Guía básica para la simulación de Monte Carlo

Juan Carlos López Agüí

AENORediciones

Índice

1 Introducción

Para los objetivos de esta guía básica, *simulación* será sinónimo de generación de datos artificiales en un ordenador. Entre sus objetivos podemos destacar los siguientes (DAVISON, A.C. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge. 2003 {12}):

a) estudiar la variabilidad que puede esperarse al muestrear en un determinado modelo estadístico,

b) evaluar el grado de adecuación de una aproximación teórica determinada,

c) realizar análisis de sensibilidad de las conclusiones de un estudio en relación con las hipótesis adoptadas,

d) aportar mayor luz respecto de un supuesto o intuición, sobre la base de que es preferible una respuesta incompleta a una cuestión correctamente formulada, que una respuesta precisa a una cuestión incorrecta,

e) proporcionar solución numérica a un determinado problema estadístico cuando la solución analítica o no existe o si existe es demasiado compleja,

f) proporcionar solución numérica a un determinado problema matemático —pero de raíz no directamente estadística— cuando la solución analítica o no existe o si existe es demasiado compleja.

g) servir de apoyo a los métodos clásicos de enseñanza de la Estadística.

La simulación es, por otra parte, una manera económica y útil de experimentación. En muchas ocasiones, el científico o el técnico se encuentran con *sistemas*

reales cuyo funcionamiento desea controlar o mejorar. Un método habitual para alcanzar esos objetivos consiste en experimentar con el sistema real, si ello es posible, e intentar utilizar los resultados de la experimentación para conocer y mejorar el funcionamiento del sistema. En muchas ocasiones, sin embargo, tal experimentación es imposible o, aun siendo posible, éticamente delicada —por ejemplo en Biología— o muy costosa, siendo conveniente disponer de algún método alternativo para ampliar el conocimiento del sistema real. Dos son las posibilidades (Figura 1.1). Construir físicamente un *prototipo*, versión simplificada del sistema real, o crear un *modelo* lógico-matemático que describa mediante un conjunto de ecuaciones —especificación— las relaciones básicas entre los principales elementos del sistema, para experimentar con ellos y no con el sistema. La simulación hace referencia en este escenario a la experimentación con el modelo, sobre todo cuando —como suele ser general— no son suficientes los procedimientos analíticos y numéricos para su estudio, ya que el modelo incluye términos de *ruido* o perturbaciones aleatorias que son esenciales en el mismo, pues sintetizan el reconocimiento de que el modelo es sólo una aproximación del sistema real.

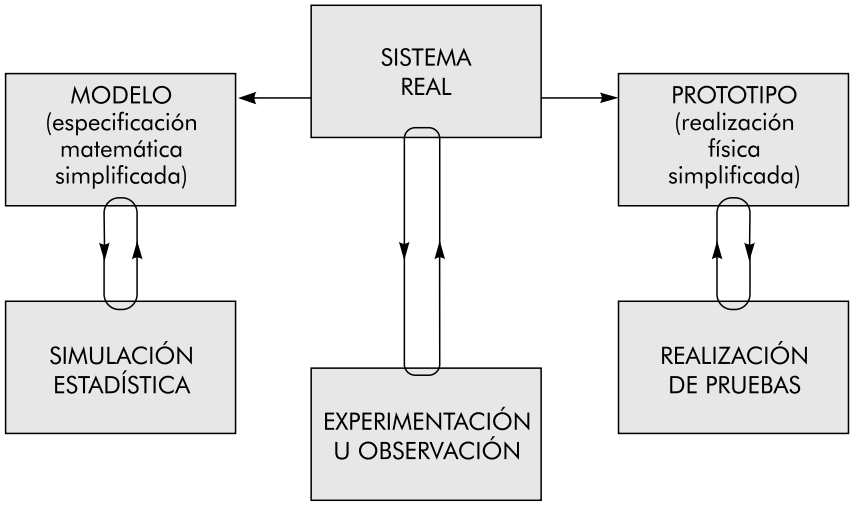

Figura 1.1. **Experimentación con el sistema, los modelos y los prototipos.**

La Figura 1.2 muestra de forma esquemática el funcionamiento de un modelo de un determinado sistema real. Existen unas entradas, tanto estocásticas como deterministas, y unas salidas. El proceso de conversión de las entradas en salidas no es una caja negra —"black box"— como sería si en ese mismo tipo de figura se

describiese el sistema real o incluso un prototipo del mismo. Por definición, las ecuaciones constitutivas del modelo —es decir, su especificación— son conocidas, de modo que el proceso interno de transformación de las variables de entrada en variables de salida es conocido y fijo una vez construido el modelo, y generalmente establecido en forma de *algoritmo*.

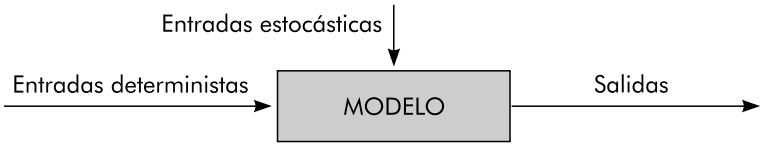

Figura 1.2. **Representación esquemática de las ecuaciones lógico-matemáticas y relaciones que constituyen el modelo.**

La *simulación* consiste en la observación y explotación de resultados que sigue a un plan estructurado de aplicación extensa del modelo. El analista hace variar de forma ordenada las entradas del modelo y obtiene como respuesta un gran número de salidas u observaciones artificiales que analiza estadísticamente para extraer conclusiones del propio modelo y extrapolarlas al sistema real para prever su comportamiento.

Los procedimientos o experimentos de simulación pueden diferir dependiendo de si el conjunto de salidas u observaciones potenciales es discreto o continuo. Además, dichas observaciones pueden ser estáticas o dinámicas, y en este último caso como función continua o discreta del tiempo. Por otro lado, las medidas del comportamiento del sistema también pueden variar enormemente según el modelo.

Con tal abanico de posibilidades no es fácil al principio realizar una taxonomía de los procesos de simulación. Sin embargo, la mayoría de los experimentos de simulación, una vez construido el modelo que simplifica el sistema real, se pueden adaptar al siguiente esquema (RÍOS INSÚA, D., RÍOS INSÚA, S. y MARTÍN, J. *Simulación. Métodos y aplicaciones*. RA-MA. 1997 {37}):

1. Obtención de observaciones básicas de una fuente de números aleatorios.
2. Transformación de las observaciones básicas en entradas al modelo, según las especificaciones del mismo para las entradas estocásticas y deterministas.
3. Transformación de las entradas —deterministas y estocásticas— en salidas.
4. Estimación de las medidas o pautas de comportamiento del sistema mediante el análisis estadístico de las salidas del modelo.

1.1. Principios de la simulación de Monte Carlo

Dentro del contexto general de los procesos de simulación que acaba de ser expuesto, los método calificados de "Monte Carlo" inciden en la última fase del esquema general de los experimentos de simulación, constituyendo unos métodos de estimación bastante potentes de parámetros de interés del sistema real. Para llevar a cabo esa estimación el método de Monte Carlo explota ampliamente la analogía entre probabilidad y volumen. La Estadística Matemática formaliza la noción intuitiva de probabilidad de un suceso identificándola con su volumen o medida relativa en relación con el del universo de posibles resultados de un experimento aleatorio. El método de Monte Carlo utiliza esa identificación en la dirección opuesta, es decir calculando el "volumen" de un conjunto e interpretando dicho volumen como una probabilidad. En el caso más simple eso significa llevar a cabo un muestreo aleatorio del universo de resultados posibles, hacer el recuento de los resultados que pertenecen a un determinado conjunto, calcular la fracción de los resultados pertenecientes a dicho conjunto con respecto al número total de resultados generados, y tomar dicha fracción como una *estimación* del volumen de dicho conjunto. Dentro de unas hipótesis bastante generales, la *ley de los grandes números* nos asegura que esa estimación converge al verdadero valor del volumen del conjunto a medida que aumenta el número de resultados generados artificialmente. Además, y de forma crucial, el *teorema central del límite* facilita información sobre la magnitud del error de estimación cuando el tamaño de la muestra generada es finito, como por otra parte simpre va a suceder.

De gran utilidad para la aplicación de la metodología Monte Carlo resulta también la identificación de la probabilidad de un suceso con la esperanza matemática de cierta función que pasa a ser, por mor de esa identificación, la característica de mayor interés de una variable aleatoria o de una función de ella. Sea X una variable aleatoria unidimensional, $\mathrm{f}(x)$ su función de densidad y D el soporte de dicha variable, de modo que

$$\int_D \mathrm{f}(x)\, dx = 1. \qquad [1.1]$$

Como es bien sabido, la esperanza matemática de dicha variable aleatoria es

$$\mathrm{E}[X] = \int_D x\mathrm{f}(x)\, dx, \qquad [1.2]$$

y la esperanza matemática de la función $g(X)$, que puede ser considerada también como una nueva variable aleatoria obtenida por transformación de la primera,

$$E[g(X)] = \int_D g(x) f(x)\, dx. \qquad [1.3]$$

Sea ahora S un subconjunto de D, $S \subset D$. La probabilidad del suceso $X \in S$ es, como se sabe,

$$\Pr(X \in S) = \int_S f(x)\, dx, \qquad [1.4]$$

donde la integral está extendida al conjunto de valores contenidos en S.

Considérese la *función indicadora* de pertenencia a un determinado conjunto S, definida en la forma:

$$\mathbb{I}_S(x) = \mathbb{I}[x \in S] = \begin{cases} 1 & \text{si} \quad x \in S \\ 0 & \text{si} \quad x \notin S \ (\text{ó} \ x \in \overline{S}). \end{cases} \qquad [1.5]$$

Nótese que con ayuda de esta función indicadora la probabilidad $\Pr(X \in S)$ puede también expresarse en la forma

$$\Pr(X \in S) = \int_D \mathbb{I}_S(x) f(x)\, dx, \qquad [1.6]$$

pues por mero desarrollo de la misma, considerando que $S \cup \overline{S} = D$, se obtiene

$$\int_D \mathbb{I}_S(x) f(x)\, dx = \int_S 1 \cdot f(x)\, dx + \int_{\overline{S}} 0 \cdot f(x)\, dx = \int_S f(x)\, dx, \qquad [1.7]$$

que concuerda con [1.4] como se quería demostrar.
De modo que tanto [1.4] como [1.6] son expresiones válidas y alternativas de $\Pr(X \in S)$, pero [1.6] es un caso particular de [1.3], es decir es una esperanza matemática, concretamente de la función $\mathbb{I}_S(x)$, por lo que se llega a la importante conclusión,

$$\Pr(X \in S) = E[\mathbb{I}_S(x)], \qquad [1.8]$$

que identifica la probabilidad de que X pertenezca a S con la esperanza matemática de la función indicadora de pertenencia de X a S. Esta expresión es clave en la

aplicación de la metodología Monte Carlo para la estimación de probabilidades. Nótese que aunque los desarrollos precedentes se han hecho para el caso de que X sea una variable aleatoria continua, los mismos son inmediatamente generalizables al caso de que X sea discreta sin más que sustituir las integrales por sumas, haciendo que $f(x)$ pase a ser la función de probabilidad de X en lugar de su función de densidad.

Nótese también que las expresiones anteriores son fácilmente generalizables al caso de que X sea una variable aleatoria multidimensional —en ese caso se representaría por $\mathbf{X}$—. En efecto, las integrales simples serían ahora múltiples y la función indicadora, ahora representada por $\mathbb{I}_S(\mathbf{x})$, sería multivariable. No se enuncian las expresiones resultantes por brevedad y por resultar obvias a nuestro juicio.

Estas ideas simples tienen una aplicación inmediata en el campo científico. Supóngase que se desea calcular la integral

$$\alpha = \int_0^1 g(x)\, dx, \qquad [1.9]$$

siendo $g(x)$ una función de la que se va a suponer que no tiene primitiva expresable analíticamente, lo que hace que α no sea calculable por métodos analíticos —en particular por la regla de Barrow—. El camino tradicional seguido en tales casos es el de utilizar algún método de integración numérica para calcular α con una precisión prefijada. Muchos son los métodos competidores, pudiéndose mencionar entre otros la aproximación de Riemann, la regla del trapecio o de Simpson, el procedimiento de Newton-Cotes, el método de Romberg, etc. Se trata en todos los casos de métodos *deterministas* aplicados a un problema de enunciado puramente determinista como es [1.9].

Sin embargo, a pesar de que el sistema real es determinista y de que los mejores procedimientos de resolución de [1.9] son también en este caso los del análisis numérico —de base determinista—, es posible construir un modelo de simulación de base estocástica que permita *estimar* α mediante la metodología Monte Carlo.

Considérese al efecto la variable aleatoria auxiliar $X \equiv U \sim \text{Unif}(0\,;1)$ distribuida uniformemente en el intervalo $[0\,;1]$. La función de densidad en dicho caso es:

$$f(x) = \begin{cases} 0 & ; \quad x \leqslant 0, \\ 1 & ; \quad 0 < x \leqslant 1, \\ 0 & ; \quad 1 < x. \end{cases} \qquad [1.10]$$

Eso hace que se pueda interpretar [1.9], es decir la integral a calcular, como la esperanza matemática de $g(X)$, siendo X una variable aleatoria uniforme en el intervalo $[0;1]$:

$$\alpha = E[g(X)], \qquad [1.11]$$

$$X \sim \text{Unif}(0;1). \qquad [1.12]$$

La Figura 1.3 muestra el modelo de simulación sugerido para estimar el parámetro del sistema real α mediante la metodología Monte Carlo.

Figura 1.3. **Modelo de simulación sugerido para calcular [1.9].**

Se generaría un número m elevado de extracciones aleatorias e independientes de la distribución $\text{Unif}(0;1)$ —en el siguiente epígrafe se verá cómo hacerlo—, designadas por $x_1, x_2, \ldots, x_m$, se evaluaría la función $g(x)$ para cada uno de esos valores y se obtendría la estimación Monte Carlo de α mediante:

$$\hat{\alpha}_m = \frac{1}{m} \sum_{i=1}^{m} g(x_i). \qquad [1.13]$$

Nótese que si $g(\cdot)$ es integrable en $[0;1]$, entonces por la ley fuerte de los grandes números se obtiene el importante resultado de que $\hat{\alpha}_m$ tiende a α con probabilidad 1 cuando m tiende a ∞.

Si además el cuadrado de $g(x)$ es integrable en $[0;1]$ y se define,

$$\sigma_g^2 = \text{var}[g(X)] = \int_0^1 [g(x) - \alpha]^2 \, dx, \qquad [1.14]$$

entonces el error de estimación $\hat{\alpha}_m - \alpha$ se distribuye, por el *teorema central del límite*, aproximadamente normal con media cero y varianza $\sigma_g{}^2/m$, mejorando la aproximación a medida que aumenta m.

Las dos propiedades referidas pueden expresarse formalmente así:

$$\lim_{m\to\infty} \hat{\alpha}_m = \alpha, \quad \text{con probabilidad 1,} \qquad [1.15]$$

$$\sqrt{m}(\hat{\alpha}_m - \alpha) \rightsquigarrow \mathrm{N}(0\,;\sigma_{\mathrm{g}}{}^2) \quad \text{en distribución.} \qquad [1.16]$$

En particular, este último resultado muestra que el término de error, $\hat{\alpha}_m - \alpha$, tiene un error típico —desviación típica de $\hat{\alpha}_m - \alpha$— igual a $\sigma_{\mathrm{g}}/\sqrt{m}$ que va haciéndose más y más pequeño a medida que aumenta m. Así pues, la velocidad de convergencia de $\hat{\alpha}_m$ hacia α está gobernada por la potencia $m^{-1/2}$, que es característica del método de Monte Carlo y que es independiente de la dimensión de la variable aleatoria X —sería la misma velocidad si el problema planteado fuera el cálculo de una integral triple, por ejemplo—. Esto mismo también se expresa diciendo que la aproximación es de orden $m^{-1/2}$, $\mathrm{O}(m^{-1/2})$.

Nótese que eso significa que si se desea aumentar al doble la precisión en la estimación de α, será necesario multiplicar por cuatro el número de puntos muestreados —o lo que es igual, el número de variables aleatorias uniformes generadas—. Aún más, si se desea ampliar en un dígito significativo la precisión del resultado habrá que multiplicar por 100 el número de puntos muestreados.

Esas modestas prestaciones en lo que se refiere a velocidad de convergencia son características de la metodología Monte Carlo. De hecho cualquier método de integración numérica de los mencionados anteriormente supera con creces en eficacia al método de Monte Carlo como procedimiento de integración para la resolución del problema propuesto, que es unidimensional. Sin embargo las cosas cambian si aumenta la dimensión de la integral, pues el método de Monte Carlo no ve afectada su velocidad de convergencia, mientras que en los métodos numéricos se ve perjudicada drásticamente. Y en los modelos de simulación habituales la dimensión de la integral puede ser importante. Además, la metodología Monte Carlo tiene una raíz estocástica tan evidente que la hace preferible en los modelos con componente estocástico, aunque en muchos casos también podría aplicarse en los métodos de integración numéricos deterministas.

De vuelta a los resultados [1.15] y [1.16], especialmente a este último, nótese que no pueden ser aplicados directamente para la construcción de intervalos de confianza para α, porque σ_{g} no es conocida "a priori" —véase por [1.14] que su cálculo precisa de α, que es precisamente la cantidad a estimar por ser desconocida—. Este inconveniente se soslaya en parte al calcular

$$s_{\mathrm{g}} = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} [\mathrm{g}(x_i) - \hat{\alpha}_m]^2}, \qquad [1.17]$$

para lo que sólo es necesario utilizar resultados del modelo de simulación. Con ayuda de s_{g} se puede establecer un intervalo de confianza de nivel $100(1-\delta)\%$ para α. Dicho intervalo, que tiene el carácter de aproximado, es $\hat{\alpha}_m \pm z_{\delta/2}\, s_{\mathrm{g}} / \sqrt{m}$, pues al ser m un número elevado en las aplicaciones de los modelos de simulación está justificado utilizar el cuantil $z_{\delta/2}$ de la $\mathrm{N}(0\,;1)$. Se puede así poner,

$$\Pr\left(\hat{\alpha}_m - z_{\delta/2} \frac{s_{\mathrm{g}}}{\sqrt{m}} \leqslant \alpha \leqslant \hat{\alpha}_m + z_{\delta/2} \frac{s_{\mathrm{g}}}{\sqrt{m}}\right) \simeq 1 - \delta\,, \qquad [1.18]$$

de modo que la estimación puntual $\hat{\alpha}_m$ de la cantidad α queda enriquecida al ser acompañada de acotaciones del tipo [1.18], o al menos del error típico $s_{\mathrm{g}}/\sqrt{m}$, que informan del grado de precisión de la estimación $\hat{\alpha}_m$.

Todos los resultados que se acaban de exponer provienen de la discusión y resolución de la integral [1.9] pero son mucho más generales de lo que pudiera suponerse por estar ligados a un enunciado tan específico. Y ello a pesar de que el enunciado estaba preparado para que los límites de la integral coincidieran con los de la distribución uniforme auxiliar, de modo que dicha integral pudiera ser directamente identificada con una esperanza matemática.

No obstante, el hecho de que los límites de la integral sean diferentes no supone un problema serio. Si se desea calcular la integral

$$\alpha = \int_a^b \mathrm{h}(y)\, dy, \qquad [1.19]$$

se puede utilizar el cambio de variable $y = a + (b-a)x$, con $dy = (b-a)\, dx$, y disponer la integral [1.19] en la forma,

$$\alpha = (b-a) \int_0^1 \mathrm{h}[a + (b-a)x]\, dx = \int_0^1 \mathrm{g}(x)\, dx, \qquad [1.20]$$

con $\mathrm{g}(x) = (b-a)\mathrm{h}[a + (b-a)x]$. Esta integral coincide con [1.9], lo que hace aplicable inmediamente la metodología de Monte Carlo para su resolución. Incluso

si alguno de los límites no es finito existe una transformación adecuada para utilizar los resultados anteriores. Sea por ejemplo la integral

$$\alpha = \int_{0}^{\infty} \mathrm{t}(y)\, dy, \qquad [1.21]$$

y considérese el cambio de variable $y = 1/(1+x)$, para el que $dy = [-1/(1+x)^2]\, dx$. Ahora se puede poner

$$\alpha = \int_{0}^{1} \mathrm{t}\left(\frac{1}{1+x}\right) \frac{dx}{(1+x)^2} = \int_{0}^{1} \mathrm{g}(x)\, dx, \qquad [1.22]$$

con $\mathrm{g}(x) = \mathrm{t}[1/(1+x)]/(1+x)^2$ y aplicar los resultados precedentes.

Otro cambio de variable que produce resultados análogos es $y = -\ln(1-x)$, con $dy = dx/(1-x)$. Aplicando esta transformación a [1.21] se llega a [1.9], siendo en este caso $\mathrm{g}(x) = \mathrm{t}[-\ln(1-x)]/(1-x)$.

Cuando la integral es del tipo

$$\alpha = \int_{a}^{\infty} \mathrm{t}(y)\, dy, \qquad [1.23]$$

el cambio apropiado es $y = a - \ln(1-x)$, con $dy = dx/(1-x)$. La integral [1.23] se transforma de nuevo en [1.9], con $\mathrm{g}(x) = \mathrm{t}[a - \ln(1-x)]/(1-x)$.

Las integrales cuyo límite inferior es $-\infty$ y el límite superior b, como

$$\alpha = \int_{-\infty}^{b} \mathrm{m}(y)\, dy, \qquad [1.24]$$

se transforman fácilmente en integrales del tipo [1.9] mediante el cambio de variable $y = b + \ln(x)$ con $dy = dx/x$. Es claro que en tal caso la función $\mathrm{g}(x)$ de [1.9] sería $\mathrm{g}(x) = \mathrm{m}[b + \ln(x)]/x$.

Por último, las integrales del tipo

$$\alpha = \int_{-\infty}^{\infty} \mathrm{n}(y)\, dy \qquad [1.25]$$

se pueden transformar en integrales del tipo [1.9] mediante el cambio de variable $y = \ln(x) - \ln(1-x) = \ln[x/(1-x)]$. Para este cambio de variable se verifica que $dy = dx/[x(1-x)] = dx/(x - x^2)$ y $\mathrm{g}(x) = \mathrm{n}[\ln[x/(1-x)]]/(x - x^2)$.

Como ha podido comprobarse, la mayor parte de las integrales definidas simples pueden resolverse mediante simulación de Monte Carlo a partir de extracciones de la variable aleatoria uniforme Unif$(0\,;1)$. La situación es bastante más compleja si se trata de una integral múltiple con dimensión $d > 1$, pues el cambio de variables que transforma el recinto S en que se extiende la integral

$$\alpha = \int\int \cdots \int_S g(x_1, x_2, ..., x_d)\, dx_1\, dx_2 \cdots dx_d, \qquad [1.26]$$

en el hipercubo $[0\,;1]^d$, puede ser muy complejo o no existir. En realidad dicho cambio de variables no es necesario si se dispone de algún procedimiento para generar extracciones uniformemente distribuidas en el subconjunto S del soporte D, $S \subset D$. Esa tarea puede ser tan compleja como la de encontrar el cambio de variables que transforme S en $[0\,;1]^d$ y que permita simplificar el proceso de extracción de las variables aleatorias uniformes pues todas ellas pasan a ser Unif$_d[0\,;1]$, es decir uniformes multivariantes de dimensión d.

No se piense sin embargo que la metodología Monte Carlo exige ineludiblemente la generación de variables aleatorias uniformes. Eso ha sido así en el ejemplo introductorio de cálculo de la integral [1.9]. Pero ni la metodología Monte Carlo se aplica exclusivamente al cálculo de integrales, ni el cálculo de éstas tiene por que ser conducido siempre de la manera en que ha sido expuesto.

En muchas ocasiones los modelos de simulación (véase la Figura 1.2) contienen entradas estocásticas que son variables aleatorias con distribución de probabilidad cualquiera: normal o gaussiana, exponencial, logística, etc. En estas circunstancias es necesario disponer de algoritmos para generar tales tipos de distribuciones. Curiosamente tales algoritmos utilizan como materia prima extracciones de variables aleatorias uniformes para transformarlas a continuación en extracciones de las distribuciones deseadas, como si inevitablemente siempre hubiera que recurrir a la distribución uniforme —lo cual es cierto, por otra parte—.

En otras ocasiones, aunque un determinado problema pueda resolverse haciendo uso exclusivo de la distribución uniforme, como por ejemplo las integrales [1.9] y [1.26], puede ser interesante muestrear de una distribución auxiliar h(x) diferente de la uniforme con el objeto de mejorar la precisión de la estimación de Monte Carlo para un tamaño muestral dado. Esta es la idea del llamado *muestreo por importancia*, que busca disminuir la varianza del error de la estimación Monte Carlo concentrando el muestreo en zonas de «importancia» en relación con la estima-

ción. Para fijar ideas se comentará a continuación la aplicación del *muestreo por importancia* al cálculo de integrales.

Supóngase que se quiere estimar como antes la integral [1.9]. Si la función g(x) difiere mucho de la función de densidad de la distribución uniforme —en forma, no en altura o valor— entonces var[g(x)] dado por [1.14] toma un valor elevado, incidiendo negativamente en la precisión de la estimación $\hat{\alpha}_m$. La Figura 1.4 es útil para entender que el hecho de que el muestreo se lleve a cabo en la distribución uniforme tiene una relación directa en la varianza de g(x) y por tanto en la precisión de la estimación. Se ha elegido una función g(x) particular marcadamente decreciente en el intervalo (0 ; 1). Nótese que α representa a la vez el área bajo la curva g(x) en el intervalo (0 ; 1) y el valor medio de g(x) en dicho intervalo por ser de ancho unidad.

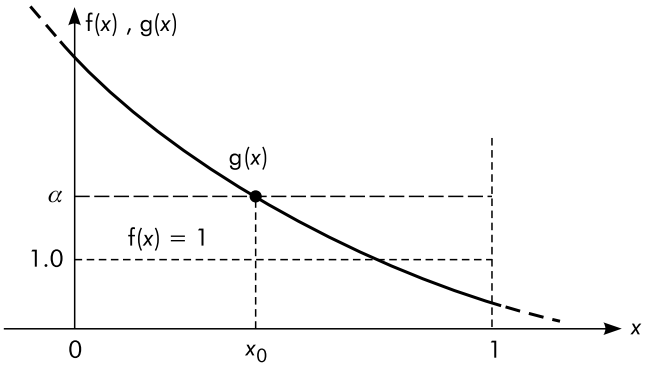

Figura 1.4. **Representación conjunta de la función g(x) y de la distribución uniforme, f(x) = 1, $0 \leqslant x \leqslant 1$, que intervienen en la integral [1.9].**

Esta última es la razón por la que α aparezca representado en la Figura 1.4 como una ordenada que, aunque desconocida por hipótesis, existe, y su inclusión es útil porque facilita la exposición. Es evidente que la mayor parte del área que representa α se encuentra en la parte más a la izquierda del intervalo (0 ; 1), donde las ordenadas de g(x) son altas, mientras que una parte pequeña de dicha área se encuentra en la zona de la derecha del intervalo (0 ; 1) por la razón contraria. Eso significa que cuando una extracción aleatoria de la Unif(0 ; 1) ha ocurrido en una parte de la izquierda del intervalo (0 ; 1) la contribución de dicha extracción al área estimada $\hat{\alpha}_m$, medida por g(x_i)/m según [1.13], tiende a sobrevalorar la contribución media, y que la contribución de dicha extracción a s_g^2 —véase [1.17]—

tiende a ser muy alta. De forma parecida, cuando una extracción aleatoria de la Unif$(0\,;1)$ tiene lugar en la parte derecha del intervalo $(0\,;1)$, la contribución de dicha extracción al área estimada $\hat{\alpha}_m$ tiende a infravalorar la contribución media pero a aumentar el valor de ${s_g}^2$.

Si el muestreo se hubiera concentrado en los aledaños de x_0 (véase la Figura 1.4) las cosas hubieran ido mejor, pues al ser próximos a α los valores $g(x_i)$ correspondientes a las extracciones realizadas, sus contribuciones individuales a $\hat{\alpha}_m$ habrían estado cerca de la media, $\hat{\alpha}_m/m$, y la contribución de $g(x_i)$ a ${s_g}^2$ habría sido pequeña. Sin embargo tal tipo de muestreo privilegiado, o *muestreo por importancia* no puede darse utilizando la distribución uniforme, que tenderá a repartir uniformemente, como es lógico, el conjunto de todas las extracciones en el intervalo $(0\,;1)$. Pero existe, afortunadamente, una manera indirecta de llevar a cabo ese muestreo por importancia, que consiste en ponderar de manera diferente las contribuciones de $g(x_i)$ en $\hat{\alpha}_m$ y en ${s_g}^2$. Esa ponderación distinta de la cuota uniforme $1/m$ se puede llevar a cabo utilizando otra métrica de probabilidad en [1.9] diferente de la uniforme. El cambio de métrica no es especialmente difícil como se verá a continuación.

El desarrollo siguiente se realiza partiendo de la expresión [1.3] que permite hacerlo más general prácticamente sin esfuerzo adicional. Únicamente se necesita aclarar que la esperanza matemática es un operador que utiliza una determinada métrica de probabilidad, la de la distribución cuya función de densidad interviene en el cálculo de dicha esperanza. En la expresión [1.3] la métrica viene inducida por $f(x)$ y tal vez hubiera sido más adecuado expresar dicha integral en la forma

$$\alpha = E_f[g(X)] = \int_D g(x)f(x)\,dx. \qquad [1.27]$$

Si no se hizo así —y en general no se hace así—, es porque el contexto suele dejar perfectamente clara la métrica que se utiliza en el cálculo de la esperanza matemática por lo que resulta en general innecesario reflejarla en el operador esperanza, como se ha hecho en [1.27] disponiendo como subíndice del operador el símbolo de la función de densidad con la que se integra.

Como el método del muestreo por importancia utiliza diferentes métricas para el cálculo de esperanzas matemáticas, conviene utilizar el criterio de nomenclatura expresado en [1.27] por ser más informativo.

Sea ahora $h(x)$ una función de densidad auxiliar, cuyo soporte se va a designar por H, de modo que $D \subset H$, es decir que el soporte D de $f(x)$ queda incluido en H o, lo que es igual, que $f(x) > 0$ implica que $h(x) > 0$ para todo x. Esta exigencia también implica que $h(x) > 0$ para $x \in D$, de modo que la siguiente transformación en [1.27] es válida:

$$\alpha = E_f[g(X)] = \int_D \left[\frac{g(x)f(x)}{h(x)}\right] h(x)\, dx. \qquad [1.28]$$

En esta última integral $h(x)$ ha sustituido a $f(x)$ como métrica de probabilidad, y si no fuera porque la integral está extendida al conjunto D, diferente de H que es el soporte de $h(x)$, parecería que representa la esperanza matemática de la función que integra el corchete bajo la métrica $h(x)$. La esperanza matemática aludida en esta última frase es en realidad

$$E_h\left[\frac{g(x)f(x)}{h(x)}\right] = \int_H \left[\frac{g(x)f(x)}{h(x)}\right] h(x)\, dx, \qquad [1.29]$$

extendida al conjunto H, soporte de $h(x)$. El hecho de que los conjuntos D y H puedan ser diferentes —recuérdese que $D \subset H$— hace pensar que las integrales de las expresiones [1.28] y [1.29] tienen un valor diferente. Esa primera impresión es sin embargo apresurada. En efecto, al partir H en la unión de conjuntos disjuntos $H = D \cup (H - D) = D \cup (H \cap \bar{D})$, la integral [1.29] puede descomponerse en la suma

$$\int_H \left[\frac{g(x)f(x)}{h(x)}\right] h(x)\, dx = \int_D \left[\frac{g(x)f(x)}{h(x)}\right] h(x)\, dx + \int_{H\cap\bar{D}} \left[\frac{g(x)f(x)}{h(x)}\right] h(x)\, dx, \quad [1.30]$$

y si se observa que por ser D el soporte de $f(x)$ debe ser necesariamente $f(x) = 0$ para $x \in \bar{D}$ y por tanto para $x \in H \cap \bar{D}$, se concluye que la última integral de este desarrollo es nula, por lo que se obtiene:

$$\alpha = E_f[g(X)] = \int_D \left[\frac{g(x)f(x)}{h(x)}\right] h(x)\, dx = E_h\left[\frac{g(x)f(x)}{h(x)}\right]. \qquad [1.31]$$

Este resultado es la clave del procedimiento de reducción de varianza denominado *muestreo por importancia*. Establece que la integral α es a la vez la esperanza matemática de $g(X)$ calculada con la métrica $f(x)$ y la esperanza matemática de $g(X)f(X)/h(X)$ calculada con la métrica $h(x)$.

Si se designan por x_{f1}, x_{f2}, ..., x_{fm} ciertas extracciones aleatorias e independientes de la distribución $f(x)$, y por x_{h1}, x_{h2}, ..., x_{hm} ciertas extracciones aleatorias e independientes de la distribución $h(x)$, se pueden construir dos estimaciones diferentes de α, que van a ser designadas por $\hat{\alpha}_{f,m}$ y $\hat{\alpha}_{h,m}$:

$$\hat{\alpha}_{f,m} = \frac{1}{m} \sum_{i=1}^{m} g(x_{fi}), \qquad [1.32]$$

$$\hat{\alpha}_{h,m} = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{f(x_{hi})}{h(x_{hi})} \right] g(x_{hi}). \qquad [1.33]$$

Estas dos estimaciones de Monte Carlo son realizaciones de estimadores centrados en α por [1.31], pero la varianza del error de estimación puede ser muy diferente en uno y otro caso. Una elección adecuada de la distribución auxiliar $h(x)$ puede hacer que la varianza del error de estimación $\hat{\alpha}_{h,m} - \alpha$ sea sensiblemente inferior a la varianza del error de estimación $\hat{\alpha}_{f,m} - \alpha$. Para ello es conveniente que la densidad $h(x)$ sea próxima al producto de $g(x)f(x)$ por una constante de proporcionalidad —que no es necesario determinar—. En tal caso los términos del sumatorio [1.33] serán próximos a esa constante de proporcionalidad aludida y como consecuencia la varianza de $f(x)g(x)/h(x)$ será pequeña. Nótese que en el caso particular de que $h(x) \propto f(x)g(x)$ todos los términos de [1.33] serían constantes y la varianza nula. Sin embargo en este caso la constante de proporcionalidad que al ser aplicada al producto $f(x)g(x)$ la hace coincidir con $h(x)$ es $1/\alpha$, el inverso de la integral a calcular, de modo que esa elección tan precisa de $h(x)$ —cuando es posible— es equivalente al cálculo de la integral solicitada y no es necesario continuar con ningún tipo de muestreo. Ni que decir tiene que una elección desafortunada de $h(x)$ puede conducir a un resultado distinto del deseado: que la varianza de $\hat{\alpha}_{h,m} - \alpha$ sea superior a la de $\hat{\alpha}_{f,m} - \alpha$.

En el caso particular de que $f(x)$ sea la función de densidad de la distribución uniforme Unif$(0;1)$, la integral [1.27] coincide con [1.9] pues $f(x) = 1$ para $0 < x \leqslant 1$ y $f(x) = 0$ para x fuera del intervalo $(0;1]$. En ese caso las extracciones x_{fi} de la expresión [1.32] son uniformes Unif$(0;1)$. Nótese que la expresión [1.33] en tal caso no puede simplificarse sustituyendo $f(x_{hi})$ por la unidad, pues esto sucede sólo en el caso de que $0 < x_{hi} \leqslant 1$, pero como el soporte H de $h(x)$ puede contener valores de x_{hi} fuera del intervalo $(0;1]$, en tales casos será $f(x_{hi}) = 0$, por lo que $f(x_{hi})$ debe permanecer como tal en [1.33] a pesar de que $f(x)$ sea la distribución uniforme. Sólo cuando el soporte H de $h(x)$ coincida con

el intervalo $(0;1]$ será posible simplificar la expresión [1.33] utilizando el hecho de que en tal caso es $f(x_{hi}) = 1$.

Al comparar las expresiones [1.32] y [1.33] se observa que el muestreo por importancia es equivalente a calcular $\hat{\alpha}_{h,m}$ ponderando los valores $g(x_{hi})$ con los pesos $f(x_{hi})/mh(x_{hi})$. Algunos de estos pesos pueden ser cero —cuando $x_{hi} \in \bar{D}$— y otros en cambio relativamente elevados, siendo en tal caso x_{hi} extracciones importantes por su contribución al cálculo de α. Nótese que a pesar de la apariencia de [1.33], $\hat{a}_{h,m}$ no será en general una media ponderada de los valores $g(x_{hi})$, y ello porque los pesos de las diferentes observaciones $g(x_{hi})$ no tienen por qué sumar la unidad.

Un caso de gran importancia es el de la integral [1.6] que identifica la probabilidad de un suceso con la esperanza matemática de la función indicadora de dicho suceso. Comparando [1.6] y [1.27] se deduce que en este caso es $g(x) = \mathbb{I}_S(x)$, de modo que la elección ideal de $h(x)$, que coincide con el producto $g(x)f(x)$ dividido por el valor de la integral, resulta ser

$$h(x)_{\text{ideal}} = \frac{g(x)f(x)}{\alpha} = \frac{\mathbb{I}_S(x)f(x)}{\Pr(X \in S)} = f(x|S), \qquad [1.34]$$

es decir la función de densidad de X condicionada por el suceso $X \in S$. Una elección de $h(x)$ lo más próxima que se pueda a $f(x|S)$ provocará una disminución notable de la varianza del error de estimación.

El *muestreo por importancia* pone de manifiesto que en muchos casos será conveniente muestrear de distribuciones $h(x)$ cualesquiera y no sólo de la distribución uniforme. No obstante, el muestreo de la distribución uniforme es muy importante la en simulación de Monte Carlo, tanto porque es la base de la resolución de algunos problemas típicos de la metodología Monte Carlo —como la integración y la optimización de funciones— como por ser un paso previo para la generación de otro tipo de variables aleatorias, en tanto que los algoritmos ideados para llevar a cabo esa generación se apoyan siempre en la extracción de variables aleatorias uniformemente distribuidas.

La importancia que tiene el muestreo de la distribución uniforme en la simulación estadística hace aconsejable que se le dedique un epígrafe, aunque el estudio del mismo se puede omitir sin pérdida importante de continuidad. Si así se hace, se puede pasar directamente al capítulo 3 dando por hecho que los ordenadores son

capaces de generar secuencias de números aleatorios —en realidad pseudoaleatorios— del tipo U_1, U_2, ..., U_i, que se comportan como extracciones independientes de la distribución uniforme Unif$(0;1)$. En Excel^{TM}, por ejemplo, cada vez que se ejecuta la función ALEATORIO( ) la máquina devuelve un número aproximadamente aleatorio de la secuencia referida.

El cometido de esta publicación es analizar el método de Monte Carlo como herramienta de muestreo artificial. Dicho método ha sido utilizado extensamente para validar los criterios de certificación de muchos productos del sector de la construcción. El autor ofrece la revisión de algunos de los procedimientos más sencillos utilizados como herramientas prácticas para el análisis de datos mediante hojas de cálculo de Excel™.

Uno de sus objetivos concretos es la formalización de la simulación estadística –entendiendo como *simulación* la generación de datos artificiales en un ordenador– en el área de control de calidad y, en especial, en el desarrollo de criterios de aceptación de materiales. Juan Carlos López Agüí se adentra, con una propuesta personal, en el fascinante terreno de la simulación de la distribución gaussiana o normal. Además, el lector encontrará una mención especial a la simulación de los estadísticos de orden más conocidos y de amplio uso en control de calidad.

Este es un título recomendado para el conjunto de profesionales relacionados con el control de calidad de los materiales, aunque puede ser utilizado para otras aplicaciones estadísticas que usan la simulación de Monte Carlo de modo general.

Juan Carlos López Agüí es Doctor Ingeniero de Caminos, Canales y Puertos y Licenciado en Ciencias Económicas y Empresariales. Es profesor en la Universidad Politécnica de Madrid y ponente y autor de numerosos artículos, libros y publicaciones. Dentro del entorno de AENOR, es Presidente del comité AEN/CTN 80 y miembro de la Comisión Permanente. Actualmente es Presidente del Comité Europeo de Normalización (CEN), Director General del Instituto Español del Cemento y sus Aplicaciones (IECA), Vicepresidente del Consejo de la Asociación Científico-Técnica del Hormigón Estructural (ACHE) y Presidente del Grupo de Trabajo "Estadística aplicada al hormigón estructural" (ACHE).

ISBN: 978-84-8143-532-0

AENOR Asociación Española de Normalización y Certificación