Tema 3

Variables aleatorias

3.1 De la medida en espacios de probabilidad

En este capítulo vamos a estudiar **variables aleatorias**. De nuevo, sabiendo que estamos ante alumnos de Informática, sentimos la necesidad de justificar para qué y por qué estudiamos tal concepto. Podemos justificar su estudio desde diversos ángulos. Uno podría ser el siempre imprescindible y convincente de las aplicaciones; ¿para qué estudiar, si no, algo que no es relevante para la disciplina por la que uno siente **pasión**? Otro ángulo igualmente importante pero menos evidente a primera vista es el conceptual; ¿para qué estudiar algo que no tiene envergadura intelectual ni profundidad? ¡Abajo con lo irrelevante! Vamos a dar inmediatamente la definición de variable aleatoria y a partir de ella discutir estos dos ángulos.

Hasta ahora nos hemos centrado en las probabilidades de sucesos (en especial, de sucesos finitos) y en general en cómo definir dichas probabilidades. En su momento, examinamos el comportamiento en media del algoritmo de ordenación *quicksort*, la paradoja de Monty Hall, la regla de Laplace, las probabilidades condicionadas, el teorema de Bayes, la aplicación de este teorema al diagnóstico médico, entre otros. El siguiente paso es medir sobre un espacio de probabilidad una cantidad de interés. Dicha necesidad de medir estuvo motivada en su origen por las aplicaciones y la definición de variable aleatoria no viene más que a formalizar dicha idea. He aquí su definición formal; la figura 3.1 muestra una interpretación gráfica de esa definición (A es un suceso del espacio E en la figura).

Definición 3.1.1 Sea (E, Ω, P) un espacio de probabilidad. Una variable aleatoria sobre un espacio de probabilidad es una función definida sobre el espacio de resultados E.

Es una definición sencilla, el lector no podrá sino coincidir con nosotros en esto, pero ¿por qué es necesaria? Acudamos por un instante a la definición de matemáticas para explicar esa necesidad. Aunque no hay consenso definitivo sobre la definición de las matemáticas —y probablemente nunca lo habrá—, las definiciones modernas[1] se podrían sintetizar diciendo que las matemáticas son el **estudio abstracto de la cantidad, la estructura, los patrones, el espacio y el cambio**. Y es en el primer término, la cantidad, y en el último, el cambio, que encontramos las razones de la importancia de las variables aleatorias. En Informática necesitamos medir, y **medir en presencia de la incertidumbre**, porque los fenómenos son lo suficientemente complejos y porque la cantidad de datos es inmanejablemente grande.

[1]Esta definición está tomada de la Wikipedia: https://en.wikipedia.org/wiki/Mathematics.

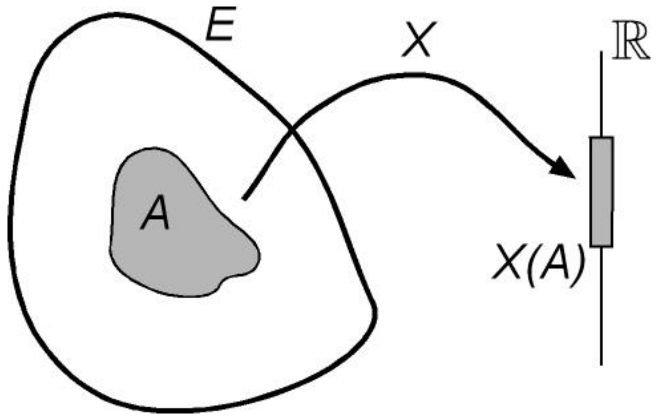

Figura 3.1: Esquema de la definición de variable aleatoria

Ejemplos de medidas relevantes en informática son el tiempo de ejecución de algoritmos, los códigos autocorrectores, los paquetes corruptos en redes, la sobrecarga de servidores web, los fallos en los discos duros, el reconocimiento óptico de textos, la estimación del error en la conversión analógica-digital, entre otros[2]. Esta lista debería ser suficiente en cuanto a las aplicaciones, pero el lector puede investigar más por su cuenta y convencerse rápidamente de que las variables aleatorias son ubicuas. Por otro lado, ese afán de medir está incluido en el mencionado estudio de la cantidad y el cambio del que se hablaba en la definición de las matemáticas. Por tanto, las variables aleatorias surgen de la necesidad de medir sobre espacios de probabilidad. Cuando se mide sobre conjuntos normales las medidas reciben el nombre de **funciones**. El hecho de que una variable aleatoria sea la generalización del concepto función ya justifica su importancia teórica.

Volvamos a la definición de variable aleatoria, que indudablemente merece más análisis por nuestra parte. Esencialmente, es una función definida sobre un espacio de probabilidad. Recordamos que una función $f : A \longrightarrow B$ es una correspondencia tal que todo elemento $a \in A$ tiene una única imagen $f(a) \in B$, donde B es un conjunto de números. Vemos que la definición de función requiere que la medida sea única, que es algo lógico e imprescindible.

3.2 Variables aleatorias

Ilustraremos con un ejemplo la definición de variable aleatoria dada más arriba. Se tiran dos dados no cargados y se observan los puntos que aparecen cara arriba (este ejemplo ya ha sido desarrollado sin variables aleatorias en el ejemplo 2.4.15). Para este experimento, el espacio muestral es $E = A \times A$, donde $A = \{1, \dots, 6\}$. El espacio de probabilidad es la terna (E, Ω, P), donde $\Omega = \mathcal{P}(\Omega)$ y la probabilidad de cualquier suceso elemental (a, b), con $a, b \in A$, es $1/36$. Definamos una variable aleatoria X sobre este espacio de probabilidad. Supongamos que no nos interesan los puntos que salen en los dados, sino su suma, que es una función sobre E. Ahora tenemos la **variable aleatoria** $X :\longrightarrow \mathbb{N}$. Esta variable aleatoria produce un **nuevo espacio de probabilidad**, que designaremos por (B, Ω_B, P_X), donde:

[2]Esta lista está tomada del artículo en que se describe el contenido del libro *A course on probability theory for computer scientists*, de Mehran Sahami (Universidad de Stanford); véase http://dl.acm.org/citation.cfm?id=1953245. Solo se han tomado los ejemplos más relevantes.

- B es la imagen de E por X, esto es, el conjunto $B = \{X(e) \in B \mid e \in E\}$;
- Ω_B o los sucesos en B son las partes de $\mathcal{P}(B)$;
- y finalmente la función de probabilidad se define por $P_X(b) = P\left(\{e \in E \,|\, X(e) = b\}\right)$.

En nuestro caso concreto, del dado, la imagen de B o el conjunto de valores que toma X es el conjunto $B = \{2, 3, \ldots, 12\}$, que son las posibles sumas de las caras del dado. Como conjunto de sucesos tomamos las partes de B, es decir, cualquier subconjunto de B es un suceso. Y, por último, las probabilidades de cada valor que toma X son las siguientes:

$$\begin{array}{lll}
P_X(2) = \frac{1}{36} & P_X(6) = \frac{5}{36} & P_X(10) = \frac{3}{36} \\
P_X(3) = \frac{2}{36} & P_X(7) = \frac{6}{36} & P_X(11) = \frac{2}{36} \\
P_X(4) = \frac{3}{36} & P_X(8) = \frac{5}{36} & P_X(12) = \frac{1}{36} \\
P_X(5) = \frac{4}{36} & P_X(9) = \frac{4}{36} &
\end{array}$$

Nota 3.2.1 Formalmente, las probabilidades P y P_X son dos funciones diferentes definidas sobre dos espacios muestrales diferentes, E y $X(E)$. No obstante, se suele indicar ambas probabilidades simplemente con la letra P. No hay confusión posible con esta identificación, pues los sucesos determinan a qué probabilidad nos referimos. Nosotros seguiremos esta convención y usaremos P para ambas probabilidades.

Ejercicio 3.2.2 Justificad de dónde salen las probabilidades anteriores.

Ejercicio 3.2.3 Se tiran tres monedas no cargadas. Se quiere estudiar la variable aleatoria número de cruces en las tres tiradas. Identificad el espacio de probabilidad asociado a la variable aleatoria y calculad sus probabilidades.

Ejercicio 3.2.4 Un experimento consiste en tirar un dado no cargado repetidamente. Definimos la variable aleatoria número de tiradas hasta que sale 5 (incluida esta última). Construid el espacio de probabilidad de esta variable aleatoria.

Las variables aleatorias se clasifican en dos grupos diferentes que determinan sus características y cálculo. Si una variable aleatoria toma valores en un conjunto finito o en una sucesión infinita, entonces se dice que es una **variable aleatoria discreta**. Si toma valores en todos los puntos de un intervalo, se dice que es una **variable aleatoria continua**. Esencialmente, los datos discretos se cuentan y los datos continuos se miden.

Ejercicio 3.2.5 Comprobad que las variables aleatorias de los tres últimos ejemplos son discretas.

Ejercicio 3.2.6 Proporcionad un ejemplo de vuestra invención de una variable aleatoria continua. Justificad que es así.

La diferencia entre las variables discretas y las continuas es tan marcada que cada uno tiene un estudio por separado. Empezamos con las variables aleatorias discretas.

3.2.1 Variables aleatorias discretas

Consideremos un espacio de probabilidad (E, Ω, P), donde el espacio muestral sea discreto. Estos espacios de probabilidad se caracterizan directamente por su **función de masa**, de la que damos su definición formal a continuación.

Definición 3.2.7 Sea un espacio de probabilidad (E, Ω, P), donde E es un conjunto discreto, pongamos $E = \{e_i \,|\, i \in I\}$, siendo I un subconjunto de $\mathbb{N}$. Dada la variable aleatoria X, su función de masa son las probabilidades $p_i = P(X(e_i))$, donde $i \in I$.

A la pareja (e_i, p_i) se le suele llamar **distribución de probabilidad**.

Ejercicio 3.2.8 ¿Podría explicar el lector por qué en la definición anterior aparece la expresión "siendo I un subconjunto de $\mathbb{N}$"? ¿Está seguro el lector de que comprende la definición de conjunto discreto? Si no es así, búsquelo en internet.

Aplicando los axiomas de la probabilidad, los números $p_i, i \in I$ cumplen dos propiedades inmediatas, a saber: (1) $p_i \geq 0$, para todo $i \in I$; (2) la suma de todos los p_i da 1. De hecho, el recíproco es cierto y todo conjunto de números que cumplen (1) y (2) son función de masa de alguna variable aleatoria.

Ejercicio 3.2.9 Se tienen los siguientes conjuntos de números p_i. Comprobad si son realmente una función de masa.

(a) $p_n = \dfrac{1}{6}\left(\dfrac{5}{6}\right)^n$, donde $n = 0, 1, 2, \ldots$;

(b) $\{p_n\}$ es el conjunto $\left\{\dfrac{37}{57}, \dfrac{43}{48}, \dfrac{5}{130}, \dfrac{31}{111}, 0, 0\right\}$;

(c) $p_n = \dfrac{1}{2}\left(\dfrac{3}{4}\right)^n$, donde $n = 1, 2, 3, \ldots$.

Ejercicio 3.2.10 Un envío de 8 ordenadores contienen 3 que son defectuosos. Una empresa compra 2 de estos 8 ordenadores. Hallad la función de masa para la variable aleatoria número de ordenadores defectuosos en la compra.

Problema 3.2.11 Un jugador extrae dos bolas a la vez de una urna que contiene 3 blancas y 2 verdes. El jugador pierde 3 € si saca las dos bolas blancas, gana 1 € si obtiene una bola blanca y otra verde y queda en paz en otro caso. Se pide:

(a) Definir una variable aleatoria X que describa la ganancia del jugador. Hallar su función de masa y representarla.

(b) Hallar la distribución de probabilidad de X.

(c) Calcular $P(-3 \leq X \leq 0)$ y $P(-4 < X < 0)$.

Problema 3.2.12 Sea p_n, con $n \in \mathbb{N}$, una función de masa de una variable aleatoria. Probad que $\lim_{n\to\infty} p_n = 0$.

3.2.2 Variables aleatorias continuas

El lector se podría estar preguntando en este momento por qué es necesaria la distinción entre variables discretas y continuas. La razón es técnica y a la vez profunda. Para decirlo con contundencia y presteza, la razón es que todos los infinitos no son iguales. Cuando consideramos una variable discreta, esta puede tomar valores en un conjunto finito o infinito del tipo de $\mathbb{N}$ (llamado **infinito numerable**). Con este tipo de conjuntos se pueden sumar probabilidades del tipo $\sum_{n=0}^{\infty} p_n$, es decir, por series. Sin embargo, si la variable toma valores en un intervalo, entonces resulta que el infinito del intervalo es más grande que el infinito numerable[3]. Ello hace que no sea posible tomar los sumatorios como en el caso discreto; hay que acudir a otro mecanismo matemático más general: las integrales. Los sumatorios del mundo discreto serán las integrales en el mundo continuo y lo que eran funciones de masa definidas sobre puntos aislados de un espacio discreto ahora serán funciones definidas en $\mathbb{R}$ o en algún intervalo suyo.

La introducción de las variables continuas en la probabilidad permitió resolver muchos problemas prácticos, sobre todo ciertos problemas geométricos. Posteriormente, se encontraron aplicaciones en multitud de contextos, incluida la informática. En informática gráfica, los métodos de Monte Carlo y la generación de números aleatorios son dos de los campos de más frecuente e intensa aplicación.

Una variable aleatoria continua X se caracteriza por su **función de densidad** $f(x)$. Damos en este momento su definición formal. A partir de ahora cuando hablemos de intervalos incluirá el propio $\mathbb{R}$, considerando a este como el intervalo $(-\infty, \infty)$.

Definición 3.2.13 Sea un espacio de probabilidad (E, Ω, P), donde E es un intervalo (a, b). Dada la variable aleatoria continua X, su función de densidad es una función $f : \mathbb{R} \longrightarrow \mathbb{R}$ que cumple las dos propiedades siguientes:

(1) $f(x) \geq 0$ en (a, b) y $f(x) = 0$ fuera de él;

(2) la integral $\int_a^b f(x)dx$ vale 1.

La función de densidad no tiene por qué tener una definición cerrada en términos de una sola fórmula; con frecuencia, es una función a trozos.

Pongamos un ejemplo sencillo para ilustrar esta definición. consideremos la función $f(x)$

$$f(x) = \begin{cases} 2x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

¿Es una función de densidad? Enseguida se ve que la primera condición se cumple, pues la función $2x$ es mayor o igual que 1 en el intervalo $[0, 1]$. Para la segunda condición, tenemos:

$$\int_0^1 xdx = \left[x^2\right]_0^1 = 1 - 0 = 1$$

Por tanto, se cumple la segunda condición y la función sí es de densidad.

[3]Para el lector interesado, existe toda una jerarquía de infinitos. Es algo que se conoce desde finales del siglo XIX gracias a Cantor y otros matemáticos. El infinito más pequeño es el de $\mathbb{N}$ y se designa por $\aleph_0$. El siguiente es el de los números reales y sus intervalos y se designa por $\aleph_1$. Se sabe que $\aleph_0 < \aleph_1$.

Problema 3.2.14 Dado que las funciones de densidad están relacionadas con las integrales por vía de la condición (2) de la definición 3.2.13, interpretad la conexión entre variable aleatoria, probabilidad e integral.

Lo que el lector se preguntará —con lógica impaciencia— es cómo se calcula la probabilidad de un cierto intervalo. En el caso de las variables discretas, la función de masa nos servía para ello; bastaba con sumar las probabilidades asociadas al suceso en cuestión. En el caso continuo la probabilidad de un intervalo (a, b) es la *suma infinita* de la integral de la función densidad. **La probabilidad de un intervalo** se define como sigue.

Definición 3.2.15 Si X es una variable aleatoria continua con función de densidad $f(x)$, la probabilidad de un intervalo $(a, b]$ es

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

Más adelante se justificará esta definición, en particular por qué el primer menor del intervalo es estricto y el segundo no.

Ejercicio 3.2.16 Dada una variable aleatoria continua X con función de densidad $f(x)$, probad lo siguiente:

(a) La probabilidad de un punto es 0;

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

Problema 3.2.17 La definición de probabilidad de intervalo, ¿es compatible con los axiomas de la probabilidad y con el resto de conceptos del capítulo 2? Comprobad que es así.

Problema 3.2.18 La longitud de ciertas piezas, en cm, es una variable aleatoria X con densidad

$$\begin{cases} kx & \text{si } 0 < x < 2 \\ k(4-x) & \text{si } 2 \leq x < 4 \\ 0 & \text{en otro caso} \end{cases}$$

Hallad:

(a) El valor de k para que f sea función de densidad. Representadla.

(b) La probabilidad de que una pieza, tomada al azar, mida entre 1 y 3 cm, sabiendo que mide más de 2 cm.

3.3 Momentos de variables aleatorias

Las variables aleatorias pueden presentar muchas formas, tanto en la forma en que toman los valores como en la distribución de probabilidades de estos. Para intentar caracterizarla se