Y=m)}{P(X+Y=m)} \\ &= \frac{P(Y=k, X=m-k)}{P(X+Y=m)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{m-k}p^{m-k}(1-p)^{n-(m-k)}}{\binom{2n}{m}p^m(1-p)^{2n-m}} \\ &= \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}, \end{aligned}$$

es decir, $Y|X+Y=m \sim H(m,2n,n)$*. La* $E(Y|X+Y=m)$ *valdrá*

$$E(Y|X+Y=m) = \frac{nm}{2n} = \frac{m}{2}.$$

La esperanza condicionada posse todas las propiedades inherentes al concepto de esperanza anteriormente estudiadas. A título de ejemplo podemos recordar,

PEC1) La esperanza condicionada es un operador lineal,

$$E[(ag_1(X)+bg_2(X))|Y] = aE[g_1(X)|Y] + bE[g_2(X)|Y].$$

En particular,

$$E[(aX+b)|Y] = aE(X|Y) + b.$$

PEC2) $P(a \le X \le b) = 1 \Longrightarrow a \le E(X|Y) \le b.$

PEC3) $P(g_1(X) \le g_2(X)) = 1 \Longrightarrow E[g_1(X)|Y] \le E[g_2(X)|Y].$

PEC4) $E(c|Y) = c$, para c constante.

Momentos de todo tipo de la distribución condicionada se definen de forma análoga a como hicimos en el caso de la distribución absoluta y gozan de las mismas propiedades. Existen, no obstante, nuevas propiedades derivadas de las peculiares características de este tipo de distribuciones y de que $E[g(X)|Y)]$, en tanto que función de Y, es una variable aleatoria. Veámoslas.

Proposición 3.5 *Si* $E(g(X))$ *existe, entonces*

$$E\left(E[g(X)|Y)]\right) = E(g(X)). \tag{3.21}$$

Demostración.- Supongamos el caso continuo.

$$\begin{aligned} E\left(E[g(X)|Y)]\right) &= \int_{\mathcal{R}} E[g(X)|y)]f_y(y)dy \\ &= \int_{\mathcal{R}} \left[\int_{\mathcal{R}} g(x)f_{X|Y}(x|y)dx\right] f_y(y)dy \\ &= \int_{\mathcal{R}} \left[\int_{\mathcal{R}} g(x)\frac{f_{XY}(x,y)}{f_y(y)}dx\right] f_y(y)dy \\ &= \int_{\mathcal{R}} g(x)\left[\int_{\mathcal{R}} f_{XY}(x,y)dy\right] dx \\ &= \int_{\mathcal{R}} g(x)f_X(x)dx = E(g(X)). \end{aligned}$$

La demostración se haría de forma análoga para el caso discreto ♠

Ejemplo 3.5 *Consideremos el vector* (X,Y) *con densidad conjunta*

$$f_{XY}(x,y) = \begin{cases} \dfrac{1}{x}, & 0 < y \leq x \leq 1 \\ \\ 0, & \textit{en el resto.} \end{cases}$$

Fácilmente obtendremos que $X \sim U(0,1)$ *y que* $Y|X = x \sim U(0,x)$*. Aplicando el resultado anterior podemos calcular* $E(Y)$*,*

$$E(Y) = \int_0^1 E(Y|x)f_X(x)dx = \int_0^1 \frac{1}{2}xdx = \frac{1}{4}.$$

Ejemplo 3.6 *Un trabajador está encargado del correcto funcionamiento de n máquinas situadas en linea recta y distantes una de otra l metros. El trabajador debe repararlas cuando se averían, cosa que sucede con igual probabilidad para todas ellas e independientemente de una a otra. El operario puede seguir dos estrategias:*

1. *acudir a reparar la máquina estropeada y permanecer en ella hasta que otra máquina se avería, desplazándose entonces hacia ella, o*

2. *situarse en el punto medio de la linea de máquinas y desde allí acudir a la averiada, regresando nuevamente a dicho punto cuando la avería está resuelta.*

Si X es la distancia que recorre el trabajador entre dos averías consecutivas, ¿cuál de ambas estrategias le conviene más para andar menos?

Se trata, en cualquiera de las dos estrategias, de obtener la $E(X)$ *y elegir aquella que la proporciona menor. Designaremos por* $E_i(X)$ *la esperanza obtenida bajo la estrategia* $i = 1, 2$*.*

> ***Estrategia 1.-*** *Sea* A_k *el suceso,* el operario se encuentra en la máquina *k. Para obtener* $E_1(X)$ *recurriremos a la propiedad anterior, pero utilizando como distribución condicionada la que se deriva de condicionar respecto del suceso* A_k*. Tendremos* $E_1(X) = E(E(X|A_k))$*.*

Para obtener $E(X|A_k)$ *tengamos en cuenta que si* i *es la próxima máquina averiada,* $P(A_i) = 1/n$, $\forall i$ *y el camino a recorrer será*

$$X|A_k = \begin{cases} (k-i)l, & si\ i \leq k, \\ \\ (i-k)l, & si\ i > k. \end{cases}$$

Así pues,

$$E(X|A_k) = \frac{1}{n}\left(\sum_{i=1}^{k}(k-i)l + \sum_{i=k+1}^{n}(i-k)l\right) = \frac{l}{2n}[2k^2 - 2(n+1)k + n(n+1)].$$

Utilizando

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6},$$

obtenemos para $E_1(X)$

$$E_1(X) = E(E(X|A_k)) = \frac{1}{n}\sum_{k} E(X|A_k) = \frac{l(n^2-1)}{3n}$$

Estrategia 2.- *Para facilitar los cálculos supongamos que* n *es impar, lo que supone que hay una máquina situada en el punto medio de la linea, la* $\frac{n+1}{2}$*-ésima. Si la próxima máquina averiada es la* i *la distancia a recorrer será,*

$$X = \begin{cases} 2(\frac{n+1}{2} - i)l, & si\ i \leq \frac{n+1}{2}, \\ \\ 2(i - \frac{n+1}{2})l, & si\ i > \frac{n+1}{2}, \end{cases}$$

donde el 2 *se justifica porque el operario en esta segunda estrategia regresa siempre al punto medio de la linea de máquinas. La esperanza viene dada por,*

$$E_2(X) = \frac{2}{n}\sum_{i=1}^{n}\left|\frac{n+1}{2} - i\right| l = \frac{l(n-1)}{2}.$$

Como $E_1(X)/E_2(X) = 2(n+1)/3n \leq 1$ *si* $n \geq 2$*, podemos deducir que la primera estrategia es mejor, salvo que hubiera una sola máquina.*

El ejemplo anterior ilustra la utilidad de (3.21) en la obtención de la esperanza absoluta. El ejemplo que sigue nos muestra como, en ocasiones, el rodeo que (3.21) pueda suponer es sólo aparente porque, en ocasiones, la obtención directa de $E(X)$ es mucho más laboriosa.

Ejemplo 3.7 *De una urna con* n *bolas blancas y* m *bolas negras llevamos a cabo extracciones sucesivas sin reemplazamiento. Si queremos conocer el número esperado de bolas negras extraídas antes de la primera blanca deberemos conocer la distribución de la correspondiente variable,* X*.*

La función de probabilidad de X *vale*

$$f_X(k) = P(X = k) = \frac{\binom{m}{k}}{\binom{m+n}{k}} \times \frac{n}{m+n-k}, \quad k = 0, 1, \ldots, m, \tag{3.22}$$

donde el primer factor establece la probabilidad de que las k primeras bolas sean negras, y el segundo la de que la $k+1$-ésima bola sea blanca. Un sencillo cálculo con los números combinatorios permite escribir (3.22) de esta otra forma

$$f_X(k) = \frac{1}{\binom{m+n}{n}} \times \binom{m+n-1-k}{n-1},$$

más útil a la hora de comprobar que (3.22) es una función de probabilidad. En efecto,

$$\begin{aligned}
\sum_{k=0}^{m} f_X(k) &= \sum_{k=0}^{m} \frac{1}{\binom{m+n}{n}} \times \binom{m+n-1-k}{n-1} \\
&= \frac{1}{\binom{m+n}{n}} \sum_{k=0}^{m} \binom{m+n-1-k}{n-1} \quad \textit{cambio } [m-k=j] \\
&= \frac{1}{\binom{m+n}{n}} \sum_{j=0}^{m} \binom{n-1+j}{n-1} \\
&= \frac{\binom{m+n}{n}}{\binom{m+n}{n}} = 1.
\end{aligned}$$

El valor esperado valdrá,

$$\begin{aligned}
E(X) &= \sum_{k=0}^{m} k f_X(k) \\
&= \sum_{k=1}^{m} k \times \frac{\binom{m}{k}}{\binom{m+n}{k}} \times \frac{n}{m+n-k} \\
&= \frac{m}{m+n} \sum_{k=1}^{m} k \times \frac{\binom{m-1}{k-1}}{\binom{m+n-1}{k-1}} \times \frac{n}{m-1+n-(k-1)} \\
&= \frac{m}{m+n} \sum_{j=0}^{m-1} (j+1) f_{X_{m-1}}(j),
\end{aligned}$$

donde X_{m-1} es la variable asociada al mismo experimento cuando la urna contiene $m-1$ bolas negras y n bolas blancas. Obtenemos en definitiva la fórmula de recurrencia

$$E_m(X) = \frac{m}{m+n}(1 + E_{m-1}(X)), \tag{3.23}$$

en la que el significado de los subíndices es obvio. Observemos que si $m = 0$, $E_0(X) = 0$ *y a partir de aquí*

$$\begin{aligned} E_1(X) &= \frac{1}{n+1}(1+0) = \frac{1}{n+1} \\ E_2(X) &= \frac{2}{n+2}\left(1+\frac{1}{n+1}\right) = \frac{2}{n+1} \\ E_3(X) &= \frac{3}{n+3}\left(1+\frac{2}{n+1}\right) = \frac{3}{n+1} \\ \cdots &\quad \cdots \\ E_m(X) &= \frac{m}{n+m}\left(1+\frac{m-1}{n+1}\right) = \frac{m}{n+1}. \end{aligned}$$

La expresión (3.23) puede obtenerse más fácilmente si utilizamos la variable auxiliar Y *definida mediante*

$$Y = \begin{cases} 1, & \textit{si la primera bola es blanca;} \\ 0, & \textit{si la primera bola es negra.} \end{cases}$$

Podremos escribir

$$E_m(X) = E[E_m(X|Y)] = E_m(X|Y=1)P(Y=1) + E_m(X|Y=0)P(Y=0),$$

pero $E_m(X|Y=1) = 0$ *y* $E_m(X|Y=0) = 1 + E_{m-1}(X)$. *Sustituyendo,*

$$E_m(X) = \frac{m}{m+n}(1 + E_{m-1}(X)),$$

que coincide con (3.23).

Ejemplo 3.8 *El tiempo de descarga de un barco que transporta cereal se distribuye uniformemente entre* a *y* b *horas,* $T \sim U(a,b)$, *siempre que dicha descarga se lleve a cabo sin interrupciones. La grúa que efectúa la descarga es poco fiable y propensa a las averías. Éstas ocurren según una distribución de Poisson con media* λ *averías por hora. Cuando la grúa está estropeada se requieren* d *horas para arreglarla. Con todas estas circunstancias, queremos conocer el tiempo real medio de descarga del barco.*

Si designamos por T_R *el tiempo real de descarga del barco, lo que se nos pide es* $E(T_R)$. *Observemos que sobre dicho tiempo influyen, tanto la variable* T, *que nos proporciona el tiempo de descarga sin tener en cuenta las posibles interrupciones, como las horas que haya que añadir debido a la averías de la grúa durante el período* T, *averías cuyo número en dicho lapso de tiempo será* $N_T \sim Po(\lambda T)$.

Sabemos que $E(T_R) = E[E(T_R|T)]$. *Para calcular la esperanza condicionada hemos de tener en cuenta las posibles averías, lo que equivale a decir que hemos de condicionar a los posibles valores de* N_T. *Así,*

$$\begin{aligned} E(T_R|T) &= E\left\{E\left[E(T_R|T)|N_T\right]\right\} \\ &= \sum_{n\geq 0} E(T_R|T, N_T = n)P(N_T = n) \\ &= \sum_{n\geq 0} (T + nd)P(N_T = n) \\ &= T\sum_{n\geq 0} P(N_T = n) + d\sum_{n\geq 0} nP(N_T = n) \\ &= T(1 + d\lambda). \end{aligned}$$

Por último

$$E(T_R) = E[E(T_R|T)] = (1 + d\lambda)E(T) = \frac{(a+b)(1+d\lambda)}{2}.$$

También es posible relacionar la varianza absoluta con la varianza condicionada, aunque la expresión no es tan directa como la obtenida para la esperanza.

Proposición 3.6 *Si $E(X^2)$ existe, entonces*

$$var(X) = E(var[X|Y]) + var(E[X|Y]). \tag{3.24}$$

Demostración.- Haciendo del anterior resultado,

$$\begin{aligned}
var(X) &= E\left[(X - E(X))^2\right] = E\left\{E\left[(X - E(X))^2|Y\right]\right\} \\
&= E\left\{E\left[\left(X^2 + (E(X))^2 - 2XE(X)\right)|Y\right]\right\} \\
&= E\left\{E[X^2|Y] + E(X)^2 - 2E(X)E[X|Y]\right\} \\
&= E\left\{E[X^2|Y] - (E[X|Y])^2 + (E[X|Y])^2 + E(X)^2 - 2E(X)E[X|Y]\right\} \\
&= E\left\{var[X|Y] + (E[X|Y] - E(X))^2\right\} \\
&= E\left\{var[X|Y]\right\} + E\left\{(E[X|Y] - E(X))^2\right\} \\
&= E(var[X|Y]) + var(E[X|Y]). \qquad \spadesuit
\end{aligned}$$

Corolario 3.3 *Si $E(X^2)$ existe, por la propiedad anterior*

$$var(X) \geq var(E[X|Y]).$$

Si se verifica la igualdad, de (3.24) deducimos que $E(var[X|Y]) = 0$, y como $var[X|Y] \geq 0$, tendremos que $P(var[X|Y] = 0) = 1$, es decir

$$P\left(E\left\{(X - E[X|Y])^2|Y\right\} = 0\right) = 1,$$

y como $(X - E[X|Y])^2 \geq 0$, aplicando de nuevo el anterior razonamiento concluiremos que

$$P\left(X = E[X|Y]\right) = 1,$$

lo que supone que, con probabilidad 1, X es una función de Y puesto que $E[X|Y]$ lo es.

3.4.1. El principio de los mínimos cuadrados

Supongamos que entre las variables aleatorias X e Y existe una relación funcional que deseamos conocer. Ante la dificultad de poder hacerlo vamos a tratar de encontrar la función $h(X)$ que mejor aproxime dicha relación. El interés en semejante función es poder predecir el valor que tomará Y a partir del valor que ha tomado X.

Si $E(Y^2)$ y $E(h(X)^2)$ existen, el *principio de los mínimos cuadrados* es uno de los posibles criterios para encontrar $h(X)$. Consiste en elegir aquella $h(X)$ que minimice la cantidad

$E\left\{(Y-h(X))^2\right\}$, que no es más que la esperanza del error que cometeremos al sustituir el verdadero valor de Y por su estimación, $h(X)$. Si (X,Y) es un vector aleatorio continuo, tenemos

$$\begin{aligned} E\left\{(Y-h(X))^2\right\} &= \int_{\mathcal{R}^2} (y-h(x))^2 f_{XY}(x,y)dxdy \\ &= \int_{\mathcal{R}^2} (y-h(x))^2 f_{Y|X}(x,y)f_X(x)dydx \\ &= \int_{\mathcal{R}} \left[\int_{\mathcal{R}} (y-h(x))^2 f_{Y|X}(x,y)dy\right] f_X(x)dx. \end{aligned}$$

Que sea mínima esta expresión supone que lo es la integral interior, pero de acuerdo con una propiedad de la varianza (PV4 de la página 71) el valor que minimiza dicha integral es precisamente $h(X) = E(X|Y)$. Para el caso discreto obtendríamos un resultado análogo.

Definición 3.3 (Regresión de Y sobre X) *La relación $y = E[Y|x]$ se conoce como la regresión de Y sobre X. Análogamente $x = E[X|y]$ se conoce como la regresión de X sobre Y.*

Recta de regresión

En ocasiones, fundamentalmente por motivos de sencillez, se está interesado en aproximar la relación entre X e Y mediante una línea recta. En esta situación, y haciendo uso del principio de los mínimos cuadrados, elegiremos los parámetros de la recta de forma que

$$L(a,b) = E\left\{(Y-aX-b)^2\right\}$$

sea mínimo.

La obtención de a y b se reduce a un problema de máximos y mínimos y basta igualar a 0 las derivadas parciales $\partial L/\partial a$ y $\partial L/\partial b$. Si lo hacemos obtendremos,

$$a = \frac{cov(X,Y)}{var(X)}, \qquad b = E(Y) - aE(X).$$

La ecuación de la que se conoce como *recta de regresión* de Y sobre X tendrá por expresión,

$$Y - E(Y) = \frac{cov(X,Y)}{var(X)}(X - E(X)).$$

Por simetría, la *recta de regresión* de X sobre Y tendrá por expresión,

$$X - E(X) = \frac{cov(X,Y)}{var(Y)}(Y - E(Y)).$$

En las anteriores expresiones, las cantidades

$$\gamma_{X|Y} = \frac{cov(X,Y)}{var(Y)} \quad y \quad \gamma_{Y|X} = \frac{cov(X,Y)}{var(X)},$$

reciben el nombre de *coeficientes de regresión* de X sobre Y e Y sobre X, respectivamente. Tiene una interesante propiedad,

$$\gamma_{X|Y} \cdot \gamma_{Y|X} = \frac{(cov(X,Y))^2}{var(X)var(Y)} = \rho_{xy}^2. \tag{3.25}$$

Ejemplo 3.9 *Consideremos el vector aleatorio (X, Y) con densidad conjunta,*

$$f_{XY}(x,y) = \begin{cases} e^{-y}, & 0 < x < y < +\infty \\ 0, & \text{en el resto.} \end{cases}$$

Las correspondientes marginales vienen dadas por

$$f_X(x) = \int_x^{+\infty} e^{-y} dy = \begin{cases} e^{-x}, & 0 < x < +\infty \\ 0, & \text{en el resto,} \end{cases}$$

y

$$f_Y(y) = \int_0^{y} e^{-y} dx = \begin{cases} ye^{-y}, & 0 < y < +\infty \\ 0, & \text{en el resto.} \end{cases}$$

Las distribuciones condicionadas se obtienen fácilmente a partir de las anteriores,

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{y}, & 0 < x < y \\ 0, & \text{en el resto,} \end{cases} \qquad f_{Y|X}(y|x) = \begin{cases} e^{x-y}, & x < y < +\infty \\ 0, & \text{en el resto.} \end{cases}$$

Las esperanzas condicionadas valen

$$E(X|Y) = \int_0^{y} x \frac{1}{y} dx = \frac{y}{2}, \qquad 0 < y < +\infty,$$

$$E(Y|X) = \int_x^{+\infty} y e^{x-y} dy = x + 1, \qquad 0 < x < +\infty.$$

Se trata en ambos casos de rectas, por lo que coincidirán con las rectas de regresión correspondientes. El valor absoluto del coeficiente de correlación podrá obtenerse fácilmente a partir de (3.25),

$$\rho_{XY}^2 = \gamma_{X|Y} \cdot \gamma_{Y|X} = \frac{1}{2} \cdot 1 = \frac{1}{2},$$

el signo será el mismo que el de la covarianza, que se comprueba fácilmente que vale 1. Así pues, $\rho_{XY} = \sqrt{1/2}$.

Ejemplo 3.10 (El fenómeno de la regresión a la media) *La recta de regresión nos ayuda a comprender porqué los hijos de padres muy altos tienden a ser más bajos que sus padres y los de padres muy bajos suelen superarlos en estatura. Pensemos en un padre que tiene una altura X cuando nace su hijo y preguntémonos la altura Y que alcanzará su hijo cuando tenga la edad que ahora tiene su padre.*

Si por μ_X, μ_Y, σ_X^2 y σ_Y^2 designamos, respectivamente, las medias y las varianzas de las alturas de padre e hijo, una hipótesis razonable y lógica es suponer que $\mu_X = \mu_Y = \mu$ y $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Por las propiedades de la recta de regresión, sabemos que la mejor estimación lineal de Y es $\hat{Y}$, que verifica

$$\hat{Y} - \mu = \frac{\sigma_{XY}}{\sigma^2}(X - \mu). \tag{3.26}$$

Como

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma^2},$$

de aquí $\sigma_{XY} = \rho_{XY}\sigma^2$ y sustituyendo en (3.26)

$$\hat{Y} - \mu = \rho_{XY}(X - \mu).$$

Puesto que $|\rho_{XY}| \leq 1$ y es de suponer que será positivo, concluiremos que la estatura del hijo estará más próxima a la media que la del padre.

Capítulo 4

Convergencia de sucesiones de variables aleatorias

4.1. Introducción

Los capítulos anteriores nos han permitido familiarizarnos con el concepto de variable y vector aleatorio, dotándonos de las herramientas que nos permiten conocer su comportamiento probabilístico. En el caso de un vector aleatorio somos capaces de estudiar el comportamiento conjunto de un número finito de variables aleatorias. Pero imaginemos por un momento los modelos probabilísticos asociados a los siguientes fenómenos:

1. lanzamientos sucesivos de una moneda,
2. tiempos transcurridos entre llamadas consecutivas a una misma centralita,
3. sucesión de estimadores de un parámetro cuando se incrementa el tamaño de la muestra...

En todos los casos las variables aleatorias involucradas lo son en cantidad numerable y habremos de ser capaces de estudiar su comportamiento conjunto y, tal y como siempre sucede en ocasiones similares, de conocer cuanto haga referencia al límite de la sucesión. Del comportamiento conjunto se ocupa una parte de la Teoría de la Probabilidad que dado su interés ha tomado entidad propia: la Teoría de los Procesos Estocásticos. En este capítulo nos ocuparemos de estudiar cuanto está relacionado con el límite de las sucesiones de variables aleatorias. Este estudio requiere en *primer lugar*, introducir los tipos de convergencia apropiados a la naturaleza de las sucesiones que nos ocupan, para en *segundo lugar* obtener las condiciones bajo las que tienen lugar las dos convergencias que nos interesan: la convergencia de la sucesión de variables a una constante (Leyes de los Grandes Números) y la convergencia a otra variable (Teorema Central del Límite). El estudio de esta segunda situación se ve facilitado con el uso de una herramienta conocida como *función característica* de la cual nos habremos ocupado previamente.

Dos sencillos ejemplos relacionados con la distribución Binomial no servirán de introducción y nos ayudarán a situarnos en el problema.

Ejemplo 4.1 (Un resultado de J. Bernoulli) *Si repetimos n veces un experimento cuyo resultado es la ocurrencia o no del suceso A, tal que $P(A) = p$, y si la repeticiones son independientes unas de otras, la variable X_n =*número de ocurrencias de *A, tiene una distribución $B(n, p)$. La variable X_n/n representa la frecuencia relativa de A y sabemos que*

$$E\left(\frac{X_n}{n}\right) = \frac{1}{n}E(X_n) = \frac{np}{n} = p,$$

y

$$var\left(\frac{X_n}{n}\right) = \frac{1}{n^2} var(X_n) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Si aplicamos la desigualdad de Chebyshev,

$$P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{var(X_n/n)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \xrightarrow{n} 0.$$

Deducimos que la frecuencia relativa de ocurrencias de A converge*, en algún sentido, a P(A).*

Ejemplo 4.2 (Binomial vs Poisson) *El segundo ejemplo ya fue expuesto en la página 22 y no lo repetiremos aquí. Hacía referencia a la aproximación de la distribución Binomial mediante la distribución de Poisson. Vimos que cuando tenemos un gran número de pruebas Bernoulli con una probabilidad de éxito muy pequeña de manera que* $\lim_n np_n = \lambda$, $0 < \lambda < +\infty$, *la sucesión de funciones de cuantía de las variables aleatorias* $X_n \sim B(n, p_n)$ *converge a la función de cuantía de* $X \sim Po(\lambda)$.

Dos ejemplos con sucesiones de variables Binomiales que han conducido a *límites* muy distintos. En el primero, el valor límite es la probabilidad de un suceso, y por tanto una constante; en el segundo la función de cuantía tiende a otra función de cuantía.

4.2. Tipos de convergencia

Comenzaremos por formalizar el tipo de convergencia que aparece en el primer ejemplo. Para ello, y también para el resto de definiciones, sobre un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ consideremos la sucesión de variables aleatorias $\{X_n\}$ y la variable aleatoria X.

Definición 4.1 (Convergencia en probabilidad) *Decimos que* $\{X_n\}$ *converge a X en probabilidad,* $X_n \xrightarrow{P} X$*, si para cada* $\delta > 0$*,*

$$\lim_n P\{\omega : |X_n(\omega) - X(\omega)| > \delta\} = 0.$$

No es esta una convergencia puntual como las que estamos acostumbrados a utilizar en Análisis Matemático. La siguiente sí es de este tipo.

Definición 4.2 (Convergencia casi segura o con probabilidad 1) *Decimos que* $\{X_n\}$ *converge casi seguramente*[1] *a X (o con probabilidad 1),* $X_n \xrightarrow{a.s.} X$*, si*

$$P(\{\omega : \lim_n X_n(\omega) = X(\omega)\}) = 1.$$

El último tipo de convergencia involucra a las funciones de distribución asociadas a cada variable y requiere previamente una definición para la convergencia de aquellas.

Definición 4.3 (Convergencia débil) *Sean* F_n*,* $n \geq 1$*, y F funciones de distribución de probabilidad. Decimos que la sucesión* F_n *converge débilmente*[2] *a F,* $F_n \xrightarrow{\omega} F$*, si* $\lim_n F_n(x) = F(x)$*,* $\forall x$ *que sea punto de continuidad de F.*

[1] Utilizaremos la abreviatura *a. s.*, que corresponde a las iniciales de **a**lmost **s**urely, por ser la notación más extendida

[2] Utilizaremos la abreviatura ω, que corresponde a la inicial de **w**eakly, por ser la notación más extendida

Definición 4.4 (Convergencia en ley) *Decimos que $\{X_n\}$ converge en ley a X, $X_n \xrightarrow{L} X$, si $F_{X_n} \xrightarrow{\omega} F_X$. Teniendo en cuenta la definición de F_n y F, la convergencia en ley puede expresarse también*

$$X_n \xrightarrow{L} X \Longleftrightarrow \lim_n P(X_n \leq x) = P(X \leq x),$$

$\forall x$ que sea punto de continuidad de F.

Las relaciones entre los tres tipos de convergencia se establecen en el siguiente teorema.

Teorema 4.1 (Relaciones entre convergencias) *Sean X_n y X variables aleatorias definidas sobre un mismo espacio de probabilidad entonces:*

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{P}{\to} X \Rightarrow X_n \overset{L}{\to} X$$

Demostración

1) $X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{P}{\to} X$

Para $\delta > 0$ sean $A = \{\omega : \lim_n X_n(\omega) \neq X(\omega)\}$ y $A_n = \{\omega : |X_n(\omega) - X(\omega)| > \delta\}$, entonces[3] $\limsup A_n \subset A$ y $P(\limsup A_n) = 0$, y de aquí $\lim P(A_n) = 0$.

2) $X_n \overset{P}{\to} X \Rightarrow X_n \overset{L}{\to} X$

Si x es tal que $P(X = x) = 0$, se verifica que

$$\begin{aligned} &P(X \leq x - \varepsilon) = \\ &\quad = P(X \leq x - \varepsilon, X_n \leq x) + P(X \leq x - \varepsilon, X_n > x) \\ &\quad \leq P(X_n \leq x) + P(|X_n - X| > \varepsilon) \end{aligned}$$

y

$$\begin{aligned} &P(X_n \leq x) = \\ &\quad = P(X_n \leq x, X \leq x + \varepsilon) + P(X_n \leq x, X > x + \varepsilon) \\ &\quad \leq P(X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon). \end{aligned}$$

Expresando conjuntamente las dos desigualdades tenemos:

$$\begin{aligned} &P(X \leq x - \varepsilon) - P(|X_n - X| > \varepsilon) \leq \\ &\quad \leq P(X_n \leq x) \leq P(X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon), \end{aligned}$$

pero $\lim_n P(|X_n - X| > \varepsilon) = 0$ para cualquier ε positivo por lo que:

$$P(X \leq x - \varepsilon) \leq \liminf_n P(X_n \leq x) \leq \limsup_n P(X_n \leq x) \leq P(X \leq x + \varepsilon)$$

y, puesto que $P(X = x) = 0$ es equivalente a que $F(x) = P(X \leq x)$ sea continua en x, se sigue el resultado. ♠

[3]Recordemos las definiciones

$$\liminf_n A_n = \bigcup_{n\geq 1} \bigcap_{k \geq n} A_k \quad y \quad \limsup_n A_n = \bigcap_{n\geq 1} \bigcup_{k \geq n} A_k,$$

y la siguiente cadena de desigualdades,

$$P(\liminf_n A_n) \leq \liminf_n P(A_n) \leq \limsup_n P(A_n) \leq P(\limsup_n A_n).$$

Las convergencias *casi segura* y *en probabilidad* tienen distinta naturaleza, mientras aquella es de tipo puntual, esta última es de tipo conjuntista. El ejemplo que sigue ilustra bien esta diferencia y pone de manifiesto que la contraria de la primera implicación no es cierta.

Ejemplo 4.3 (La convergencia en probabilidad $\not\Longrightarrow$ la convergencia casi segura) *Como espacio de probabilidad consideraremos el intervalo unidad dotado de la σ-álgebra de Borel y de la medida de Lebesgue, es decir, un espacio de probabilidad uniforme en [0,1]. Definimos la sucesión $X_n = \mathbf{1}_{I_n}$, $\forall n$, con $I_n = [\frac{p}{2^q}, \frac{p+1}{2^q}]$, siendo p y q los únicos enteros positivos que verifican, $p + 2^q = n$ y $0 \leq p < 2^q$. Obviamente $q = q(n)$ y $\lim_n q(n) = +\infty$. Los primeros términos de la sucesión son,*

$$\begin{array}{llll} n=1 & q=0, p=0 & X_1 = \mathbf{1}_{[0,1[} \\ n=2 & q=1, p=0 & X_2 = \mathbf{1}_{[0,\frac{1}{2}[} \\ n=3 & q=1, p=1 & X_3 = \mathbf{1}_{[\frac{1}{2},1[} \\ n=4 & q=2, p=0 & X_4 = \mathbf{1}_{[0,\frac{1}{4}[} \\ n=5 & q=2, p=1 & X_5 = \mathbf{1}_{[\frac{1}{4},\frac{1}{2}[} \\ n=6 & q=2, p=2 & X_6 = \mathbf{1}_{[\frac{1}{2},\frac{3}{4}[} \\ n=7 & q=2, p=3 & X_7 = \mathbf{1}_{[\frac{3}{4},1[} \\ \ldots\ldots & & \end{array}$$

Observemos que si $X = 0$, $\forall \delta > 0$ se tiene $\lambda\{\omega : |X_n(\omega)| > \delta\} = \lambda\{\omega : |X_n(\omega)| = 1\} = \lambda(I_n) = 2^{-q}$, $2^q \leq n < 2^{q+1}$ y $X_n \xrightarrow{\lambda} 0$; pero dada la construcción de las X_n en ningún $\omega \in [0,1]$ se verifica $\lim X_n(\omega) = 0$.

La convergencia en ley (débil) no implica la convergencia en probabilidad, como pone de manifiesto el siguiente ejemplo, lo que justifica el nombre de *convergencia débil* puesto que es la última en la cadena de implicaciones.

Ejemplo 4.4 *Consideremos una variable Bernoulli con $p = 1/2$, $X \sim B(1, 1/2)$ y definamos una sucesión de variables aleatorias, $X_n = X$, $\forall n$. La variable $Y = 1 - X$ tiene la misma distribución que X, es decir, $Y \sim B(1, 1/2)$. Obviamente $X_n \xrightarrow{L} Y$, pero como $|X_n - Y| = |2X - 1| = 1$, no puede haber convergencia en probabilidad.*

Hay, no obstante, una situación en las que la convergencia débil y la convergencia en probabilidad son equivalentes, cuando el límite es a una constante. Veámoslo.

Teorema 4.2 *Si $X_n \xrightarrow{L} c$ entonces $X_n \xrightarrow{P} c$.*

Demostración.- Si F_n es la función de distribución de X_n, la convergencia en Ley implica que $F_n(x) \to F_c(x)$ tal que

$$F_c(x) = \begin{cases} 0, & \text{si } x < c; \\ 1, & \text{si } x \geq c. \end{cases}$$

Sea ahora $\delta > 0$

$$\begin{aligned} P(|X_n - c| > \delta) &= P(X_n - c < -\delta) + P(X_n - c > \delta) \\ &= P(X_n < c - \delta) + P(X_n > c + \delta) \\ &\leq P(X_n \leq c - \delta) + 1 - P(X_n \leq c + \delta) \\ &= F_n(c - \delta) + 1 - F_n(c + \delta), \end{aligned}$$

por ser $c - \delta$ y $c + \delta$ puntos de continuidad de F_c. La convergencia débil de F_n a F_c implica que $P(|X_n - c| > \delta) \to 0$ y en consecuencia $X_n \xrightarrow{P} c$. ♠

4.3. Leyes de los Grandes Números

El nombre de *leyes de los grandes números* hace referencia al estudio de un tipo especial de límites derivados de la sucesión de variables aleatorias $\{X_n\}$. Concretamente los de la forma $\lim_n \frac{S_n - a_n}{b_n}$, con $S_n = \sum_{i=1}^n X_i$ y siendo $\{a_n\}$ y $\{b_n\}$ sucesiones de constantes tales que $\lim b_n = +\infty$. En esta sección fijaremos las condiciones para saber cuando existe convergencia a.s, y como nos ocuparemos también de la convergencia en probabilidad, las leyes se denominarán *fuerte* y *débil*, respectivamente.

Teorema 4.3 (Ley débil) *Sea $\{X_k\}$ una sucesión de variables aleatorias independientes tales que $E(X_k^2) < +\infty,\ \forall k$, y $\lim_n \frac{1}{n^2} \sum_{k=1}^n var(X_k) = 0$, entonces*

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - E(X_k)) \xrightarrow{P} 0.$$

Demostración.- Para $S_n = \frac{1}{n} \sum_{k=1}^n (X_k - E(X_k))$, $E(S_n) = 0$ y $var(S_n) = \frac{1}{n^2} \sum_{k=1}^n var(X_k)$. Por la desigualdad de Chebyshev, $\forall \varepsilon > 0$

$$P(|S_n| \geq \varepsilon) \leq \frac{var(S_n)}{\varepsilon^2} = \frac{1}{n^2 \varepsilon^2} \sum_{k=1}^{n} var(X_k),$$

que al pasar al límite nos asegura la convergencia en probabilidad de S_n a 0. ♠

Corolario 4.1 *Si las X_n son i.i.d. con varianza finita y esperanza común $E(X_1)$, entonces $\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{P} E(X_1)$.*

Demostración.- Si $var(X_k) = \sigma^2,\ \forall k$, tendremos $\frac{1}{n^2} \sum_{k=1}^n var(X_k) = \frac{\sigma^2}{n}$ que tiende a cero con n. Es por tanto de aplicación la ley débil que conduce al resultado enunciado. ♠

Este resultado fué demostrado por primera vez por J. Bernoulli para variables con distribución Binomial (véase el ejemplo 4.1), versión que se conoce como la *ley de los grandes números de Bernoulli*

El siguiente paso será fijar las condiciones para que el resultado sea válido bajo convergencia a.s.

Teorema 4.4 (Ley fuerte) *Si $\{X_k\}$ es una sucesión de variables aleatorias i.i.d. con media finita, entonces*

$$\sum_{k=1}^{n} \frac{X_k}{n} \xrightarrow{a.s.} E(X_1).$$

Corolario 4.2 *Si $\{X_k\}$ es una sucesión de variables aleatorias i.i.d. con $E(X_1^-) < +\infty$ y $E(X_1^+) = +\infty$, entonces $\frac{S_n}{n} \xrightarrow{a.s.} \infty$.*

La demostración de la *ley fuerte* es de una complejidad, aun en su versión más sencilla debida a Etemadi, fuera del alcance y pretensiones de este texto. La primera demostración, más compleja que la de Etemadi, se debe a Kolmogorov y es el resultado final de una cadena de propiedades previas de gran interés y utilidad en sí mismas. Aconsejamos vivamente al estudiante que encuentre ocasión de hojear ambos desarrollos en cualquiera de los textos habituales (Burrill, Billingsley,...), pero que en ningún caso olvide el desarrollo de Kolmogorov.

4.3.1. Aplicaciones de la ley de los grandes números

El teorema de Glivenko-Cantelli

Para las variables aleatorias $X_1, X_2, \ldots, X_n$ se define la función de distribución empírica mediante

$$F_n(x,\omega) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{]-\infty,x]}(X_k(\omega)).$$

Cuando todas las variables tienen la misma distribución, $F_n(x,\omega)$ es el estimador natural de la función de distribución común, $F(x)$. El acierto en la elección de este estimador se pone de manifiesto en el siguiente resultado.

Teorema 4.5 *Sea $\{X_k\}$ una sucesión de variables aleatorias i.i.d. con función de distribución común $F(x)$, entonces $F_n(x,\omega) \xrightarrow{a.s.} F(x)$.*

Demostración.- Para cada x, $F_n(x,\omega)$ es una variable aleatoria resultante de sumar las n variables aleatorias independientes, $\mathbf{1}_{]-\infty,x]}(X_k(\omega))$, $k = 1, \ldots, n$, cada una de ellas con la misma esperanza, $E(\mathbf{1}_{]-\infty,x]}(X_k(\omega))) = \mathcal{P}(X_k \leq x) = F(x)$. Aplicando la ley fuerte de los grandes números,

$$F_n(x,\omega) \xrightarrow{a.s.} F(x),$$

que es el resultado buscado. ♠

Este resultado es previo al teorema que da nombre al apartado y que nos permite contrastar la hipótesis de suponer que F es la distribución común a toda la sucesión.

Teorema 4.6 (Glivenko-Cantelli) *Sea $\{X_k\}$ una sucesión de variables aleatorias i.i.d. con función de distribución común $F(x)$. Hagamos $D_n(\omega) = \sup_x |F_n(x,\omega) - F(x)|$, entonces $D_n \xrightarrow{a.s.} 0$.*

La demostración, muy técnica, la omitimos y dejamos al interés del lector consultarla en el texto de Billingsley.

Cálculo aproximado de integrales por el método de Monte-Carlo

Sea $f(x) \in \mathcal{C}([0,1])$ con valores en $[0,1]$. Una aproximación al valor de $\int_0^1 f(x)dx$ puede obtenerse a partir de una sucesión de pares de variables aleatorias distribuidas uniformemente en $[0,1]$, $(X_1,Y_1), (X_2,Y_2), \ldots$. Para ello hagamos,

$$Z_i = \begin{cases} 1, & si\ f(X_i) \geq Y_i \\ 0, & si\ f(X_i) < Y_i. \end{cases}$$

Así definidas las Z_i son variables Bernoulli con parámetro $p = E(Z_i) = P(f(X_i) \geq Y_i) = \int_0^1 f(x)dx$, y aplicándoles la ley fuerte de los grandes números tendremos que

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{a.s.} \int_0^1 f(x)dx,$$

lo que en términos prácticos supone simular los pares (X_i,Y_i), $i = 1, \ldots, n$, con X_i e $Y_i \sim U(0,1)$, y calcular la proporción de ellos que caen por debajo de la gráfica $y = f(x)$.

Aplicación de la LGN a la aproximación de funciones

Sea g una función acotada definida sobre $[0,1]$, la función B_n definida sobre $[0,1]$ mediante

$$B_n(x) = \sum_{k=0}^{n} g\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k},$$

es conocida como polinomio de Bernstein de grado n.

El teorema de aproximación de Weierstrass asegura que toda función continua sobre un intervalo cerrado puede ser aproximada uniformemente mediante polinomios. Probemos dicha afirmación para los polinomios de Bernstein.

Si la función g a aproximar es continua en $[0,1]$, será uniformemente continua, entonces

$$\forall \epsilon > 0, \exists \delta > 0 \text{ tal que } |g(x) - g(y)| < \epsilon, \text{ si } |x-y| < \delta.$$

Además g estará también acotada y por tanto $|g(x)| < M, \forall x \in [0,1]$.

Sea ahora un x cualquiera en $[0,1]$,

$$\begin{aligned}
|g(x) - B_n(x)| &= \left| g(x)\sum_{k=0}^{n}\binom{n}{k}x^k(1-x)^{n-k} - \sum_{k=0}^{n} g\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}\right| \\
&\leq \sum_{k=0}^{n}\left|g(x) - g\left(\frac{k}{n}\right)\right|\binom{n}{k}x^k(1-x)^{n-k} \\
&= \sum_{|k/n-x|<\delta}\left|g(x) - g\left(\frac{k}{n}\right)\right|\binom{n}{k}x^k(1-x)^{n-k} + \\
&\quad + \sum_{|k/n-x|\geq\delta}\left|g(x) - g\left(\frac{k}{n}\right)\right|\binom{n}{k}x^k(1-x)^{n-k} \\
&\leq \epsilon + 2M\sum_{|k/n-x|\geq\delta}\binom{n}{k}x^k(1-x)^{n-k}.
\end{aligned}$$

Si $Z_n \sim B(n,x)$, el último sumatorio no es más que

$$P\left(\left|\frac{Z_n}{n} - x\right| > \delta\right) = \sum_{|k/n-x|\geq\delta}\binom{n}{k}x^k(1-x)^{n-k},$$

y tendremos

$$|g(x) - B_n(x)| \leq \epsilon + 2MP\left(\left|\frac{Z_n}{n} - x\right| > \delta\right),$$

pero por la ley de los grandes números

$$\frac{Z_n}{n} \xrightarrow{P} x \quad \textit{y por tanto} \quad P\left(\left|\frac{Z_n}{n} - x\right| > \delta\right) \longrightarrow 0,$$

lo que demuestra la convergencia uniforme de B_n a g en $[0,1]$.

4.4. Función característica

La *función característica* es una herramienta de gran utilidad en Teoría de la Probabilidad, una de sus mayores virtudes reside en facilitar la obtención de la distribución de probabilidad de

la suma de variables aleatorias y la del límite de sucesiones de variables aleatorias, situaciones ambas que aparecen con frecuencia en Inferencia Estadística.

El concepto de *función característica*, introducido por Lyapunov en una de la primeras versiones del Teorema Central del Límite, procede del Análisis Matemático donde se le conoce con el nombre de *transformada de Fourier*.

Definición 4.5 *Sea X una variable aleatoria y sea $t \in R$. La función característica de X, $\phi_X(t)$, se define como $E(e^{itX}) = E(\cos tX) + iE(\sin tX)$.*

Como $|e^{itX}| \leq 1$, $\forall t$, $\phi_X(t)$ existe siempre y está definida $\forall t \in R$. Para su obtención recordemos que,

Caso discreto.- Si X es una v. a. discreta con soporte D_X y función de cuantía $f_X(x)$,

$$\phi_X(t) = \sum_{x \in D_X} e^{itx} f_X(x). \tag{4.1}$$

Caso continuo.- Si X es una v. a. continua con función de densidad de probabilidad $f_X(x)$,

$$\phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f_X(x) dx. \tag{4.2}$$

De la definición se derivan, entre otras, las siguientes propiedades:

Pϕ1) $\phi_X(0) = 1$

Pϕ2) $|\phi_X(t)| \leq 1$

Pϕ3) $\phi_X(t)$ es uniformemente continua

En efecto,

$$\phi_X(t+h) - \phi_X(t) = \int_\Omega e^{itX}(e^{ihX} - 1)\, dP.$$

Al tomar módulos,

$$|\phi_X(t+h) - \phi_X(t)| \leq \int_\Omega |e^{ihX} - 1|\, dP, \tag{4.3}$$

pero $|e^{ihX} - 1| \leq 2$ y (4.3) será finito, lo que permite intercambiar integración y paso al límite, obteniendo

$$\lim_{h \to 0} |\phi_X(t+h) - \phi_X(t)| \leq \int_\Omega \lim_{h \to 0} |e^{ihX} - 1|\, dP = 0.$$

Pϕ4) Si definimos $Y = aX + b$,

$$\phi_Y(t) = E(e^{itY}) = E\left(e^{it(aX+b)}\right) = e^{itb}\phi_X(at)$$

Pϕ5) Si $E(X^n)$ existe, la función característica es n veces diferenciable y $\forall k \leq n$ se verifica $\phi_X^{(k)}(0) = i^k E(X^k)$

La propiedad 5 establece un interesante relación entre las derivadas de $\phi_X(t)$ y los momentos de X cuando estos existen, relación que permite desarrollar $\phi_X(t)$ en serie de potencias. En efecto, si $E(X^n)$ existe $\forall n$, entonces,

$$\phi_X(t) = \sum_{k \geq 0} \frac{i^k E(X^k)}{k!} t^k. \tag{4.4}$$

4.4.1. Función característica e independencia

Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y definimos $Y = X_1+X_2+\cdots+X_n$, por la observación (3.3) de la página 80 tendremos,

$$\phi_Y(t) = E\left(e^{it(X_1+X_2+\cdots+X_n)}\right) = E\left(\prod_{k=1}^{n} e^{itX_k}\right) = \prod_{k=1}^{n} E\left(e^{itX_k}\right) = \prod_{k=1}^{n} \phi_{X_k}(t), \tag{4.5}$$

expresión que permite obtener con facilidad la función característica de la suma de variables independientes y cuya utilidad pondremos de manifiesto de inmediato.

4.4.2. Funciones características de algunas distribuciones conocidas

Bernoulli.- Si $X \sim B(1,p)$

$$\phi_X(t) = e^0 q + e^{it}p = q + pe^{it}.$$

Binomial.- Si $X \sim B(n,p)$

$$\phi_X(t) = \prod_{k=1}^{n}(q + pe^{it}) = (q + pe^{it})^n.$$

Poisson.- Si $X \sim P(\lambda)$

$$\phi_X(t) = \sum_{x\geq 0} e^{itx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x\geq 0}\frac{(\lambda e^{it})^x}{x!} = e^{\lambda(e^{it}-1)}.$$

Normal tipificada.- Si $Z \sim N(0,1)$, sabemos que existen los momentos de cualquier orden y en particular, $E(Z^{2n+1}) = 0,\ \forall n$ y $E(Z^{2n}) = \frac{(2n)!}{2^n n!},\ \forall n$. Aplicando (4.4),

$$\phi_Z(t) = \sum_{n\geq 0}\frac{i^{2n}(2n)!}{2^n(2n)!n!}t^{2n} = \sum_{n\geq 0}\frac{\left(\frac{(it)^2}{2}\right)^n}{n!} = \sum_{n\geq 0}\frac{\left(-\frac{t^2}{2}\right)^n}{n!} = e^{-\frac{t^2}{2}}.$$

Para obtener $\phi_X(t)$ si $X \sim N(\mu,\sigma^2)$, podemos utilizar el resultado anterior y P4. En efecto, recordemos que X puede expresarse en función de Z mediante $X = \mu + \sigma Z$ y aplicando P4,

$$\phi_X(t) = e^{i\mu t}\phi_Z(\sigma t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}.$$

Observación 4.1 *Obsérvese que $Im(\phi_Z(t)) = 0$. El lector puede comprobar que no se trata de un resultado exclusivo de la Normal tipificada, si no de una propiedad que poseen todas las v. a. con distribución de probabilidad simétrica, es decir, aquellas que verifican (P(X ≥ x)=P(X ≤ −x)).*

Gamma.- Si $X \sim G(\alpha,\lambda)$, su función de densidad de probabilidad viene dada por

$$f_X(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} & \text{si } x > 0, \\ \\ 0 & \text{si } x \leq 0, \end{cases}$$

por lo que aplicando (4.2),

$$\phi_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{itx} x^{\alpha-1} e^{-\lambda x} dx,$$

que con el cambio $y = x(1 - it/\lambda)$ conduce a

$$\phi_X(t) = \left(1 - \frac{it}{\lambda}\right)^{-\alpha}.$$

Hay dos casos particulares que merecen ser mencionados:

Exponencial.- La distribución exponencial puede ser considerada un caso particular de $G(\alpha, \lambda)$ cuando $\alpha = 1$. A partir de aquí,

$$\phi_X(t) = \frac{\lambda}{\lambda - it}.$$

Chi-cuadrado.- Cuando $\alpha = n/2$ y $\lambda = 1/2$, decimos que X tiene una distribución χ^2 con n grados de libertad, $X \sim \chi^2_n$. Su función característica será

$$\phi_X(t) = (1 - 2it)^{-\frac{n}{2}}.$$

4.4.3. Teorema de inversión. Unicidad

Hemos obtenido la función característica de una v. a. X a partir de su distribución de probabilidad, pero es posible proceder de manera inversa por cuanto el conocimiento de $\phi_X(t)$ permite obtener $F_X(x)$.

Teorema 4.7 (Fórmula de inversión de Lévy) *Sean $\phi(t)$ y $F(x)$ las funciones característica y de distribución de la v. a. X y sean $a \leq b$ sendos puntos de continuidad de F, entonces*

$$F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt.$$

Que puede también expresarse

$$F(x_0 + h) - F(x_0 - h) = \lim_{T\to\infty} \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-itx_0} \phi(t) dt,$$

donde $h > 0$ y $x_0 + h$ y $x_0 - h$ son puntos de continuidad de F.

Demostración.- Sea

$$\begin{aligned} J &= \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-itx_0} \phi(t) dt \\ &= \frac{1}{\pi} \int_{-T}^{T} \frac{\sin ht}{t} e^{-itx_0} \left[\int_R e^{itx} dF(x) \right] dt \\ &= \frac{1}{\pi} \int_{-T}^{T} \left[\int_R \frac{\sin ht}{t} e^{it(x-x_0)} dF(x) \right] dt, \end{aligned}$$

donde $\phi(t) = E(e^{itX}) = \int_R e^{itx} dF(x)$. Pero

$$\left| \frac{\sin ht}{t} e^{it(x-x_0)} \right| \leq \left| \frac{\sin ht}{t} \right| \leq h$$

y como T es finito J será también finito y podemos aplicar el teorema de Fubini que nos permite el cambio en el orden de integración. Es decir

$$\begin{aligned} J &= \frac{1}{\pi} \int_R \left[\int_{-T}^{T} \frac{\sin ht}{t} e^{it(x-x_0)} dt \right] dF(x) \\ &= \frac{1}{\pi} \int_R \left[\int_{-T}^{T} \frac{\sin ht}{t} (\cos t(x-x_0) + i \sin t(x-x_0)) dt \right] dF(x) \\ &= \frac{2}{\pi} \int_R \left[\int_{0}^{T} \frac{\sin ht}{t} \cos t(x-x_0) dt \right] dF(x). \end{aligned}$$

Aplicando la fórmula $\sin\alpha \cos\beta = \frac{1}{2}(\sin(\alpha+\beta) + \sin(\alpha-\beta))$, con $\alpha = ht$ y $\beta = t(x-x_0)$,

$$\begin{aligned} J &= \frac{2}{\pi} \int_R \left[\int_{0}^{T} \frac{1}{2t} (\sin t(x-x_0+h) - \sin t(x-x_0-h)) dt \right] dF(x) \\ &= \frac{1}{\pi} \int_R \left[\int_{0}^{T} \frac{\sin t(x-x_0+h)}{t} dt - \int_0^T \frac{\sin t(x-x_0-+h)}{t} dt \right] dF(x) \\ &= \frac{1}{\pi} \int_R g(x,T) dF(x). \end{aligned}$$

Por lo que respecta a la función $g(x,T)$ observemos que

$$\lim_{T\to\infty} \int_0^T \frac{\sin x}{x} dx = \frac{\pi}{2}$$

y está acotada para $T > 0$. Por otra parte

$$\int_0^T \frac{\sin \alpha x}{x} dx = \int_0^{\alpha T} \frac{\sin u}{u} du$$

y en consecuencia

$$\lim_{T\to\infty} \frac{2}{\pi} \int_0^T \frac{\sin \alpha x}{x} dx = \begin{cases} 1, & \alpha > 0; \\ 0, & \alpha = 0; \\ -1, & \alpha < 0. \end{cases}$$

Aplicándolo a $g(x,T)$,

$$\lim_{T\to\infty} \frac{1}{\pi} g(x,T) = \begin{cases} 0, & x < x_0 - h; \\ \frac{1}{2}, & x = x_0 - h; \\ 1, & x_0 - h < x < x_0 + h; \\ \frac{1}{2}, & x = x_0 + h; \\ 0, & x > x_0 + h. \end{cases}$$

Como $|g(x,T)|$ está acotada podemos hacer uso de los teoremas de convergencia para permutar integración y paso al límite, con lo que tendremos

$$\begin{aligned}\lim_{T\to\infty} J &= \lim_{T\to\infty} \frac{1}{\pi}\int_R g(x,T)dF(x)\\ &= \frac{1}{\pi}\int_R \lim_{T\to\infty} g(x,T)dF(x)\\ &= \int_{x_0-h}^{x_0+h} dF(x) = F(x_0+h)-F(x_0-h).\end{aligned}$$

♠

Este resultado permite obtener $F(x)$ en cualquier x que sea punto de continuidad de F. Basta para ello que en $F(x)-F(y)$ hagamos que $y\to-\infty$ a través de puntos de continuidad. Como $F_X(x)$ es continua por la derecha, la tendremos también determinada en los puntos de discontinuidad sin más que descender hacia ellos a través de puntos de continuidad.

Si la variable es continua, un corolario del anterior teorema permite obtener la función de densidad directamente a partir de la función característica.

Corolario 4.3 *Si $\phi(t)$ es absolutamente integrable en R, entonces la función de distribución es absolutamente continua, su derivada es uniformemente continua y*

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{2\pi}\int_R e^{-itx}\phi(t)dt.$$

Demostración.- Del desarrollo del teorema anterior tenemos que

$$\begin{aligned}\left|\frac{F(x_0+h)-F(x_0-h)}{2h}\right| &= \left|\frac{1}{2\pi}\int_R \frac{\sin ht}{ht}e^{-itx_0}\phi(t)dt\right|\\ &\le \frac{1}{2\pi}\int_R \left|\frac{\sin ht}{ht}e^{-itx_0}\phi(t)\right|dt\\ &\le \frac{1}{2\pi}\int_R |\phi(t)|dt < +\infty,\end{aligned}$$

y por tanto

$$\lim_{h\to 0}\frac{F(x_0+h)-F(x_0-h)}{2h} = f(x_0) = \frac{1}{2\pi}\int_R e^{-itx_0}\phi(t)dt.$$

Existe por tanto la derivada en todos los puntos de continuidad de F. La continuidad absoluta de F deriva de la desigualdad

$$F(x_0+h)-F(x_0-h) \le \frac{2h}{2\pi}\int_R |\phi(t)|dt,$$

cuyo segundo miembro podemos hacer tan pequeño como queramos eligiendo h adecuadamente.

Por último, la continuidad uniforme de $f(x)$ se deriva de

$$|f(x+h)-f(x)| = \frac{1}{2\pi}\left|\int_R e^{-ix}(e^{-ith}-1)\phi(t)dt\right| \le \frac{1}{2\pi}\int_R |e^{-ith}-1||\phi(t)|dt, \qquad (4.6)$$

pero

$$\begin{aligned} |e^{-ith}-1| &= |(\cos th-1)-i\sin th| \\ &= \sqrt{(\cos th-1)^2+\sin^2 th} \\ &= \sqrt{2(1-\cos th)} \\ &= 2\left|\sin\frac{th}{2}\right|, \end{aligned}$$

y sustituyendo en (4.6)

$$\begin{aligned} |f(x+h)-f(x)| &\leq \frac{1}{2\pi}\int_R 2\left|\sin\frac{th}{2}\right||\phi(t)|dt \\ &\leq \frac{1}{2\pi}\int_{|t|\leq a} 2\left|\sin\frac{th}{2}\right||\phi(t)|dt+\frac{1}{2\pi}\int_{|t|>a}2|\phi(t)|dt, \end{aligned}$$

donde a se elige de forma que la segunda integral sea menor que $\epsilon/2$, lo que es posible por la integrabilidad absoluta de ϕ. La primera integral también puede acotarse por $\epsilon/2$ eligiendo h adecuadamente.

♠

Pero este teorema tiene una trascendencia mayor por cuanto implica la *unicidad* de la función característica, que no por casualidad recibe este nombre, porque *caracteriza*, al igual que lo hacen otras funciones asociadas a X (la de distribución, la de probabilidad o densidad de probabilidad, ...), su distribución de probabilidad. Podemos afirmar que si dos variables X e Y comparten la misma función característica tienen idéntica distribución de probabilidad. La combinación de este resultado con las propiedades antes enunciadas da lugar a una potente herramienta que facilita el estudio y obtención de las distribuciones de probabilidad asociadas a la suma de variables independientes. Veámoslo con algunos ejemplos.

1) Suma de Binomiales independientes.- Si las variables $X_k \sim B(n_k,p),\ k=1,2,\ldots,m$ son independientes, al definir $X=\sum_{k=1}^m X_k$, sabemos por (4.5) que

$$\phi_X(t)=\prod_{k=1}^m \phi_{X_k}(t)=\prod_{k=1}^m (q+pe^{it})^{n_k}=(q+pe^{it})^n,$$

con $n=n_1+n_2+\cdots+n_m$. Pero siendo esta la función característica de una variable $B(n,p)$, podemos afirmar que $X\sim B(n,p)$.

2) Suma de Poisson independientes.- Si nuestra suma es ahora de variables Poisson independientes, $X_k\sim P(\lambda_k)$, entonces

$$\phi_X(t)=\prod_{k=1}^m \phi_{X_k}(t)=\prod_{k=1}^m e^{\lambda_k(e^{it}-1)}=e^{\lambda(e^{it}-1)},$$

con $\lambda=\lambda_1+\lambda_2+\cdots+\lambda_m$. Así pues, $X\sim P(\lambda)$.

3) Combinación lineal de de Normales independientes.- Si $X=\sum_{k=1}^n c_kX_k$ con $X_k\sim$

$N(\mu_k, \sigma_k^2)$ e independientes,

$$\begin{aligned}\phi_X(t) &= \phi_{\sum c_k X_k}(t) = \prod_{k=1}^{n} \phi_{X_k}(c_k t) \\ &= \prod_{k=1}^{n} e^{ic_k t\mu_k - \frac{\sigma_k^2 c_k^2 t^2}{2}} \\ &= e^{i\mu t - \frac{\sigma^2 t^2}{2}}, \end{aligned} \tag{4.7}$$

Se deduce de (4.7) que $X \sim N(\mu, \sigma^2)$ con

$$\mu = \sum_{k=1}^{n} c_k \mu_k \qquad \text{y} \qquad \sigma^2 = \sum_{k=1}^{n} c_k^2 \sigma_k^2.$$

4) Suma de Exponenciales independientes.- En el caso de que la suma esté formada por n variables independientes todas ellas $Exp(\lambda)$,

$$\phi_X(t) = \left(\frac{\lambda}{\lambda - it}\right)^n = \left(1 - \frac{it}{\lambda}\right)^{-n},$$

y su distribución será la de una $G(n, \lambda)$.

5) Cuadrado de una $N(0,1)$.- Sea ahora $Y = X^2$ con $X \sim N(0,1)$, su función característica viene dada por

$$\phi_Y(t) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{x^2}{2}(1-2it)} dx = \frac{1}{(1-2it)^{\frac{1}{2}}},$$

lo que nos asegura que $Y \sim \chi_1^2$.

Algunos de estos resultados fueron obtenidos anteriormente, pero su obtención fué entonces mucho más laboriosa de lo que ahora ha resultado.

Distribuciones en el muestreo en una población $\mathbf{N(\mu, \sigma^2)}$

Si la distribución común a las componentes de la muestra aleatoria de tamaño n es una $N(\mu, \sigma^2)$, se verifica el siguiente teorema:

Teorema 4.8 *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño* n *de una población $N(\mu, \sigma^2)$. Entonces, $\overline{X}_n \sim N(\mu, \sigma^2/n)$ y $(n-1)S_n^2/\sigma^2 \sim \chi_{n-1}^2$, y además son independientes.*

Demostración.- La distribución de $\overline{X}_n$ se deduce de (4.7) haciendo $c_k = 1/n$, $\forall k$. Tendremos que

$$\mu = \sum_{k=1}^{n} c_k \mu_k = \mu \qquad \text{y} \qquad \sigma^2 = \sum_{k=1}^{n} c_k^2 \sigma_k^2 = \frac{\sigma^2}{n}.$$

La obtención de la distribución de S_n^2 exige primero demostrar su independencia de $\overline{X}_n$. Para ello recordemos la expresión de la densidad conjunta de las n componentes de la muestra (página 85),

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{\left\{-\frac{1}{2\sigma^2}\sum_{i=i}^{n}(x_i - \mu)^2\right\}},$$

pero fácilmente se comprueba que

$$\sum_{i=i}^{n}(x_i-\mu)^2=\sum_{i=i}^{n}(x_i-\overline{x}_n)^2+n(\overline{x}_n-\mu)^2. \tag{4.8}$$

Sustituyendo en la anterior expresión

$$\begin{aligned} f(x_1,x_2,\ldots,x_n) &= \frac{1}{\sigma^n(2\pi)^{n/2}}e^{\left\{-\frac{1}{2\sigma^2}\sum_{i=i}^{n}(x_i-\overline{x}_n)^2-\frac{n}{2\sigma^2}(\overline{x}_n-\mu)^2\right\}} \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}}e^{-\left(u/2\sigma^2\right)}e^{-\left(n/2\sigma^2\right)(\overline{x}_n-\mu)^2}, \end{aligned}$$

con $u=\sum(x_i-\overline{x}_n)^2$.

Hagamos ahora el cambio

$$x_i=\overline{x}_n+v_i\sqrt{u},\quad i=1,2,\ldots,n.$$

Como

$$\sum_{i=1}^{n}v_i=0 \qquad \text{y} \qquad \sum_{i=1}^{n}v_i^2=1, \tag{4.9}$$

dos de las nuevas variables serán función de las restantes. Resolviendo las ecuaciones de (4.9) para v_{n-1} y v_n, una de las dos soluciones es

$$v_{n-1}=\frac{A-B}{2} \qquad \text{y} \qquad v_n=\frac{A+B}{2},$$

con

$$A=-\sum_{i=1}^{n-2}v_i \qquad \text{y} \qquad B=\left(2-3\sum_{i=1}^{n-2}v_i^2-\sum_{i\neq j}v_iv_j\right)^2.$$

Así pues, podemos expresar cada x_i en función de $(v_1,\ldots,v_{n-2},\overline{x}_n,u)$ de la siguiente forma,

$$\begin{aligned} x_i&=\overline{x}+v_i\sqrt{u},\quad i=1,\ldots,n-2, \\ x_{n-1}&=\overline{x}_n+\frac{A-B}{2}\sqrt{u},\quad x_n=\overline{x}_n+\frac{A+B}{2}\sqrt{u}. \end{aligned} \tag{4.10}$$

El Jacobiano de (4.10), laborioso pero sencillo de obtener, tiene la forma

$$|J|=u^{(n-2)/2}h(v_1,\ldots,v_{n-2}),$$

donde la expresión exacta de h no es necesaria a los efectos de obtener la densidad conjunta de $(v_1,\ldots,v_{n-2},\overline{x}_n,u)$. Utilizando del Teorema 2.5 (página 60), podemos escribir

$$g(v_1,\ldots,v_{n-2},\overline{x}_n,u)=\frac{1}{\sigma^n(2\pi)^{n/2}}u^{n-2}e^{-\left(u/2\sigma^2\right)}e^{-\left(n/2\sigma^2\right)(\overline{x}_n-\mu)^2}h(v_1,\ldots,v_{n-2}),$$

que no es más que el producto de las tres densidades siguientes

$$\begin{aligned} g_1(u) &= c_1u^{n-2}e^{-\left(u/2\sigma^2\right)} \\ g_2(\overline{x}_n) &= c_2e^{-\left(n/2\sigma^2\right)(\overline{x}_n-\mu)^2} \\ g_3(v_1,\ldots,v_{n-2}) &= c_3h(v_1,\ldots,v_{n-2}). \end{aligned}$$

Esta factorización nos permite afirmar que las variables $U = (n-1)S_n^2$, $\overline{X}_n$ y $(V_1, V_2, \ldots, V_{n-2})$ son independientes.

Volvamos ahora a la distribución de S_n^2. De (4.8) podemos escribir,

$$\sum_{i=i}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum_{i=i}^{n}\left(\frac{X_i-\overline{X}_n}{\sigma}\right)^2+\left(\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\right)^2=(n-1)\frac{S_n^2}{\sigma^2}+\left(\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{\sigma}\right)^2. \quad (4.11)$$

Como $Y_i = (X_i - \mu)/\sigma$, $i = 1, \ldots, n$ y $Z = \sqrt{n}\left(\overline{X}_n - \mu\right)/\sigma$ son todas ellas variables $N(0,1)$, tendremos que

$$\sum_{i=i}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi_n^2 \qquad \text{y} \qquad \sqrt{n}\left(\frac{\overline{X}_n-\mu}{\sigma}\right)^2\sim\chi_1^2,$$

pero como $(n-1)S_n^2$ y $\overline{X}_n$ son independientes, también lo serán $(n-1)S_n^2/\sigma^2$ y $\overline{X}_n$, y al calcular funciones características en ambos miembros de (4.11),

$$(1-2it)^{n/2}=(1-2it)^{1/2}\phi_{(n-1)S_n^2/\sigma^2}.$$

De donde,

$$\phi_{(n-1)S_n^2/\sigma^2}=(1-2it)^{(n-1)/2},$$

y

$$(n-1)\frac{S_n^2}{\sigma^2}\sim\chi_{n-1}^2.$$

♠

4.4.4. Teorema de continuidad de Lévy

Se trata del último de los resultados que presentaremos y permite conocer la convergencia de una sucesión de v. a. a través de la convergencia puntual de la sucesión de sus funciones características.

Teorema 4.9 (Directo) *Sea $\{X_n\}_{n\geq 1}$ una sucesión de v. a. y $\{F_n(x)\}_{n\geq 1}$ y $\{\phi_n(t)\}_{n\geq 1}$ las respectivas sucesiones de sus funciones de distribución y características. Sea X una v. a. y $F_X(x)$ y $\phi(t)$ sus funciones de distribución y característica, respectivamente. Si $F_n \xrightarrow{w} F$ (es decir, $X_n \xrightarrow{L} X$), entonces*

$$\phi_n(t)\longrightarrow\phi(t), \quad \forall t\in R.$$

Resultado que se completa con el teorema inverso.

Teorema 4.10 (Inverso) *Sea $\{\phi_n(t)\}_{n\geq 1}$ una sucesión de funciones características y $\{F_n(x)\}_{n\geq 1}$ la sucesión de funciones de distribución asociadas. Sea $\phi(t)$ una función continua en 0, si $\forall t \in R$, $\phi_n(t) \to \phi(t)$, entonces*

$$F_n\xrightarrow{w}F,$$

donde $F(x)$ en una función de distribución cuya función característica es $\phi(t)$.

Este resultado permite, como decíamos antes, estudiar el comportamiento límite de sucesiones de v. a. a través del de sus funciones características, generalmente de mayor sencillez. Sin duda una de sus aplicaciones más relevantes ha sido el conjunto de resultados que se conocen como *Teorema Central del Límite* (TCL), bautizados con este nombre por Lyapunov que pretendió así destacar el papel *central* de estos resultados en la Teoría de la Probabilidad.

4.5. Teorema Central de Límite

Una aplicación inmediata es el Teorema de De Moivre-Laplace, una versión temprana del TCL, que estudia el comportamiento asintótico de una $B(n,p)$.

Teorema 4.11 (De Moivre-Laplace) *Sea $X_n \sim B(n,p)$ y definamos $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$. Entonces*

$$Z_n \xrightarrow{L} N(0,1).$$

Demostración.- Aplicando los resultados anteriores, se obtiene

$$\phi_{Z_n}(t) = \left((1-p)e^{-it\sqrt{\frac{p}{n(1-p)}}} + pe^{it\sqrt{\frac{(1-p)}{np}}}\right)^n,$$

que admite un desarrollo en serie de potencias de la forma

$$\phi_{Z_n}(t) = \left[1 - \frac{t^2}{2n}(1+R_n)\right]^n,$$

con $R_n \to 0$, si $n \to \infty$. En consecuencia,

$$\lim_{n\to\infty} \phi_{Z_n}(t) = e^{-\frac{t^2}{2}}.$$

La unicidad y el teorema de continuidad hacen el resto. ♠

Observación 4.2 *Lo que el teorema afirma es que si $X \sim B(n,p)$, para n suficientemente grande, tenemos*

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq x\right) \simeq \Phi(x),$$

donde $\Phi(x)$ es la función de distribución de la $N(0,1)$.

¿De qué forma puede generalizarse este resultado? Como ya sabemos $X_n \sim B(n,p)$ es la suma de n v. a. i.i.d., todas ellas Bernoulli ($Y_k \sim B(1,p)$), cuya varianza común, $var(Y_1) = p(1-p)$, es finita. En esta dirección tiene lugar la generalización: variables independientes, con igual distribución y con varianza finita.

Teorema 4.12 (Lindeberg) *Sean $X_1, X_2, \ldots, X_n$, v.a. i.i.d. con media y varianza finitas, μ y σ^2, respectivamente. Sea $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ su media muestral, entonces*

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{var(\overline{X}_n)}} \xrightarrow{L} N(0,1).$$

Demostración.- Teniendo en cuenta la definición de $\overline{X}_n$ podemos escribir

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{k=1}^n Z_k,$$

con $Z_k = (X_k - \mu)/\sigma$, variables aleatorias i.i.d. con $E(Z_1) = 0$ y $var(Z_1) = 1$. Aplicando Pϕ4 y (4.5) tendremos

$$\phi_{Y_n}(t) = \left[\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Pero existiendo los dos primeros momentos de Z_1 y teniendo en cuenta (4.4), $\phi_{Z_1}(t)$ puede también expresarse de la forma

$$\phi_{Z_1}(t) = 1 - \frac{t^2}{2n}(1 + R_n),$$

con $R_n \to 0$, si $n \to \infty$. En consecuencia,

$$\phi_{Y_n}(t) = \left[1 - \frac{t^2}{2n}(1 + R_n)\right]^n.$$

Así pues,

$$\lim_{n\to\infty} \phi_{Y_n}(t) = e^{-\frac{t^2}{2}},$$

que es la función característica de una $N(0,1)$. ♠

Observemos que el Teorema de De Moivre-Laplace es un caso particular del Teorema de Lindeberg, acerca de cuya importancia se invita al lector a reflexionar porque lo que en él se afirma es, ni más ni menos, que sea cual sea la distribución común a las X_i, su media muestral $\overline{X}_n$, adecuadamente tipificada, converge a una $N(0,1)$ cuando $n \to \infty$.

El teorema de Lindeberg, que puede considerarse el teorema central del límite básico, admite una generalización en la dirección de relajar la condición de equidistribución exigida a las variables. Las llamadas condiciones de Lindeberg y Lyapunov muestran sendos resultados que permiten eliminar aquella condición.

Ejemplo 4.5 (La fórmula de Stirling para aproximar n!) *Consideremos una sucesión de variables aleatorias $X_1, X_2, \ldots,$, independientes e idénticamente distribuidas, Poisson de parámetro $\lambda = 1$. La variable $S_n = \sum_{i=1}^n X_i$ es también Poisson con parámetro $\lambda_n = n$. Si $Z \sim N(0,1)$, para n suficientemente grande el TCL nos permite escribir,*

$$\begin{aligned}
P(S_n = n) &= P(n-1 < S_n \leq n) \\
&= P\left(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leq 0\right) \\
&\approx P\left(-\frac{1}{\sqrt{n}} < Z \leq 0\right) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-1/\sqrt{n}}^{0} e^{-x^2/2} dx \\
&\approx \frac{1}{\sqrt{2\pi n}},
\end{aligned}$$

en donde la última expresión surge de aproximar la integral entre $[-1/\sqrt{n}, 0]$ de $f(x) = e^{-x^2/2}$ mediante el área del rectángulo que tiene por base el intervalo de integración y por altura el $f(0) = 1$.

Por otra parte,

$$P(S_n = n) = e^{-n}\frac{n^n}{n!}.$$

Igualando ambos resultado y despejando n! se obtiene la llamada fórmula de Stirling,

$$n! \approx n^{n+1/2} e^{-n} \sqrt{2\pi}.$$

4.5.1. Una curiosa aplicación del TCL: estimación del valor de π

De Moivre y Laplace dieron en primer lugar una *versión local* del TCL al demostrar que si $X \sim B(n,p)$,

$$P(X=m)\sqrt{np(1-p)} \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \tag{4.12}$$

para n suficientemente grande y $x = \frac{m-np}{\sqrt{np(1-p)}}$. Esta aproximación nos va a servir para estudiar la credibilidad de algunas aproximaciones al número π obtenidas a partir del problema de la *aguja de Buffon*.

Recordemos que en el problema planteado por Buffon se pretende calcular la probabilidad de que una aguja de longitud l, lanzada al azar sobre una trama de paralelas separadas entre si una distancia a, con $a > l$, corte a alguna de las paralelas. Puestos de acuerdo sobre el significado de *lanzada al azar*, la respuesta es

$$P(corte) = \frac{2l}{a\pi},$$

resultado que permite obtener una aproximación de π si, conocidos a y l, sustituimos en $\pi = \frac{2l}{aP(corte)}$ la probabilidad de corte por su estimador natural *la frecuencia relativa de corte*, p, a lo largo de n lanzamientos. Podremos escribir, si en lugar de trabajar con π lo hacemos con su inverso,

$$\frac{1}{\pi} = \frac{am}{2ln},$$

donde m es el número de cortes en los n lanzamientos.

El año 1901 Lazzarini realizó 3408 lanzamientos obteniendo para π el valor 3,1415929 con ¡¡6 cifras decimales exactas!!. La aproximación es tan buena que merece como mínimo alguna pequeña reflexión. Para empezar supongamos que el número de cortes aumenta en una unidad, las aproximaciones de los inversos de π correspondientes a los m y $m+1$ cortes diferirían en

$$\frac{a(m+1)}{2ln} - \frac{am}{2ln} = \frac{a}{2ln} \geq \frac{1}{2n},$$

que si $n \approx 5000$, da lugar a $\frac{1}{2n} \approx 10^{-4}$. Es decir, un corte más produce una diferencia mayor que la precisión de 10^{-6} alcanzada. *No queda más alternativa que reconocer que Lazzarini tuvo la suerte de obtener* exactamente *el número de cortes, m, que conducía a tan excelente aproximación.* La pregunta inmediata es, *¿cuál es la probabilidad de que ello ocurriera?*, y para responderla podemos recurrir a (4.12) de la siguiente forma,

$$P(X=m) \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(m-np)^2}{2np(1-p)}} \leq \frac{1}{\sqrt{2\pi np(1-p)}}.$$

Por ejemplo, si $a = 2l$ entonces $p = 1/\pi$ y para $P(X=m)$ obtendríamos la siguiente cota

$$P(X=m) \leq \sqrt{\frac{\pi}{2n(\pi-1)}}.$$

Para el caso de Lazzarini n=3408 y $P(X=m) \leq 0{,}0146$, $\forall m$. *Parece ser que Lazzarini era un hombre de suerte, quizás demasiada.*

Capítulo 5

Simulación de variables aleatorias

5.1. Introducción

En el juego del dominó en sus distints versiones, cada jugador elige al azar 7 fichas de entre las 28. Calcular la probabilidad de ciertas composiciones de dichas 7 fichas puede ser tarea sencilla o, en ocasiones, prácticamente imposible por su complejidad. Así, si nos piden la probabilidad de que el jugador no tenga entre sus fichas ningún 6, la fórmula de Laplace nos da como resultado

$$p = \frac{\binom{21}{7}}{\binom{28}{7}} = 0,0982.$$

Si el jugador está jugando a *"la correlativa"*, le interesa saber con qué probabilidad, p_c, elegirá 7 fichas que puedan colocarse correlativamente una tras otra. Dicha probabilidad es matemáticamente intratable.

Un problema de imposibilidad práctica semejante nos ocurriría si quisiéramos calcular la probabilidad de tener un mazo de cartas ordenado de tal forma que permitiera hacer un solitario directamente. ¿Hemos de conformarnos con no poder dar respuesta a situaciones como las planteadas u otras similares que puedan surgir?

La ley fuerte de los grandes números, LFGN, (ver Teorema 4.4 en la página 101) y la potencia de cálculo de los ordenadores de sobremesa actuales no permiten abordar el problema empíricamente. En efecto, si podemos simular la elección al azar de 7 fichas de entre las 28 y comprobar después si es posible colocarlas correlativamente una tras otra (éxito), repitiendo el proceso de simulación definiremos una variable X_i ligada a la simulación *i-ésima* que tomará valores

$$X_i = \begin{cases} 1, & \text{si la colocación es posible;} \\ 0, & \text{en caso contrario.} \end{cases}$$

Así definida, $X_i \sim B(1, p_c)$, y es independiente de cualquier otra X_j. Si llevamos a cabo n simulaciones, por la LFGN se verificará

$$\sum_{k=1}^{n} \frac{X_i}{n} \xrightarrow{a.s.} p_c,$$

y podremos aproximar el valor de p_c mediante la frecuencia relativa de éxitos que hayamos obtenido.

Un programa adecuado en cualquiera de los lenguajes habituales de programación permitirá averiguar si las 7 fichas son correlativas, pero ¿cómo simular su elección aleatoria entre las 28 fichas del dominó? Podríamos proceder como sigue:

1. ordenamos las fichas con algún criterio, por ejemplo, comenzar por las blancas ordenadas según el otro valor que las acompaña, continuar por los 1's ordenados según el valor que les acompaña y así sucesivamente hasta finalizar con el 6 doble,
2. las numeramos a continuación del 1 al 28, y
   a) extraemos al azar, sin repetición, 7 números del 1 al 28, las correspondientes fichas constituirán nuestra elección; o bien,
   b) permutamos aleatoriamente los 28 números y las fichas correspondientes a los 7 primeros constituirán nuestra elección.

Queda, no obstante, por resolver la forma de simular la extracción de los 7 números o de permutar aleatoriamente los 28. En cualquier caso necesitamos poder generar con el ordenador números aleatorios.

5.2. Generación de números aleatorios

La generación de números aleatorios con el ordenador consiste en una subrutina capaz de proporcionar valores de una variable aleatoria $X \sim U(0,1)$. Cabe preguntarse si una sucesión de valores obtenida de semejante forma es aleatoria. La respuesta es inmediata y decepcionante, no. En realidad, los números generados mediante cualquiera de las subrutinas disponibles a tal efecto en los sistemas operativos de los ordenadores podemos calificarlos como *pseudoaleatorios* que, eso sí, satisfacen los contrastes de uniformidad e independencia.

La mayoría de los generadores de números aleatorios son del tipo *congruencial*. A partir de un valor inicial, X_0, denominado semilla, y enteros fijos no negativos, a, c y m, calculan de forma recursiva

$$X_{i+1} = aX_i + c \ \ (\text{módulo } m), \qquad i = 1, 2, \ldots, n. \tag{5.1}$$

Como valores de la $U(0,1)$ se toman los

$$U_i = \frac{X_i}{m}.$$

La expresión (5.1) justifica la denominación de pseudoaleatorios que les hemos otorgado. En primer lugar, porque si iniciamos la generación siempre con el mismo valor de X_0 obtendremos la misma sucesión de valores y, claro está, nada más alejado de la aleatoriedad que conocer de antemano el resultado. En segundo lugar, porque el mayor número de valores distintos que podemos obtener es precisamente m. Se deduce de aquí la conveniencia de que m sea grande.

Puede demostrarse que una elección adecuada de a, c y m proporciona valores que, aun no siendo aleatorios, se comportan como si lo fueran a efectos prácticos porque, como ya hemos dicho, satisfacen las condiciones de uniformidad e independencia.

Si, como es nuestro caso en el ejemplo del dominó, deseamos elegir al azar números entre 1 y 28, recordemos que si $U_i \sim U(0,1)$, $kU_i \sim U(0,k)$ y por tanto,

$$P(j-1 < kU_i \leq j) = \frac{1}{k}, \quad j = 1, 2, \ldots, k.$$

Si definimos $N_k = [kU_i] + 1$, donde $[\,\cdot\,]$ representa la parte entera del argumento, N_k es una variable discreta que tiene la distribución deseada, uniforme en $\{1, 2, \ldots, k\}$.

5.3. Técnicas generales de simulación de variables aleatorias

La generación de una variable $U(0,1)$ es un primer paso necesario pero no suficiente, como el ejemplo del dominó ha puesto de manifiesto. Nuestro objetivo era generar valores de una uniforme discreta sobre el soporte $\{1,2,\ldots,k\}$ y lo hemos conseguido previa simulación de una $U(0,1)$. En muchos procesos industriales o en el estudio de fenómenos experimentales, un estudio previo de su comportamiento requiere simular cantidades aleatorias que siguen, por lo general, una distribución diferente de la uniforme. ¿Cómo hacerlo? A continuación presentamos tres métodos para simular variables aleatorias.

Antes de responder digamos que el origen de la simulación de variables aleatorias, conocida como *simulación de Montecarlo*, se debe a von Neumann y Ulam que lo utilizaron por primera vez durante la segunda guerra mundial para simular el comportamiento de la difusión aleatoria de neutrones en la materia fisionable. El trabajo, relacionado con la fabricación de la primera bomba atómica, se desarrollaba en el laboratorio de Los Álamos y Montecarlo fue el código secreto que ambos físicos le dieron.

5.3.1. Método de la transformación inversa

Este método se basa en el resultado de la Proposición 2.3 de la página 59. Si $U \sim U(0,1)$, F es una función de distribución de probabilidad y definimos

$$F^{-1}(x) = \text{ínf}\{t:\ F(t) \geq x\},$$

la variable $X = F^{-1}(U)$ verifica que $F_X = F$.

Este resultado permite generar valores de una variable aleatoria con función de distribución dada, sin más que obtener la antiimagen mediante F de valores de una $U(0,1)$. Veamos como generar una exponencial de parámetro λ.

Ejemplo 5.1 (Generación de $X \sim Exp(\lambda)$) *La densidad de una Exponencial de parámetro λ es*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0; \\ 0, & \text{en el resto,} \end{cases}$$

y su función de distribución,

$$F(x) = \begin{cases} 0, & x < 0; \\ 1 - e^{-\lambda x}, & x \geq 0;. \end{cases}$$

De aquí,

$$1 - e^{-\lambda X} = U \implies X = -\frac{1}{\lambda}\ln(1-U),$$

nos permite generar valores de una $Exp(\lambda)$. Observemos que $1-U$ es también $U(0,1)$ con lo que

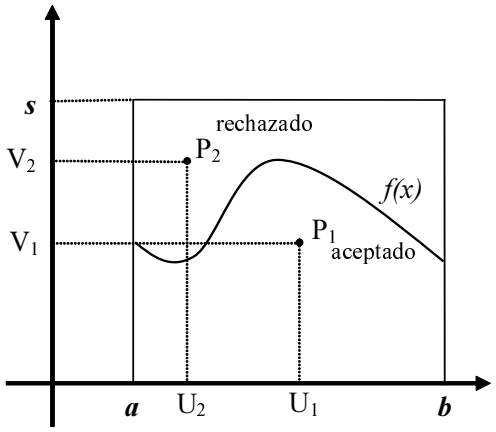
$$X = -\frac{1}{\lambda}\ln U,$$

genera también valores de una $Exp(\lambda)$.

Recordemos que si $Y \sim Ga(n, 1/\lambda)$ entonces Y es la suma de n *exponenciales independientes de parámetro λ (página 64). Aplicando el resultado anterior podremos generar valores de una $Ga(n, 1/\lambda)$ a partir de* n *$U(0,1)$, $U_1, U_2, \ldots, U_n$ mediante la expresión*

$$Y = -\sum_{i=1}^{n} \frac{1}{\lambda}\ln U_i = -\frac{1}{\lambda}\ln\left(\prod_{i=1}^{n} U_i\right).$$

5.3.2. Método de aceptación-rechazo



Si queremos generar una variable aleatoria X con densidad $f(x)$, cuyo soporte es el intervalo $[a,b]$, podemos proceder tal como ilustra la figura. Si s es una cota superior de $f(x)$, $f(x) \leq s$, $\forall x \in [a,b]$, generamos aleatoriamente un punto (U,V) en el rectángulo $[a,b] \times [0,s]$. Si $V \leq f(U)$ hacemos $X = U$, en caso contrario rechazamos el punto y generamos uno nuevo.

La primera pregunta que surge es si el número de simulaciones necesarias para obtener un valor de X será finito. Veamos que lo es con probabilidad 1. Para ello, designemos por A la sombra de $f(x)$ y por $\{P_n, n \geq 1\}$ una sucesión de puntos elegidos al azar en $[a,b] \times [0,s]$. La probabilidad de que uno cualquiera de ellos, P_i, no cumpla con la condición $V \leq f(U)$, equivale a $\{P_i \notin A\}$ y vale

$$P(P_i \notin A) = \frac{s(b-a) - |A|}{s(b-a)} = 1 - \frac{1}{s(b-a)},$$

pues $|A| = \int_a^b f(x)dx = 1$. Como la elección es al azar, la probabilidad de que ninguno de ellos esté en A,

$$P\left(\cap_{n\geq 1}\{P_n \notin A\}\right) = \lim_{n\to\infty} \left(1 - \frac{1}{s(b-a)}\right)^n = 0,$$

y por tanto con probabilidad 1 el número necesario de simulaciones para generar un valor de X será finito.

Queda ahora por comprobar que X se distribuye con densidad $f(x)$. En efecto,

$$P(X \leq x) = \sum_{n\geq 1} P(X \leq x, X = U_n),$$

donde

$$\{X \leq x, X = U_n\} = \{P_1 \notin A, \ldots, P_{n-1} \notin A, P_n \in [a,x] \times [0,s].$$

Al tomar probabilidades,

$$\begin{aligned} P(X \leq x) &= \sum_{n\geq 1} P(X \leq x, X = U_n) \\ &= \sum_{n\geq 1} \left(1 - \frac{1}{s(b-a)}\right)^{n-1} \frac{\int_a^x f(x)dx}{s(b-a)} \\ &= \int_a^x f(x)dx. \end{aligned}$$

Ejemplo 5.2 *Una variable $X \sim Be(4,3)$ tiene por densidad,*

$$f(x) = \begin{cases} 60x^3(1-x)^2, & 0 \leq x \leq 1; \\ 0, & \textit{en el resto.} \end{cases}$$

Su función de distribución será un polinomio de grado 6, lo que dificulta la generación de una variable aleatoria con esta distribución mediante el método de la transformación inversa. Podemos, en su lugar, recurrir al método de aceptación-rechazo. Tomemos para ello $s = \max_{0\leq x\leq 1} f(x) = f(0{,}6) = 2{,}0736$. *El procedimiento consistirá en generar*

$$(U, V) \sim U([0, 1] \times [0; 2{,}0736]),$$

pero como $kV \sim U(0, k)$ *si* $V \sim U(0, 1)$, *bastará con generar sendas* $U(0, 1)$ *y hacer* $X = U$ *si*

$$2{,}0736V \leq 60U^3(1-U)^2 \implies V \leq \frac{60U^3(1-U)^2}{2{,}0736}.$$

Generalización del método de aceptación-rechazo

El método anterior puede generalizarse utilizando una función $t(x)$ que mayore a $f(x)$, $\forall x$, en lugar de una cota superior. Si $t(x)$ es tal que

$$1 \leq \int_{-\infty}^{\infty} t(x)dx = c < +\infty,$$

la función $g(x) = t(x)/c$ es una densidad. El método exige que podamos generar con facilidad una variable aleatoria Y con densidad $g(x)$ y, en ese caso, los pasos a seguir son,

1. Generamos Y cuya densidad es $g(y) = t(y)/c$.
2. Generamos $U \sim U(0, 1)$.
3. Si $U \leq f(Y)/t(Y)$, hacemos $X = Y$, en caso contrario reiniciamos el proceso desde 1.

La X así generada tiene por densidad $f(x)$ porque

$$\{X \leq x\} = \{Y_N \leq x\} = \{Y \leq x | U \leq f(Y)/t(Y)\},$$

donde N designa el número de la iteración en la que se ha cumplido la condición. Al tomar probabilidades

$$P(X \leq x) = P(Y \leq x | U \leq f(Y)/t(Y)) = \frac{P(Y \leq x, U \leq f(Y)/t(Y))}{P(U \leq f(Y)/t(Y))}.$$

Como Y y U son independientes, su densidad conjunta será

$$h(y, u) = g(y), \quad 0 \leq u \leq 1, \quad -\infty < y < \infty.$$

De aquí,

$$\begin{aligned} P(X \leq x) &= \frac{1}{P(U \leq f(Y)/t(Y))} \int_{-\infty}^{x} g(y) \left[\int_0^{f(y)/cg(y)} du \right] dy \\ &= \frac{1}{cP(U \leq f(Y)/t(Y))} \int_{-\infty}^{x} f(y)dy. \end{aligned} \tag{5.2}$$

Pero

$$\lim_{x\to\infty} P(X \leq x) = 1 = \frac{1}{cP(U \leq f(Y)/t(Y))} \int_{-\infty}^{\infty} f(y)dy = \frac{1}{cP(U \leq f(Y)/t(Y))}. \tag{5.3}$$

Sustituyendo en (5.2),

$$P(X \leq x) = \int_{-\infty}^{x} f(y)dy.$$

Respecto al número N de iteraciones necesarias para obtener un valor de X, se trata de una variable geométrica con $p = P(U \leq f(Y)/t(Y))$. De (5.3) deducimos que $P(U \leq f(Y)/t(Y)) = 1/c$ y $E(N) = c$, además

$$P(N = \infty) = \lim_{n\to\infty} P(N > n) = \lim_{n\to\infty} \sum_{j\geq n+1} \frac{1}{c}\left(1 - \frac{1}{c}\right)^{j-1} = \lim_{n\to\infty} \left(1 - \frac{1}{c}\right)^{n} = 0,$$

con lo que N será finito con probabilidad 1.

Ejemplo 5.3 (Simulación de una N(0,1). Método general de aceptación-rechazo) *Recordemos en primer lugar la densidad de $Z \sim N(0,1)$,*

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

Si definimos $X = |Z|$, su densidad será

$$f_X(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty. \tag{5.4}$$

Vamos a utilizar el método general de aceptación-rechazo para simular valores de la N(0,1) utilizando como paso previo la generación de valores de X con densidad (5.4). Para ello tomaremos como función auxiliar $t(x) = \sqrt{2e/\pi}e^{-x}$, $x \geq 0$, porque mayora a $f_X(x)$ en todo su dominio,

$$\frac{f_X(x)}{t(x)} = \frac{\sqrt{2/\pi}e^{-x^2/2}}{\sqrt{2e/\pi}e^{-x}} = \exp\left\{-\frac{(x-1)^2}{2}\right\} \leq 1, \quad x \geq 0.$$

Por otra parte,

$$\int_0^{\infty} t(x)dx = \sqrt{\frac{2e}{\pi}},$$

y por tanto

$$g(x) = \frac{t(x)}{\sqrt{2e/\pi}} = e^{-x},$$

que es la densidad de una Exponencial con parámetro $\lambda = 1$. De acuerdo con el procedimiento antes descrito,

1. *Generaremos sendas variables aleatorias, Y y U, la primera $Exp(1)$ y la segunda $U(0,1)$.*
2. *Si*

$$U \leq \frac{f_X(Y)}{t(Y)} = \exp\left\{-\frac{(Y-1)^2}{2}\right\},$$

haremos $X = Y$, en caso contrario comenzaremos de nuevo.

Una vez obtenido un valor de X el valor de Z se obtiene haciendo $Z = X$ o $Z = -X$ con probabilidad 1/2.

El procedimiento anterior puede modificarse si observamos que aceptamos Y si

$$U \leq \exp\{-(Y-1)^2/2\} \iff -\ln U \geq (Y-1)^2/2.$$

Pero de acuerdo con el ejemplo 5.1 $-\ln U \sim Exp(1)$ con lo que la generación de valores de una $N(0,1)$ se pude llevar a cabo mediante,

1. *Generaremos sendas variables aleatoria Y_1, Y_2, ambas $Exp(1)$.*

2. *Si $Y_2 \geq (Y_1 - 1)^2/2$, hacemos $X = Y_1$, en caso contrario comenzaremos de nuevo.*

3. *Generamos $U \sim U(0,1)$ y hacemos*

$$Z = \begin{cases} X, & si\ U \leq 1/2; \\ -X, & si\ U > 1/2. \end{cases}$$

Ejemplo 5.4 (Simulación de una N(0,1). El método polar) *Si $X \sim N(0,1)$ e $Y \sim N(0,1)$ y son independientes, su densidad conjunta viene dada por*

$$f_{XY}(x,y) = \frac{1}{2\pi} \exp\left\{-\frac{x^2+y^2}{2}\right\}.$$

Consideremos ahora las coordenadas polares del punto (X,Y), $R = \sqrt{X^2+Y^2}$ y $\Theta = \arctan Y/X$. Para obtener su densidad conjunta, necesitamos las transformaciones inversas, $X = R\cos\Theta$ e $Y = R\sin\Theta$. El correspondiente jacobiano vale $J_1 = R$ y su densidad conjunta,

$$f_{R\Theta}(r,\theta) = \begin{cases} \dfrac{1}{2\pi} r \exp\left\{-\dfrac{r^2}{2}\right\}, & r \geq 0,\ 0 \leq \theta \leq 2\pi, \\ \\ 0, & en\ el\ resto. \end{cases}$$

De aquí se obtienen fácilmente las densidades marginales de R^2 y Θ que resultan ser $R^2 \sim Exp(1/2)$ y $\Theta \sim U(0,2\pi)$. Haciendo uso del resultado del ejemplo 5.1, si generamos dos variables $U80,1)$, U_1 y U_2,

$$R^2 = -2\ln U_1 \sim Exp(1/2)$$
$$\Theta = 2\pi U_2 \sim U(0,2\pi),$$

y de aquí

$$X = (-2\ln U_1)^{1/2}\cos 2\pi U_2$$
$$Y = (-2\ln U_1)^{1/2}\sin 2\pi U_2,$$

son $N(0,1)$ e independientes.

5.3.3. Simulación de variables aleatorias discretas

El método más extendido para simular variables discretas es el de la transformación inversa. Si la variable X cuyos valores queremos simular tiene por función de distribución $F(x)$, entonces

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p_i,$$

con $x_i \in D_X$, el soporte de la variable, y $p_i = P(X = x_i)$.

Si suponemos que los valores del soporte están ordenados, $x_1 < x_2 < \ldots$, el algoritmo consiste en los pasos siguientes:

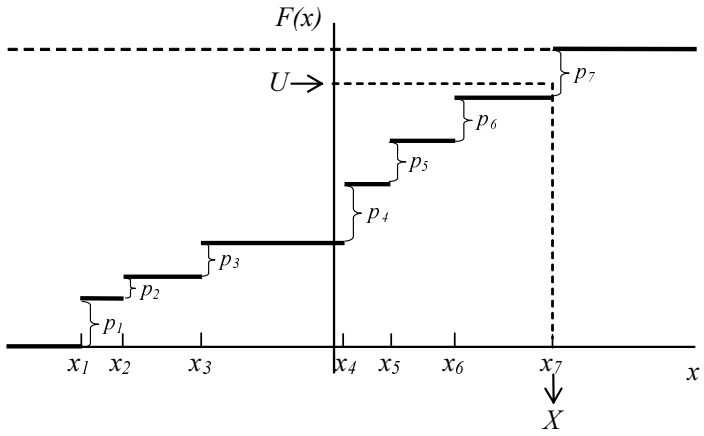
1. Generamos $U \sim U(0,1)$.

2. Hacemos $X = k$, con $k = \text{mín}\{i; U \leq F(x_i)\}$.

La variable así generada se distribuye como deseamos. En efecto,

$$P(X = x_i) = P\left(F(x_{i-1} < U \leq F(x_i)\right) = F(x_i) - F(x_{i-1}) = p_i.$$

La figura ilustra gráficamente el algoritmo para una variable discreta con soporte $D_X = \{x_1, x_2, \ldots, x_7\}$.



A continuación presentamos la adaptación de este algoritmo para la simulación de las variables discretas más conocidas. En algunos casos, el algoritmo no resulta reconocible porque la forma de la función de cuantía permite formas de búsqueda más ventajosas que lo enmascaran.

Ejemplo 5.5 (Simulación de una variable Bernoulli) *Si $X \sim B(1,p)$, el algoritmo que sigue es muy intuitivo y equivale al algoritmo de la transformación inversa si intercambiamos U por $1-U$.*

1. *Generamos $U \sim U(0,1)$.*
2. *Hacemos*

$$X = \begin{cases} 1, & si\ U \leq p; \\ 0, & si\ U > p. \end{cases}$$

Ejemplo 5.6 (Simulación de una variable uniforme) *Para generar X con $D_X = \{x_1, x_2, \ldots, x_n\}$ con $P(X = x_i) = 1/n$, $\forall i$,*

1. *Generamos $U \sim U(0,1)$.*
2. *Hacemos $X = x_k$, con $k = 1 + \lfloor nU \rfloor$ donde $\lfloor s \rfloor$, es la parte entre por defecto de* s.

Ejemplo 5.7 (Simulación de una variable Binomial) *Si $X \sim B(n,p)$, entonces X es la suma de* n *Bernoullis independientes con igual parámetro, bastará por tanto repetir* n *veces el algoritmo del ejemplo 5.5 y tomar $X = \sum_{i=1}^{n} X_i$, la suma de los* n *valores generados*

Ejemplo 5.8 (Simulación de una variable Geométrica) *Recordemos que $X \sim Ge(p)$, X es el número de pruebas Bernoulli necesarias para alcanzar el primer éxito, de aquí, $P(X = i) = p(1-p)^{i-1}$, $i \geq 1$, y $P(X > i) = \sum_{j>i} P(X = j) = (1-p)^i$. Tendremos que*

$$F(k-1) = P(X \leq k-1) = 1 - P(X > k-1) = 1 - (1-p)^{k-1}.$$

Aplicando el algoritmo de la transformación inversa

1. *Generamos* $U \sim U(0,1)$.
2. *Hacemos* $X = k$, *tal que* $1-(1-p)^{k-1} < U \leq 1-(1-p)^k$, *que podemos también escribir* $(1-p)^k < 1-U \leq (1-p)^{k-1}$. *Como* $1-U \sim U(0,1)$, *podemos definir* X *mediante*

$$\begin{aligned} X &= \text{mín}\{k; (1-p)^k < U\} \\ &= \text{mín}\{k; k\ln(1-p) < \ln U\} \\ &= \text{mín}\left\{k; k > \frac{\ln U}{\ln(1-p)}\right\}. \end{aligned}$$

El valor que tomaremos para X *será,*

$$X = 1 + \left\lfloor \frac{\ln U}{\ln(1-p)} \right\rfloor. \tag{5.5}$$

Ejemplo 5.9 (Simulación de una variable Binomial Negativa) *Para generar valores de* $X \sim BN(r,p)$ *recordemos que* X *puede expresarse como suma de* r *variables independientes todas ellas Geométricas de parámetro* p *(véase el Ejemplo 3.2 de la página 81). Bastará por tanto utilizar el siguiente algoritmo:*

1. *Simulamos* $X_1, X_2, \ldots, X_r$ *todas ellas* $Ge(p)$, *utilizando para ello la expresión (5.5).*
2. *hacemos* $X = X_1 + X_2 + \ldots + X_r$.

Ejemplo 5.10 (Simulación de una variable Poisson) *La simulación de una variable* $X \sim Po(\lambda)$ *se basa en la propiedad que liga la distribución de Poisson con la Exponencial de igual parámetro. Como señalábamos en la página 31, al estudiar la ocurrencia de determinado suceso a largo del tiempo, por ejemplo las partículas emitidas por un mineral radiactivo durante su desintegración, el tiempo que transcurre entre dos ocurrencias consecutivas es una variable aleatoria* $Y \sim Exp(\lambda)$ *y el número de ocurrencias que se producen en el intervalo [0,t] es* $X_t \sim Po(\lambda t)$, *donde* λ *es el número medio de ocurrencias por unidad de tiempo. De acuerdo con esto, el algoritmo de simulación es el siguiente:*

1. *Simulamos* $U_1, U_2, U_3, \ldots$, *todas ellas* $U(0,1)$.
2. *Hacemos* $X = N - 1$, *donde*

$$N = \text{mín}\left\{n; \prod_{i=1}^{n} U_i < e^{-\lambda}\right\}.$$

La variable así obtenida se distribuye como una $Po(\lambda)$. *En efecto,*

$$\begin{aligned} X+1 &= \text{mín}\left\{n; \prod_{i=1}^{n} U_i < e^{-\lambda}\right\}. \\ X &= \text{máx}\left\{n; \prod_{i=1}^{n} U_i \geq e^{-\lambda}\right\} \\ &= \text{máx}\left\{n; \sum_{i=1}^{n} \ln U_i \geq -\lambda\right\} \\ &= \text{máx}\left\{n; \sum_{i=1}^{n} -\ln U_i \leq \lambda\right\}, \end{aligned}$$

pero como vimos en el Ejemplo 5.1, $-\ln U_i \sim Exp(1)$*, con lo que* X *puede interpretarse com el mayor número de* $Exp(1)$ *cuya suma es menor que* λ*. Ahora bien, exponenciales independientes de parámetro 1 equivalen a tiempos de espera entre ocurrencias de un suceso con una ratio media de ocurrencias de 1 suceso por unidad de tiempo, lo que permite interpretar* X *como el número y. Así,* $X \sim Po(\lambda)$*.*