Temas de matemáticas

Elementos de simulación estocástica

Luis Rincón
David J. Santana

Elementos de simulación estocástica

Luis Rincón
Departamento de Matemáticas
Facultad de Ciencias UNAM
Circuito Exterior de CU
04510 México CDMX

—

David Josafat Santana Cobian
División Académica de Ciencias Básicas
Universidad Juárez Autónoma de Tabasco

La presente obra se financió con recursos del proyecto
PAPIME PE102321 “Estadística y Simulación” de la DGAPA, UNAM.

La presente obra se financió con recursos del proyecto PAPIME PE102321
"Estadística y Simulación" de la DGAPA, UNAM

Elementos de simulación estocásica
1a edición digital, 2024

ISBN: 978-607-587-029-8

Diseño de portada: Eliete Martín del Campo Treviño

Este libro ha sido dictaminado por pares académicos y sometido a aprobación
del Comité Editorial de la Facultad de Ciencias de la UNAM.

Impreso y hecho en México.

Prólogo

El presente texto contiene una introducción a diversos temas de la simulación estocástica. Está dirigido a estudiantes de licenciatura de Actuaría, Matemáticas y Matemáticas Aplicadas, y más generalmente, a estudiantes de otras carreras del área científica como Física, Ciencias de Datos y las distintas Ingenierías, quienes pueden tomar esta asignatura como optativa dentro del área de probabilidad y estadística en los últimos semestres de sus planes de estudio. Como requisitos, se espera que el lector haya cursado un primer semestre en probabilidad y otro en estadística matemática, e idealmente un primer curso de procesos estocásticos.

Una simulación es una recreación o representación de un proceso o sistema que evoluciona en el tiempo. Para tener tal representación se requiere contar con un modelo que capture las características del sistema que es de nuestro interés. Como el modelo debe evolucionar en el tiempo de manera preferentemente no determinista, es necesario incorporar elementos aleatorios. Al estudio de la generación de estos elementos azarosos se le llama simulación estocástica. El texto contiene material suficiente para ser cubierto en un semestre y cubre temas como el concepto de número pseudoaleatorio, diversos métodos para generar valores de variables y vectores aleatorios, la integración Monte Carlo, algunos métodos basados en cadenas de Markov y el algoritmo EM para la estimación de parámetros ante la falta o censura de datos.

Nuestro tratamiento es elemental pero riguroso en el sentido de que buscamos proveer enunciados y demostraciones de las fórmulas, y la justificación de los procedimientos. Es nuestra expresa intención no usar ningún lenguaje de programación particular para ilustrar los algoritmos que se estudian. Más bien, se proveen los procedimientos en lenguaje natural, los cuales pueden ser implementados con relativa facilidad por parte del lector en el lenguaje de programación de su gusto, conocimiento o popularidad del momento. Sin embargo, con el fin de realizar una primera ilustración de una programación básica para algunos de los algoritmos presentados, se agregaron en el apéndice códigos en el lenguaje R, el cual tiene actualmente y un gran respaldo en la comunidad estadística.

Agradecemos a la DGAPA UNAM por el apoyo otorgado a través del proyecto PAPIME PE102321 “Estadística y simulación”, mediante el cual pudo ser posible la edición de este trabajo. Asimismo, agradecemos muy sinceramente a los árbitros de este trabajo por sus valiosos comentarios y a los Comités de Publicaciones del Departamento de Matemáticas y de la Facultad de Ciencias por su excelente labor editorial. Finalmente, es un grato deber agradecer a Brayan Ricardez Gómez por su apoyo en la elaboración de algunas de las gráficas que aparecen a lo largo del texto y en el fortalecimiento del número de ejercicios propuestos.

Los autores
Julio 2024

Contenido

Capítulo 1

Introducción

Una simulación es una recreación de un proceso o sistema. Regularmente tal proceso o sistema evoluciona en el tiempo y se utiliza un modelo que capture aquellas características del sistema real que son de nuestro interés, y que deseamos reproducir. El sistema real puede ser demasiado grande para interaccionar con él, o bien puede ser muy costoso o muy riesgoso, así es que la imitación del sistema real puede ser de utilidad.

La simulación se usa para recrear posibles escenarios de un sistema sin interactuar con el sistema real. Con frecuencia se usan computadoras para simular la evolución del sistema, y uno de sus usos principales es el entrenamiento de las personas para desarrollar habilidades que lleven a tener comportamientos adecuados, o tomar decisiones convenientes, en los diferentes escenarios del modelo del sistema. Por ejemplo, los campos de entrenamiento de los soldados les permiten simular situaciones reales de batallas y desarrollar las habilidades requeridas. Un tanque de agua puede servir como un ambiente simulado para que los astronautas aprendan a moverse con eficiencia y manejar instrumentos en ambientes de gravedad cero. Estos son dos ejemplos en los que una computadora no se usa directamente en la recreación del sistema, aunque lo más común es que los sistemas computacionales jueguen un papel importante en la simulación. Los simuladores de vuelo o de operación de maquinaria especializada y muchos de los videojuegos son ejemplos en donde la simulación se lleva a cabo enteramente en sistemas de cómputo. Otros ejemplos son: simulador de

juegos de casino, simulador de inversiones, simulador de negocios, simulador de clima, simulador de enfermedades infecciosas, etc.

En particular, en la simulación estocástica algunas de las variables del modelo se consideran aleatorias y esto permite que la evolución en el tiempo del sistema simulado no sea determinista. Así es como parece que es el mundo real pues, siendo éste tan complejo, es imposible tener siempre control de cada variable y muchos de los acontecimientos parecen azarosos. Por ejemplo, para aplicar diversos métodos de la estadística matemática, en muchas ocasiones se requiere contar con los valores de una muestra aleatoria, es decir, valores numéricos generados al azar de manera independiente uno del otro y que siguen una distribución de probabilidad dada. Por otro lado, la simulación de las trayectorias de un proceso estocástico puede ayudar a estimar algunas cantidades o características complejas del proceso.

Considerando las herramientas matemáticas, definiciones y conceptos necesarios, la simulación estocástica es una área de la probabilidad aplicada. Los métodos dentro de esta área buscan aproximar o estimar resultados numéricos que son difíciles de calcular de forma analítica, o bien, que son resultado de fenómenos o experimentos aleatorios. Se pueden encontrar aplicaciones en diversas áreas de la ciencia como la física, la química y la biología, por mencionar algunas. Dentro de la probabilidad y la estadística, podemos realizar aplicaciones en problemas relacionados a procesos estocásticos, teoría del riesgo, finanzas cuantitativas, muestreo, econometría, series de tiempo, estadística bayesiana, etc.

A continuación se mencionan algunas consideraciones que ayudarán a entender el rumbo del presente trabajo.

- Las distintas técnicas que se estudian en simulación estocástica parten de obtener primero una muestra de valores llamados números aleatorios.
- En las aplicaciones de simulación estocástica es indispensable la generación de valores de variables aleatorias con distintas distribuciones de probabilidad.
- El fundamento principal de las aproximaciones hechas en simulación

estocástica es la llamada ley de los grandes números.

Teniendo en cuenta estos pilares, se buscará dar respuesta a las siguientes preguntas guía: ¿Cómo generar valores de una variable aleatoria? ¿De qué forma pueden ser usados tales valores para aproximar probabilidades y valores esperados? ¿Cuál es el número de simulaciones que se deben realizar y cuál es la confiabilidad de las aproximaciones producidas? ¿Cómo pueden ser aprovechadas estas técnicas para realizar estimaciones estadísticas?

La lista de preguntas anterior podría ser más larga, pero la mayoría de las aplicaciones de la simulación estocástica resuelve problemas derivados de una combinación de estas cuestiones. En la siguiente sección presentaremos algunos primeros ejemplos de aplicación de la simulación estocástica. Estos ejemplos buscan despertar el interés del lector para emprender el estudio de estos temas.

1.1. Ejemplos de aplicaciones

En esta sección, se mencionan algunos ejemplos comunes que tienen impacto directo en el contenido de este trabajo. Se muestra la flexibilidad del uso de simulación para resolver problemas de probabilidad, cálculo integral y estadística, por mencionar algunos.

Eventos condicionados

Consideremos el experimento aleatorio consistente en lanzar un dado equilibrado y denotemos por W el resultado del lanzamiento. Es claro que la variable W puede tomar los valores: $1, \ldots, 6$. Después de esto, se lanza una moneda equilibrada en W ocasiones. Supondremos que los lados de la moneda se denominan "cara" y "cruz". Se registra el número de resultados "cruz" obtenidos y al número de veces que aparece este resultado le llamaremos Y. Esta nueva variable Y puede tomar los valores $0, 1, \ldots, W$. El problema consiste en calcular el número esperado de cruces que se obtienen al final del experimento, es decir, se desea calcular $E(Y)$.

Este problema es fácil de resolver si se conoce el siguiente resultado de pro-

babilidad condicional llamado esperanza iterada:

$$E(Y) = E(E(Y \mid W)). \tag{1.1}$$

Usaremos esta identidad para resolver con facilidad el problema planteado. Notemos que

$$\begin{aligned} W &\sim \text{unif}\,\{1, \ldots, 6\} \\ Y \mid (W = w) &\sim \text{bin}\,(w, 1/2). \end{aligned}$$

Al lanzar w veces la moneda se tiene que $E(Y \mid W = w) = w/2$, y esto implica que $E(Y \mid W) = W/2$. En la Sección 4.2 que inicia en la página 149, el lector encontrará una exposición breve sobre este tipo de esperanzas. Regresando al problema planteado, tenemos que $E(W) = (1 + \cdots + 6)/6 = 7/2$. Por lo tanto,

$$E(Y) = E(E(Y \mid W)) = E(W/2) = (1/2)(7/2) = 7/4.$$

De forma alternativa, se puede encontrar una aproximación a la solución del problema haciendo varias simulaciones del lanzamiento del dado y la moneda. La ley de los grandes números garantiza que conforme más veces se repita el experimento, el promedio de los resultados se aproximará más al resultado exacto. Más específicamente, se puede obtener una aproximación a $E(Y)$ mediante el siguiente algoritmo:

Algoritmo para estimar $E(Y)$

Sea n el número de veces que se realiza el experimento aleatorio.
Para $t = 1, 2, \ldots, n$,

- Genere un valor w_t de una variable aleatoria $W \sim \text{unif}\,\{1, \ldots, 6\}$.
- Genere un valor y_t de una variable aleatoria $Y \sim \text{bin}\,(w_t, 1/2)$.

De esta manera se obtienen los valores $y_1, \ldots, y_n$ y se propone la aproximación

$$E(Y) \approx \overline{y}_n := \frac{1}{n} \sum_{t=1}^{n} y_t. \tag{1.2}$$

En la Figura 1.1 se puede observar la evolución del promedio (1.2) conforme aumenta el tamaño n de la muestra. Esta gráfica se obtuvo al implementar

el algoritmo anterior en el lenguaje R. Una observación importante es que en el cálculo de (1.2) no importa el orden de los sumandos. Por lo anterior, el primer paso del algoritmo puede ser simplificado si se genera un número de resultados que sea múltiplo de 6, debido a la distribución uniforme del resultado del lanzamiento del dado. Por ejemplo, si $n = 6\,m$, entonces la muestra sería conformada con m valores iguales a 1, m valores iguales a 2, y así hasta m valores iguales a 6.

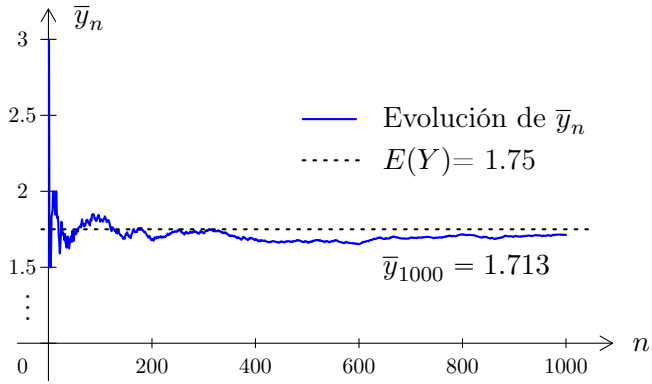


Figura 1.1: Evolución del promedio $\overline{y}_n$ y su aproximación al valor $E(Y) = 1.75$.

Cadenas de Markov y tiempos de arribo

Recordemos que un proceso estocástico es, a grandes rasgos, una sucesión de variables aleatorias dependientes de un parámetro temporal. Podemos clasificar a los procesos estocásticos de acuerdo a las propiedades de dependencia entre las variables aleatorias que lo conforman. Uno de los procesos estocásticos más estudiados y con más aplicaciones son los llamados procesos de Markov o cadenas de Markov. Su principal característica es que el valor del proceso en un momento determinado sólo depende del pasado más próximo. Para entender el ejemplo que se expone a continuación, el lector debe estar familiarizado con las cadenas de Markov en tiempo discreto. Para una revisión breve de este tema véase la Sección 5.1 en la página 190.

Consideremos la siguiente pregunta: ¿cuántos lanzamientos de una moneda justa son necesarios para obtener dos "caras" consecutivas? Para encontrar la solución podemos ayudarnos de la cadena de Markov $\{X_n : n = 0, 1, \ldots\}$ homogénea en el tiempo, con espacio de estados $\{0, 1, 2\}$ y definida mediante la siguiente matriz de probabilidades de transición

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}.$$

La variable X_n se define como la longitud de la racha de "caras" al tiempo n y sus valores tienen la siguiente interpretación:

a) $X_n = 0$ cuando en el lanzamiento n se ha obtenido el resultado "cruz".

b) $X_n = 1$ cuando en el lanzamiento n se ha obtenido el resultado "cara" y en el lanzamiento anterior se ha obtenido "cruz".

c) $X_n = 2$ cuando en el lanzamiento n, y en el anterior, se ha obtenido el resultado "cara", o bien, cuando $X_{n-1} = 2$, debido a que el estado 2 se define como estado absorbente.

Para responder a la pregunta planteada, realizando un análisis del primer paso, se pueden obtener ecuaciones recursivas, en las cuales se desconocen los tiempo esperados de arribo al estado 2 desde cada estado inicial. Alternativamente, es posible aproximar la solución haciendo uso de simulaciones de las trayectorias de la cadena, con la finalidad de promediar el número de pasos necesarios para arribar al estado 2. En el siguiente capítulo se estudiarán métodos para generar valores de una variable aleatoria con distribución Bernoulli y de esa manera se podrán realizar simulaciones de lanzamientos de una moneda, con lo cual se podría realizar un algoritmo para obtener la aproximación mencionada.

Integración de funciones

Otra aplicación interesante de la simulación es la aproximación de integrales. La integración de funciones que no tienen función primitiva (antiderivada) es un problema que requiere el uso de métodos numéricos para encontrar

una solución aproximada. De forma alternativa, se pueden encontrar aproximaciones expresando a la integral como un valor esperado. Veamos, por ejemplo, la siguiente integral que no posee solución analítica,

$$\int_0^1 e^{-x^2} dx.$$

Puede observarse que esta integral coincide con el valor esperado de la transformación de una variable aleatoria X con distribución unif $(0, 1)$, es decir, si $f(x)$ denota la función de densidad de la distribución indicada, entonces

$$\int_0^1 e^{-x^2} dx = \int_0^1 e^{-x^2} f(x)\, dx = E\left(e^{-X^2}\right).$$

Haciendo uso de la ley de los grandes números, se puede aproximar este valor esperado usando un promedio, generando previamente valores de la transformación $\exp\left(-X^2\right)$, a partir de la obtención de valores de la variable aleatoria X con distribución unif $(0, 1)$. En la Sección 1.2 se retoma este ejemplo y en el Capítulo 3 se estudiará con más detalle esta aplicación.

Hay que notar que si la integral tuviera límites de integración distintos, también sería posible expresarla en términos de un valor esperado. Veamos, por ejemplo, la siguiente integral y consideremos ahora una variable aleatoria $Y \sim$ unif $(0, 2)$. Si $f(x)$ es la función de densidad de la distribución mencionada, entonces

$$\int_0^2 e^{-x^2} dx = 2 \int_0^2 e^{-x^2} f(x)\, dx = 2\, E\left(e^{-Y^2}\right).$$

Aplicaciones en estadística

En estadística y en todas las áreas relacionadas al análisis de datos, puede ser necesario producir muestras aleatorias numéricas (observe la Definición 1.1), por ejemplo, para obtener estimaciones de los parámetros de algún modelo que involucre variables aleatorias. Adicional a lo anterior, en modelos de regresión o de series de tiempo puede resultar de utilidad generar distintos escenarios para observar el comportamiento de estos modelos cuando se hacen variar sus parámetros (análisis de sensibilidad).

Por otro lado, cuando se realizan pruebas de hipótesis, en ocasiones se tiene un estadístico de prueba para el cual no es posible obtener el valor exacto de su función de distribución. Lo anterior provoca que no se puedan calcular ni el valor exacto del *p-value*, ni los valores críticos para concluir si se rechaza o no se rechaza una hipótesis nula. En estos casos, se pueden generar muestras del estadístico y dar valores aproximados a las cantidades de interés.

En finanzas, las llamadas opciones financieras, son acuerdos o contratos para cubrirse ante un cambio brusco en los precios de algún bien. La modelación estadística de estos precios y su simulación son útiles para obtener aproximaciones a las primas o precios que deben pagarse por estas opciones financieras (valuación de opciones financieras).

En el área de seguros, al generar simulaciones de variables aleatorias que modelan frecuencia de reclamaciones o que modelan el tamaño del monto reclamado en cada evento, se pueden obtener aproximaciones a ciertas cantidades que sirven para evaluar la solvencia de una compañía. Es decir, para medir su capacidad para enfrentar tales reclamaciones.

Otra aplicación puede darse dentro del análisis de datos cuando se tiene que trabajar con bases de datos incompletas, es decir, con datos faltantes. Existen algoritmos para intentar estimar estos datos e involucran directamente la simulación estocástica. Dos técnicas relacionadas a este problema y que son muy conocidas son el algoritmo EM y el algoritmo conocido como *data augmentation*.

Además de las aplicaciones anteriores, también existen las llamadas técnicas de remuestreo para realizar estimaciones de características que tienen que ver con la función de distribución de una muestra aleatoria. En particular es de interés estimar la desviación estándar cuando se tiene poca información de la distribución original. En este caso, son muy populares las técnicas conocidas como jackknife y bootstrap.

En el Capítulo 6 se estudiarán con más detalle estas técnicas.

1.2. Elementos de probabilidad

Es esta sección se hace una breve revisión de algunas definiciones y resultados de probabilidad y estadística que posiblemente ya conozca el lector. Se hace énfasis en ellos por ser indispensables para entender el contenido de los siguientes capítulos.

Valor esperado

El valor esperado es ampliamente estudiado en los cursos básicos de probabilidad y es un buen punto de partida para repasar algunas definiciones relacionadas. Aunque no se estará mencionando, se supondrá que las variables aleatorias que se utilicen están definidas en una espacio de probabilidad $(\Omega, \mathcal{F}, P)$, donde la medida de probabilidad P está caracterizada por la función de distribución. La media, esperanza, o valor esperado de una variable aleatoria X con función de distribución $F(x)$, se define y denota por

$$E(X) = \int_{-\infty}^{\infty} x \, dF(x), \tag{1.3}$$

cuando la integral del valor absoluto es convergente. Esta definición está expresada en términos de la integral de Riemann-Stieltjes y es válida incluso en casos donde no existe la función de densidad para distribuciones continuas. El soporte de X se define y denota como el conjunto $\mathcal{C} = \{x : P(X = x) > 0\}$ en el caso discreto, y como $\mathcal{C} = \{x : f(x) > 0\}$ en el caso absolutamente continuo, donde $f(x)$ es la función de densidad de X. Se tienen los siguientes casos particulares de la expresión (1.3):

1. Si X es una variable aleatoria discreta, entonces

$$E(X) = \sum_{x \in \mathcal{C}} x \, P(X = x).$$

2. Si X es absolutamente continua, entonces

$$E(X) = \int_{\mathcal{C}} x \, f(x) \, dx.$$

En el caso multivariado y considerando funciones que dependen de varias variables aleatorias, se tienen las siguientes expresiones, la primera para el

caso discreto y la segunda para el caso continuo.

$$E\left(g(X_1,\ldots,X_n)\right) = \sum_{x_1\in\mathcal{C}_1}\cdots\sum_{x_n\in\mathcal{C}_n} g(x_1,\ldots,x_n)\,P(X_1=x_1,\ldots,X_n=x_n),$$

$$E\left(g(X_1,\ldots,X_n)\right) = \int_{\mathcal{C}_1}\cdots\int_{\mathcal{C}_n} g(x_1,\ldots,x_n)\,f(x_1,\ldots,x_n)\,dx_n\cdots dx_1,$$

en donde f es la función de densidad conjunta. El valor esperado cumple las propiedades de un operador lineal, es decir, cumple que

$$E\left(\sum_{i=1}^{n}\alpha_i\,X_i\right) = \sum_{i=1}^{n}\alpha_i\,E\left(X_i\right), \tag{1.4}$$

en donde $\alpha_1,\ldots,\alpha_n$ son constantes. Además, si para una sucesión de variables aleatorias $X_1, X_2, \ldots$ existe la variable aleatoria $\lim_{n\to\infty} X_n$, entonces

$$E\left(\lim_{n\to\infty} X_n\right) = \lim_{n\to\infty} E\left(X_n\right), \tag{1.5}$$

siempre y cuando se cumpla alguna de las dos hipótesis siguientes:

a) La sucesión es monótona, creciente o decreciente, c.s. A este resultado se le conoce como el teorema de convergencia monótona.

b) Existe una variable aleatoria Y tal que $P\left(|X_i| \leqslant Y\right) = 1$, para $i = 1, 2, \ldots$ y $E(Y) < \infty$. Este es el teorema de convergencia dominada.

Como un caso particular, observemos que el valor esperado de la suma infinita de variables aleatorias, en general, no es la suma infinita de los valores esperados, pero cuando las variables aleatorias son tales que $X_i \geqslant 0$ c.s. para $i = 1, 2, \ldots$ y $X_1 + X_2 + \cdots < \infty$ c.s., entonces por el teorema de convergencia monótona,

$$E\left(\sum_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} E(X_i). \tag{1.6}$$

Varianza

Al considerar que una variable aleatoria es igual a un número real después de la realización de un experimento aleatorio intrínseco cuyo resultado es

un elemento del espacio muestral Ω, la varianza puede tomarse como una medida de incertidumbre. Lo anterior en el sentido de que a mayor varianza, mayor incertidumbre en los resultados de cada experimento aleatorio. Con relación al valor esperado, recordemos que la varianza de una variable aleatoria X se define y denota como

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right]. \tag{1.7}$$

La varianza es un caso particular de la covarianza de la variable aleatoria de X con ella misma. Recuérdese que para cualquier par de variables aleatorias X y Y, su covarianza se define y denota como

$$\mathrm{Cov}(X, Y) = E\left[\,(X - E(X))\,(Y - E(Y))\,\right]. \tag{1.8}$$

Aplicando las propiedades de la linealidad del valor esperado, la covarianza se puede escribir de la siguiente manera,

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y),$$

y, como consecuencia,

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = E(X^2) - E^2(X).$$

Se puede demostrar que si $X_1, \ldots, X_n$ son variables aleatorias independientes con $\mathrm{Var}(X_i) < \infty$, entonces

$$\mathrm{Var}\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i^2 \,\mathrm{Var}\,(X_i)\,. \tag{1.9}$$

Omitiendo la hipótesis de independencia entre las variables aleatorias, la fórmula general es la siguiente: si $\alpha_1, \ldots, \alpha_n$ son números reales,

$$\mathrm{Var}\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i^2 \,\mathrm{Var}\,(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \alpha_i\,\alpha_j \mathrm{Cov}(X_i, X_j). \tag{1.10}$$

Muestra aleatoria

El uso de muestras aleatorias (m.a.) es de gran importancia en la estadística y en la simulación estocástica. Una muestra aleatoria es una sucesión de

variables aleatorias que cumplen ciertas condiciones que especificamos a continuación.

Definición 1.1

a) Una muestra aleatoria es una sucesión de variables aleatorias $X_1, \ldots, X_n$ que son independientes e idénticamente distribuidas.

b) Al valor entero $n \geqslant 1$ se le conoce como el tamaño de la muestra.

c) Llamaremos muestra aleatoria numérica a una sucesión de números reales $x_1, \ldots, x_n$ que son realización de una muestra aleatoria $X_1, \ldots, X_n$.

En el caso de una muestra aleatoria numérica, ésta puede ser el resultado de varios escenarios. Por ejemplo, los valores podrían provenir de mediciones aplicadas a una población o podrían ser registros de algún fenómeno. Los datos obtenidos pueden ser analizados para conocer características de la población o fuente de información. Es posible aplicar pruebas de hipótesis, estimaciones puntuales o por intervalo, de parámetros desconocidos. Por otro lado, se pueden aplicar pruebas de bondad de ajuste para ajustar alguna distribución a la muestra. Las muestras aleatorias también pueden ser resultado de simular valores de una variable aleatoria conociendo su distribución. Este segundo caso, donde es necesario generar muestras aleatorias con distintas distribuciones, es el objetivo de la primera parte de este trabajo.

Definición 1.2

a) Sea $X_1, \ldots, X_n$ una muestra aleatoria. La media muestral se define como la variable aleatoria

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{1.11}$$

b) Sea $x_1, \ldots, x_n$ una muestra aleatoria numérica de realizaciones de una muestra aleatoria. Al número $\overline{x}_n$ que aparece abajo se le llama media muestral numérica,

$$\overline{x}_n := \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1.12}$$

Hay que notar que se tienen dos casos que a veces se tratan de forma indistinta en algunos métodos y, sin embargo, se requiere tener claro el significado de cada uno. Por un lado, la media muestral (1.11) es una variable aleatoria y, como tal, se le puede calcular su valor esperado y su varianza. Por el otro lado, la media muestral numérica (1.12) es un número real y es una realización de la variable aleatoria (1.11).

Proposición 1.1 *Sea $X_1, \ldots, X_n$ una muestra aleatoria con media μ y varianza σ^2, ambas finitas. Entonces*

1. $E\left(\overline{X}_n\right) = \mu.$
2. $\text{Var}\left(\overline{X}_n\right) = \dfrac{\sigma^2}{n}.$

Demostración. Para el valor esperado, se aplican las propiedades de linealidad. Para la varianza se aplica la hipótesis de independencia para calcular la varianza de la suma como suma de las varianzas.

1. $E\left(\overline{X}_n\right) = E\left(\dfrac{1}{n} \sum_{i=1}^{n} X_i\right) = \dfrac{1}{n} \sum_{i=1}^{n} E\left(X_i\right) = \mu.$

2. $\text{Var}\left(\overline{X}_n\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right) = \frac{\sigma^2}{n}$.

■

De los resultados anteriores, es inmediato observar que:

a) $\lim_{n\to\infty} E\left(\overline{X}_n\right) = \mu$.

b) $\lim_{n\to\infty} \text{Var}\left(\overline{X}_n\right) = 0$.

Es decir, entre más grande sea el tamaño de una muestra aleatoria, menor varianza tendrá la media muestral $\overline{X}_n$. En consecuencia, las estimaciones que se obtengan usando este valor tendrán más precisión.

Ley de los grandes números

El resultado indispensable que justifica varias de las aplicaciones de simulación estocástica es la llamada ley de los grandes números (LGN). En la siguiente proposición se enuncia la versión fuerte de esta ley.

Proposición 1.2 *Sea $X_1, \ldots, X_n$ una muestra aleatoria con media común $\mu < \infty$. Entonces*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu \quad \text{c.s.} \tag{1.13}$$

La versión débil establece con la convergencia es en probabilidad y la ley fuerte dispone que la convergencia es casi segura (c.s.). Hay que recordar que la convergencia casi segura consiste en que, si existe un conjunto $A \subseteq \Omega$, tal que $\lim_{n\to\infty} \overline{X}_n(a) \neq \mu$, para $a \in A$, entonces $P(A) = 0$. Esto implica que

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1. \tag{1.14}$$

Para una exposición más detallada sobre convergencia casi segura y la LGN, puede consultarse, por ejemplo, el texto de A. Gut [18]. En el caso

de aplicar una función de manera individual a cada una de las variables de una muestra aleatoria $X_1, \ldots, X_n$, el promedio de la función también converge casi seguramente a la media de la función evaluada en cualquiera de las variables. La siguiente proposición establece lo anterior con más precisión.

Proposición 1.3 *Sea $X_1, \ldots, X_n$ una muestra aleatoria y sea g una función definida sobre el conjunto de valores que toman estas variables aleatorias y tal que $E(g(X)) < \infty$. Entonces*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = E(g(X)), \quad \text{c.s.} \tag{1.15}$$

La siguiente igualdad es fundamental en la simulación estocástica porque nos muestra que la LGN también se aplica para aproximar probabilidades simulando eventos,

$$P(A) = E(1_A), \tag{1.16}$$

en donde A es un evento y el término 1_A representa la función indicadora de dicho evento, es decir,

$$1_A = \begin{cases} 1 & \text{si sucede } A, \\ 0 & \text{si no sucede } A. \end{cases}$$

Por lo que,

$$E(1_A) = 1 \cdot P(1_A = 1) + 0 \cdot P(1_A = 0) = P(1_A = 1) = P(A).$$

Por lo anterior, cuando se dificulta el cálculo exacto de $P(A)$, y el evento A se define usando variables aleatorias, entonces una alternativa es generar una muestra de valores de las variables involucradas y determinar en cada caso si sucede o no el evento A. Por lo que se estará generando una muestra aleatoria tipo Bernoulli con las funciones indicadoras $1_{A_1}, \ldots, 1_{A_n}$, en cada caso con probabilidad de éxito $P(A)$. De esta manera, la aproximación estará dada por

$$P(A) \approx \frac{1}{n} \sum_{i=1}^{n} 1_{A_i}.$$

Proposición 1.4 *Sea A un evento de interés que sucede o no sucede de acuerdo a alguna condición relacionada con una o varias variables aleatorias. Sean $1_{A_1}, \ldots, 1_{A_n}$ una sucesión de n funciones indicadoras, obtenidas evaluando la condición de ocurrencia de los eventos A_i a partir de la simulación de las variables aleatorias relacionadas. Entonces*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1_{A_i} = P(A), \quad \text{c.s.} \tag{1.17}$$

A continuación se muestra un ejemplo sencillo de aplicación de la LGN, en este caso aproximando el valor de una integral de Riemann expresada como un valor esperado.

Ejemplo 1.1 *Supongamos que $x_1, \ldots, x_n$ es una muestra aleatoria numérica que proviene de una variable aleatoria X con distribución* unif $(0, 1)$. *Haciendo uso de esta muestra, se aproximará el valor de la integral*

$$\int_0^1 e^{-x^2}\, dx.$$

Definamos la función constante $f(x) = 1$, para $0 < x < 1$. Esta es la función de densidad de la distribución unif $(0, 1)$. *Entonces*

$$\begin{aligned} \int_0^1 e^{-x^2}\, dx &= \int_0^1 e^{-x^2} f(x)\, dx \\ &= E\left(e^{-X^2}\right) \\ &\approx \frac{1}{n} \sum_{i=1}^{n} e^{-x_i^2}. \end{aligned}$$

■

En la Sección 1.3 se estudiará la manera en la que se puede obtener la muestra aleatoria $x_1, \ldots, x_n$ necesaria en el ejemplo anterior, y en el siguiente capítulo se estudiarán mecanismos generales para generar muestras aleatorias de cualquier distribución.

Teorema central del límite

La suma de un gran número de variables aleatorias independientes e idénticamente distribuidas tiende a tener una distribución normal. Lo anterior, es una consecuencia del resultado conocido como teorema central del límite (TCL) que se enuncia a continuación.

Teorema 1.1 *Sea $X_1, \ldots, X_n$ una muestra aleatoria con media μ y varianza $\sigma^2 < \infty$. Para cualquier número real x,*

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant x\right) = \Phi(x), \tag{1.18}$$

en donde $\overline{X}_n = (X_1 + \cdots + X_n)/n$ y $\Phi(x)$ es la función de distribución normal estándar.

El TCL indica que la variable aleatoria $Z = \left(\overline{X}_n - \mu\right)/\left(\sigma/\sqrt{n}\right)$ tiende a tener una distribución $\mathrm{N}(0,1)$ cuando el tamaño de la muestra n es grande. Como consecuencia directa, tanto $\overline{X}_n$ como $\sum_{i=1}^n X_i$, tienden a tener distribución normal para un tamaño de muestra grande y con ciertos valores para los parámetros. Como en el caso de la LGN, el TCL tiene interesantes aplicaciones numéricas. Por ejemplo, a continuación se ilustra una forma en la que se puede aplicar el TCL para obtener aproximaciones a probabilidades de eventos.

Ejemplo 1.2 *Supongamos que $X_1, \ldots, X_n$ es una sucesión de variables aleatorias independientes, cada una con distribución Bernoulli y con probabilidad de éxito p. En notación compacta, $X_i \sim \mathrm{Ber}\,(p)$ para $i = 1, \ldots, n$. Dado un número real $x > 0$ fijo, el problema que se plantea es calcular la probabilidad*

$$P\left(\sum_{i=1}^n X_i \leqslant x\right).$$

Se conoce que la suma de las variables X_i tiene distribución $\mathrm{bin}\,(n,p)$, donde el cálculo de la probabilidad indicada no es fácil, particularmente cuando el valor de n es grande. Sin embargo, podemos aproximar la probabilidad

buscada aplicando el TCL de la manera siguiente:

$$P\left(\sum_{i=1}^{n} X_i \leqslant x\right) = P\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - p}{\sqrt{p\,(1-p)/n}} \leqslant \frac{x/n - p}{\sqrt{p\,(1-p)/n}}\right)$$
$$\approx \Phi\left(\frac{x/n - p}{\sqrt{p\,(1-p)/n}}\right).$$

Dados valores particulares de x y de los parámetros n y p, la última cantidad puede encontrarse, por ejemplo, en tablas de probabilidades de la distribución normal o mediante algún programa de cómputo. ▪

1.3. Generación de números pseudoaleatorios

A grandes rasgos, los números pseudoaleatorios en el intervalo $(0, 1)$ son números reales dentro de dicho intervalo, seleccionados al azar y sin preferencia por alguno. Generarlos no es fácil, ya que se debe tener un mecanismo que garantice que tales valores no sigan un patrón predecible (aleatoriedad) y que cualquier número real dentro del intervalo $(0, 1)$ sea elegible. A continuación se presenta la definición de números aleatorios con la que se trabajará de aquí en adelante.

Definición 1.3 *Diremos que un conjunto de números reales $u_1, \ldots, u_n$ son números aleatorios en el intervalo unitario $(0, 1)$ si provienen de una sucesión de variables aleatorias $U_1, \ldots, U_n$, que son independientes y que tienen idéntica distribución* unif $(0, 1)$.

En resumen, los números aleatorios son valores de una muestra aleatoria unif $(0, 1)$. Más adelante, se podrá observar que la obtención de números aleatorios representa el primer paso para producir valores de variables aleatorias con distribución distinta a la uniforme, por esta razón es muy importante el problema de cómo generarlos. Existen varios mecanismos para intentar obtenerlos, entre los cuales se tienen los métodos manuales, por

ejemplo lanzar monedas o dados, o elegir naipes al azar. También se tienen métodos que requieren algún dispositivo físico, por ejemplo una ruleta, una tómbola o algún dispositivo para registrar sonido electrónico. Estos métodos tienen problemas que han ocasionado que su uso sea menos frecuente. Entre sus desventajas está el tiempo necesario para obtener grandes muestras. Además, en estos métodos manuales y con dispositivos físicos se pueden generar muestras con dependencia y/o que pueden estar sesgadas. La dependencia entre los valores de una muestra puede producir una tendencia y el sesgo ocasiona que no se cumpla la uniformidad deseada. En la Figura 1.2, se observan dos ejemplos de gráficos de dispersión. En el izquierdo se puede observar un ejemplo de una muestra que se obtuvo preguntando valores a un grupo de alumnos y se nota que no es uniforme, mientras que en el derecho se generó la muestra con un programa de cómputo y se puede notar que no existe tendencias ni sesgo. Ambas muestras fueron de tamaño 100.

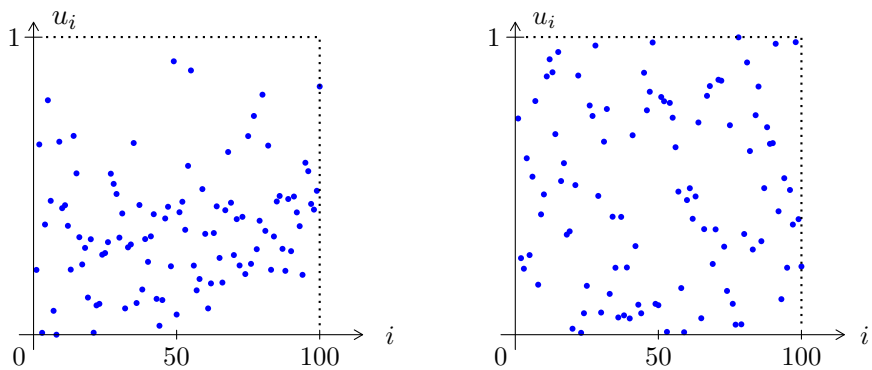


Figura 1.2: Gráficos de dispersión para dos muestras de valores en el intervalo $(0, 1)$.

Método empírico

Un primer mecanismo para intentar generar una muestra de números aleatorios $u_1, \ldots, u_n$ es el método empírico. Es decir, aplicando la intuición y proponerlos de tal forma que cada valor $u_i \in (0, 1)$. Además, buscando que

estos valores sean independientes y estén distribuidos de forma uniforme. Entonces, para que se cumpla el tercer requisito, si se hace una división del intervalo $(0, 1)$ en los intervalos $(0, 1/2)$ y $[1/2, 1)$, aproximadamente la mitad de los números deberían estar en el primer intervalo, y la otra mitad en el segundo intervalo. En general, si se realiza una partición

$$\left\{ 0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}, 1 \right\}$$

del intervalo $[0, 1]$ y en donde $m \leqslant n$, entonces debería existir aproximadamente el mismo número de valores en cada subintervalo de la partición. Por otro lado, para que exista independencia entre los valores, se tendría que evitar que se observe tendencia, o bien, patrones de repetición. Además, como los valores $u_1, \ldots, u_n$ deben provenir de la distribución unif $(0, 1)$, entonces su media muestral y su varianza muestral deberían ser, aproximadamente, $1/2$ y $1/12$, respectivamente. Recordemos que si $U_1, \ldots, U_n$ es una muestra aleatoria de la distribución unif $(0, 1)$, entonces

$$E\left(\frac{1}{n} \sum_{i=1}^{n} U_i \right) = E(U) = \frac{1}{2},$$

y

$$E\left(\frac{1}{n} \sum_{i=1}^{n} (U_i - E(U))^2 \right) = \text{Var}(U) = \frac{1}{12}.$$

Ejemplo 1.3 *Consideremos una muestra numérica consistente de los siguientes* 11 *valores:*

$$0, 0.1, 0.2, \ldots, 0.9, 1.$$

Se puede comprobar que la media muestral es igual a $1/2$ *y que la varianza muestral es* $11/100$*. Estos valores son parecidos a la media y varianza de la distribución* unif $(0, 1)$*. Sin embargo, esta muestra no es independiente porque tiene un claro patrón de incremento en cada observación, aunque parece tener la distribución requerida.* ▪

Este método empírico es útil si el orden de las observaciones no es relevante para operar con la muestra, por ejemplo, para aproximar un valor esperado como se expuso en el ejemplo del dado y la moneda de la Sección 1.1.

Números pseudoaleatorios

En la siguiente sección revisaremos algunos métodos matemáticos que hacen uso de una fórmula determinista para generar números que parecen ser aleatorios. Estrictamente hablando, se viola la hipótesis de aleatoriedad, aunque se pueden obtener muestras que son estadísticamente aleatorias, en el sentido de que se puede usar una prueba de hipótesis apropiada para no rechazar la aleatoriedad de la muestra, dado un nivel de significancia. Sin embargo, veremos que los números generados de esta manera se van a repetir en algún momento. Por lo anterior, llamaremos a éstos números pseudoaleatorios y a los algoritmos de este tipo se les llama generadores de números pseudoaleatorios (*pseudorandom number generators*). Se les denota por PRNG, por sus siglas en el idioma inglés.

Definición 1.4 *Sea $N \geqslant 1$ un número entero. Se dice que una sucesión numérica $x_1, x_2, \ldots$ tiene un patrón de repetición de periodo N si los valores $x_1, \ldots, x_N$ son todos distintos y esta misma secuencia se repite hacia adelante, es decir, $x_{1+k} = x_1, \ldots, x_{N+k} = x_N$, para $k = N, 2N, \ldots$*

La no repetición de un patrón parece ser imposible en los algoritmos matemáticos que se tienen, puesto que dependen de un valor inicial y los valores sucesivos se generan de acuerdo a una función, cuyo valor depende de los valores anteriores y que en algún momento tendrá de nuevo el valor inicial, sin embargo, es posible obtener un periodo de repetición grande.

Por lo general, los métodos generadores que se expondrán a continuación pueden ser eficientes, en el sentido de que se pueden obtener muestras de un tamaño más grande al necesario, dentro de los problemas comunes de simulación estocástica. De ahora en adelante, podemos decir que se estarán generando números pseudoaleatorios cuando se use un algoritmo matemático generador. Lo que se expone en esta sección se enfoca en la generación de números aleatorios haciendo uso de algoritmos deterministas, y aunque realmente se estén generando números pseudoaleatorios, se omitirá frecuentemente el prefijo pseudo. Así, los números aleatorios que se obtengan estarán en función de un valor inicial x_0 llamado semilla y

cada valor u_i dependerá de esta semilla. Es decir, $u_i = g(i, x_0)$, para alguna función g.

Como ya se ha mencionado, estos métodos generan valores que se repiten cada cierto tiempo. Sin embargo, se logra que las muestras obtenidas pasen las pruebas estadísticas de bondad de ajuste y de aleatoriedad. Además, se puede reproducir la obtención de la muestra al establecer el valor de una semilla. A continuación, se exponen los métodos básicos para la generación de números pseudoaleatorios. El lector interesado en un estudio más profundo y en métodos de generación de números aleatorios más sofisticados, puede consultar las siguientes referencias: J. E. Gentle [16], B. D. Ripley [48] y R. Y. Rubinstein [52], por mencionar algunas.

1.4. Generadores congruenciales

El poder generar números aleatorios absolutamente auténticos a través de alguna fórmula matemática o con alguna metodología que se pueda implementar en un programa de cómputo no parece ser posible. Sin embargo, existen algoritmos que generan sucesiones de números con un patrón de repetición grande y que pueden considerarse estadísticamente números aleatorios. En esta sección estudiaremos versiones elementales de los llamados métodos congruenciales para obtener tales números. Para entender estos métodos es conveniente recordar el concepto de congruencia entre dos números.

Definición 1.5 *Sea $M \geqslant 1$ un número entero. Se dice que dos números reales a y b son congruentes módulo M si M divide a la diferencia $a - b$. En tal caso se escribe*

$$a \equiv b \mod M. \tag{1.19}$$

De acuerdo a la definición anterior, los valores de a y b no tienen que ser enteros, pero para nuestro fines nos concentraremos sólo en ese caso. A la divisibilidad de la diferencia $(a-b)$ por el entero M se le denota por $M \mid (a-b)$. Veamos algunos ejemplos

Ejemplo 1.4 *Los siguientes ejemplos muestran que un número a puede ser congruente con b bajo distintos módulos.*

- $8 \equiv 2 \mod 6$, *pues* $6 \mid (8-2)$.
- $8 \equiv 2 \mod 3$, *pues* $3 \mid (8-2)$.
- $8 \equiv 2 \mod 1$, *pues* $1 \mid (8-2)$.

■

Se deja como ejercicio para el lector que demuestre que se cumplen las siguientes propiedades generales de la congruencia entre dos números.

Proposición 1.5 *La operación de congruencia es reflexiva, simétrica y transitiva, es decir, se cumple:*

1. $a \equiv a \mod M$.
2. *Si* $a \equiv b \mod M$, *entonces* $b \equiv a \mod M$.
3. *Si* $a \equiv b \mod M$ *y* $b \equiv c \mod M$, *entonces* $a \equiv c \mod M$.

Supongamos que $a \equiv b \mod M$, en donde $0 \leqslant a < M$. En este contexto, explicaremos a continuación una relación interesante entre el valor de a y el residuo del cociente b/M. Como $M \mid (a-b)$, entonces existe un valor $k \in \mathbb{Z}$ tal que $k\,M = a - b$, o bien, $b = a - k\,M$. En consecuencia se cumple que

$$\frac{b}{M} = \frac{a}{M} - k.$$

Debido a que $0 \leqslant a < M$, tenemos que $0 \leqslant a/M < 1$. Por lo tanto, el número a representa el residuo de la división b/M. El siguiente ejemplo muestra esta situación.

Ejemplo 1.5 *Sean* $a = 7$ *y* $b = 51$. *Entonces* $|a-b| = 44$, *luego*

- $7 \equiv 51 \mod 11$ *y como* $51/11 = 4 + 7/11$, *el residuo de* $51/11$ *es* 7.

- $7 \equiv 51 \mod 22$ *y como* $51/22 = 2 + 7/11$, *el residuo de* $51/22$ *es* 7.
- $7 \equiv 51 \mod 44$ *y como* $51/44 = 4 + 744$, *el residuo de* $51/44$ *es* 7.

■

Como convención, se escribe a veces $a = b \mod M$ para decir que a representa el residuo de b/M, en el caso cuando que $0 \leqslant a < M$.

El primer método para producir números pseudoaleatorios basado en congruencias fue propuesto en 1948 por D. H. Lehmer [35]. Este es un algoritmo para generar valores enteros entre 0 y M, y lo definiremos a continuación.

Generador congruencial lineal simple

Esta es una manera recursiva de generar una sucesión de valores $x_1, x_2, \ldots$ que pueden parecer aleatorios.

Definición 1.6 *Sea $M \geqslant 2$ un entero, y sean a y c dos constantes enteras. Dado un valor entero inicial x_0 tal que $0 \leqslant x_0 < M$, se define la sucesión numérica $x_1, x_2, \ldots$ a través del así llamado generador congruencial lineal*

$$x_i = (a\, x_{i-1} + c) \mod M, \quad i = 1, 2, \ldots \tag{1.20}$$

Es evidente que cada número generado x_i es tal que $0 \leqslant x_i < M$. En consecuencia, algunos de los valores de la sucesión pueden repetirse y que, a lo sumo, aparecerán M valores distintos. Justamente, estos valores distintos son $0, 1, \ldots, M-1$. A los parámetros del generador (1.20) se les conoce con los siguientes nombres:

- Al número entero x_0 se le llama semilla del generador.
- Al número M se le llama módulo del generador.
- Al parámetro a se le llama multiplicador del generador.

- Al parámetro c se le llama incremento del generador.

A la sucesión generada por (1.20) se le llama sucesión de Lehmer. En la Figura 1.3 se muestra la gráfica de la transformación $x_i = (a\, x_{i-1} + c)$ mód M. Se ha dibujado la gráfica con trazos continuos para una mejor visualización pero en realidad de trata de puntos que siguen el patrón lineal. La sucesión de Lehmer se obtiene al aplicar de manera iterada esta transformación a la semilla x_0 obteniendo así un sistema dinámico discreto (i.e. una trayectoria) sobre el conjunto $\{0, 1, \ldots, M-1\}$. Veamos un ejemplo.

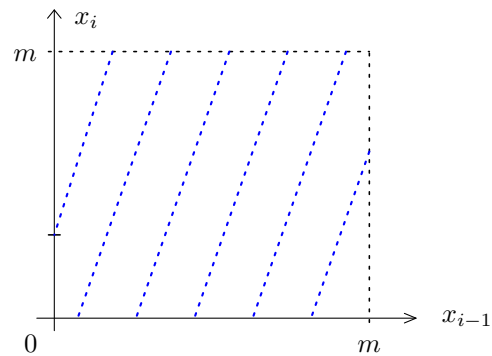


Figura 1.3: Ejemplo de gráfico x_i vs x_{i-1} para números aleatorios generados mediante una sucesión de Lehmer.

Ejemplo 1.6 *Sean* $M = 16$, $a = 5$, $c = 0$ *y* $x_0 = 1$. *Entonces*

$$\begin{aligned} x_1 &= (5 \times 1) \quad \text{mód } 16 = 5, \\ x_2 &= (5 \times 5) \quad \text{mód } 16 = 9, \\ x_3 &= (5 \times 9) \quad \text{mód } 16 = 13, \\ x_4 &= (5 \times 13) \quad \text{mód } 16 = 1, \\ &\vdots \end{aligned}$$

Para $i = 4$ se ha obtenido la semilla $x_0 = 1$ y la sucesión numérica se repite:

$$\underbrace{5, 9, 13, 1,}\ \underbrace{5, 9, 13, 1,}\ \underbrace{5, 9, 13, 1,}\ \cdots$$

■

La Definición 1.4 sobre un patrón de repetición de periodo N es equivalente a la siguiente definición de periodo para una sucesión de Lehmer. Observemos que en la sucesión de referencia no se considera el valor inicial x_0.

Definición 1.7 *Para la sucesión de Lehmer* (1.20)*, el periodo es el entero máximo $N \geqslant 1$ tal que la sucesión $x_1, \ldots, x_N$ tiene valores distintos.*

Como hemos señalado, el periodo máximo de una sucesión de Lehmer es el valor del parámetro $M \geqslant 2$. Para el último ejemplo, es claro que el periodo es $N = 4$. Veamos otro ejemplo.

Ejemplo 1.7 *Sean $M = 100$, $a = 13$, $c = 14$ y $x_0 = 1$. La sucesión que se obtiene aparece abajo. En la iteración $n = 20$ se obtiene la semilla $x_0 = 1$ y la sucesión reinicia nuevamente. El periodo es $N = 20$.*

$x_1 = 27$	$x_6 = 85$	$x_{11} = 79$	$x_{16} = 21$
$x_2 = 65$	$x_7 = 19$	$x_{12} = 41$	$x_{17} = 87$
$x_3 = 59$	$x_8 = 61$	$x_{13} = 47$	$x_{18} = 45$
$x_4 = 81$	$x_9 = 7$	$x_{14} = 25$	$x_{19} = 99$
$x_5 = 67$	$x_{10} = 5$	$x_{15} = 39$	$x_{20} = 1$

■

Números en el intervalo $(0, 1)$

Si a cada valor x_i de una sucesión de Lehmer se le divide entre el módulo M del generador, es decir, $u_i = x_i/M$, para $i = 1, 2, \ldots$ entonces la sucesión $u_1, u_2, \ldots$ toma ahora valores en el conjunto $\{0, 1/M, 2/M, \ldots, (M-1)/M\} \subseteq (0, 1)$ y, como antes, habrá a lo sumo M valores distintos.

Generador congruencial multiplicativo

Definición 1.8 *Tomando $c = 0$ en el generador lineal simple* (1.20)*, se obtiene el así llamado generador multiplicativo*

$$x_i = a\, x_{i-1} \mod M. \tag{1.21}$$

Por ejemplo, sean $a = 7^5$, $M = 2^{31} - 1$ y $x_0 = 2021$. Puede comprobarse que $x_1 = 33966947$, $x_2 = 1799311774$, $x_3 = 168268564$, ... Dividiendo estos valores entre M se obtienen los números $u_1 = 9.41 \times 10^{-7}$, $u_2 = 0.0158171$, $u_3 = 0.8378698$, ...

A continuación se presentan dos resultados que establecen las condiciones para conocer información sobre el periodo de una sucesión obtenida con los generadores congruenciales vistos. Las demostraciones pueden ser encontradas, por ejemplo, en el trabajo de T. E. Hull y A. R. Dobell [23].

Proposición 1.6 *Sean a, c y M números enteros positivos y supongamos que $0 < a < M$. El generador congruencial lineal* (1.20) *tiene periodo M si, y sólo si se cumple que:*

1. $\text{mcd}\,(c, M) = 1$.
2. $a \equiv 1 \mod q$*, para todo valor q que sea factor primo de M.*
3. *Si* $M \equiv 0 \mod 4$ *(M múltiplo de 4), entonces $a - 1$ también debe serlo.*

Sin embargo, debe observarse que un esquema que produzca periodo máximo no necesariamente garantiza que los números generados poseen las características estadísticas deseables en un conjunto de números aleatorios.

Proposición 1.7 *Un generador multiplicativo* (1.21) *tiene periodo máximo de* $N = M - 1$ *si se cumple que:*

1. M *es número primo.*
2. a *es una raíz primitiva de* M*.*

Para entender mejor la segunda condición de la proposición anterior, recordemos que se considera que a es raíz primitiva de M, si el menor valor entero q tal que

$$a^q \equiv 1 \mod M,$$

es $q = M - 1$. Veremos algunos ejemplos de raíces primitivas.

Ejemplo 1.8 *Sean* $M = 31$ *y* $a = 3$*. Entonces*

$$\begin{aligned}
a^2 - 1 &= 8 \text{ pero } 31 \text{ no divide a } 8,\\
a^3 - 1 &= 26 \text{ pero } 31 \text{ no divide a } 26,\\
a^4 - 1 &= 80 \text{ pero } 31 \text{ no divide a } 80,\\
&\vdots\\
a^{29} - 1 &= 68630377364882 \text{ pero } 31 \text{ no divide a } 68630377364882,\\
a^{30} - 1 &= 205891132094648 \text{ y } 31 \text{ sí divide } 205891132094648.
\end{aligned}$$

Por lo anterior, como $3^q \equiv 1 \mod 31$ *para* $q = 31-1$*, entonces el número* 3 *es raíz primitiva de* 31*. Ahora, tomando como semilla a* $x_0 = 13$ *y haciendo los cálculos con la relación* (1.21) *se tiene que la sucesión de Lehmer tiene los siguientes valores,* $x_1, x_2, \ldots, x_{30} = 8, 24, 10, 30, 28, 22, 4, 12, 5, 15, 14, 11, 2, 6, 18, 23, 7, 21, 1, 3, 9, 27, 19, 26, 16, 17, 20, 29, 25, 13$*. Es decir,* $x_{30} = 13$ *igual a la semilla inicial, por lo tanto, el periodo en este caso es* $N = M - 1 = 30$*.* ∎

Ejemplo 1.9 *Sean* $M = 31$ *y* $a = 2$. *Entonces*

$$\begin{aligned} a^2 - 1 &= 3 \text{ pero } 31 \text{ no divide a } 3, \\ a^3 - 1 &= 7 \text{ pero } 31 \text{ no divide a } 7, \\ a^4 - 1 &= 15 \text{ pero } 31 \text{ no divide a } 15, \\ a^5 - 1 &= 31 \text{ y } 31 \text{ sí divide } 31. \end{aligned}$$

Además, 31 también divide a los números $a^{10} - 1$, $a^{15} - 1$, $a^{20} - 1$, $a^{25} - 1$ *y* $a^{30} - 1$, *pero el menor valor de* q *tal que* $a^q \equiv 1 \mod 31$ *es* $q = 5$ *y* $M - 1 \neq 5$, *por lo tanto,* $a = 2$ *no es raíz primitiva de* $M = 31$. ∎

A partir del ejemplo anterior surge la pregunta de si existe una forma sencilla de calcular el periodo de una sucesión de Lehmer. Para el generador multiplicativo (1.21) se conoce que la respuesta es afirmativa en el caso donde M es número primo. El resultado es el siguiente.

Proposición 1.8 *Considere el generador multiplicativo* (1.21) *en donde* M *es un número primo, la semilla* x_0 *es un número entero tal que* $0 < x_0 < M$ *y el factor* a *es un entero positivo. Entonces el periodo de la sucesión de Lehmer está dado por el mínimo valor* $q \in \{1, \ldots, M-1\}$ *tal que*

$$a^q \equiv 1 \mod M.$$

Demostración. Por hipótesis, $x_{n+1} \equiv a\, x_n \mod M$, en donde $0 < x_{n+1} < M$. Luego, $x_1 \equiv a\, x_0 \mod M$ implica que

$$a\, x_1 \equiv a^2\, x_0 \mod M. \tag{1.22}$$

Además, $x_2 \equiv a\, x_1 \mod M$, por lo cual M divide a la diferencia $x_2 - a\, x_1$. Por (1.22), M divide a la diferencia $a\, x_1 - a^2\, x_0$. Entonces M divide a la suma

$$(x_2 - a\, x_1) + (a\, x_1 - a^2\, x_0) = x_2 - a^2\, x_0.$$

Se tiene entonces que

$$x_2 \equiv a^2\, x_0 \mod M.$$

Siguiendo el mismo procedimiento se llega a que

$$x_n \equiv a^n x_0 \mod M. \tag{1.23}$$

Por otro lado, la sucesión $x_0, x_1, x_2, \ldots$ se comienza a repetir por primera vez en el elemento q-ésimo (periodo) cuando $x_q = x_0$, $x_{q+1} = x_1$, $x_{q+2} = x_2, \ldots$ así que, sustituyendo $x_q = x_0$ en (1.23), se llega a

$$x_0 \equiv a^q x_0 \mod M,$$

lo cual implica que

$$a^q \equiv 1 \mod M.$$

■

Para que la proposición anterior tenga sentido, la pregunta natural es ¿siempre existe un valor $q \in \{1, \ldots, M-1\}$ tal que $a^q \equiv 1 \mod M$? Además, ¿el valor de a puede ser cualquier entero positivo? La respuesta a ambas preguntas se tiene con el siguiente resultado demostrado por Fermat en 1640 y cuya demostración es omitida.

Teorema 1.2 *Sea M un número primo y sea a un número entero positivo tal que M no divide al número a. Entonces*

$$a^{M-1} \equiv 1 \mod M. \tag{1.24}$$

Por lo anterior, siempre existe al menos $q = M - 1$ tal que $a^q \equiv 1 \mod M$, para valores de a que no son múltiplos de M. Tampoco es posible que $a = M$, ya que si así fuera se tendría que

$$\frac{M^q - 1}{M} = k,$$

para algún $k \in \mathbb{Z}$, pero eso implicaría que

$$M^{q-1} - k = \frac{1}{M},$$

lo cual es una contradicción porque $M^{q-1} - k \in \mathbb{Z}$.

Ejemplo 1.10 *Para $M = 11$, y para $a = 2, 6, 7, 8, 13, 17, 18, 19, 24, 28, 29, 30$, por mencionar algunos valores, se cumple que el mínimo valor de q tal que $a^q \equiv 1 \mod M$ es $q = M - 1 = 10$ y, por lo tanto, son raíces primitivas de M.* ▪

Generador congruencial múltiple

Esta es una extensión del generador congruencial multiplicativo (1.21). La recursión ahora se efectúa sobre 2 o más valores anteriores. Se usan la letras MRG (*Multiple Recursor Generator*) para denotar a estos generadores de números pseudoaleatorios.

Definición 1.9 *Sea $M \geqslant 2$ un entero y sean $a_1, \ldots, a_k$ constantes enteras en donde $k \geqslant 1$. Sean $x_0, x_1, \ldots, x_{k-1}$ valores enteros iniciales. El generador congruencial múltiple es*

$$x_i = (a_1 x_{i-1} + a_2 x_{i-2} + \cdots + a_k x_{i-k}) \mod M, \quad i \geqslant k. \tag{1.25}$$

Al número k se le conoce como el orden del generador multiplicativo. Por ejemplo, el generador $x_i = a\, x_{i-1} \mod M$ es de orden 1. Para un generador congruencial múltiple donde M sea un número primo, parecería ser más fácil encontrar sucesiones con un periodo grande. Sin embargo, aún se desconoce un resultado que proporcione condiciones suficientes para tal afirmación.

Ejemplo 1.11 *El generador congruencial múltiple que aparece abajo fue el objeto de estudio del trabajo de P. L'Ecuyer, F. Blouin y R. Couture de 1993. Se requiere de valores iniciales $x_0, \ldots, x_4$. El estudio puede consultarse en [36].*

$$x_i = 107374182\, x_{i-5} \mod (2^{31} - 1), \quad i \geqslant 5. \tag{1.26}$$

▪

Generador congruencial de Fibonacci

Este es un ejemplo particular de un generador congruencial múltiple.

Definición 1.10 *Sea j y k dos enteros tales que $1 \leqslant j < k$. Sean $x_0, x_1, \ldots, x_{k-1}$ valores enteros iniciales tales que $0 \leqslant x_i < M$, para $i = 0, \ldots, k-1$. El generador congruencial de Fibonacci es*

$$x_i = (x_{i-j} + x_{i-k}) \mod M, \quad i \geqslant k. \tag{1.27}$$

El nombre de este generador proviene del hecho de que cuando $j = 1$, $k = 2$, $M = \infty$ y los valores iniciales son $x_0 = 0, x_1 = 1$, el generador (1.27) produce la famosa sucesión de Fibonacci

$$x_i = x_{i-1} + x_{i-2}, \quad i \geqslant 2.$$

Ejemplo 1.12 *Sean $x_0 = 0$, $x_1 = 1$ y $M = 97$. El generador de Fibonacci dado por $x_i = (x_{i-1} + x_{i-2}) \mod M$ produce los valores*

$x_0 = 0$	$x_4 = 3$	$x_8 = 21$	$x_{12} = 47$
$x_1 = 1$	$x_5 = 5$	$x_9 = 34$	$\vdots$
$x_2 = 1$	$x_6 = 8$	$x_{10} = 55$	
$x_3 = 2$	$x_7 = 13$	$x_{11} = 89$	

Se puede comprobar que el periodo de esta sucesión es $N = 196$. En la Tabla 1.1 se muestra el periodo N para otros valores de M.

M	94	95	96	97	98	99	100	101	102	103
N	96	180	48	196	336	120	300	50	72	208

Tabla 1.1: Periodo N para distintos valores de M.

■

Esto concluye nuestra breve introducción al tema de generadores congruenciales de números pseudoaleatorios. Debe mencionarse que existen muchos otros mecanismos sofisticados basados en congruencias que logran generar

números aleatorios con características casi idóneas. Por ejemplo, en el trabajo de los matemáticos M. Matsumoto y T. Nishimura [37] de 1998 se expone el así llamado generador *Mersenne Twister* y cuyo periodo es $N = 2^{19937}-1$.

1.5. Elementos gráficos

Los diagramas de dispersión, el gráfico de u_i vs u_{i-1}, los histogramas, el gráfico de la función de distribución empírica y el correlograma, son gráficas comúnmente estudiados en los cursos de estadística descriptiva y series de tiempo. Estas gráficas son útiles para realizar un primer análisis de las características de los números pseudoaleatorios que se dispongan. Estos elementos visuales pueden ayudar a generar conjeturas que después se pueden corroborar con pruebas estadísticas. A continuación, se mencionan algunos usos sugeridos para cada gráfica. Es posible que el lector ya esté familiarizado con la mayoría de ellas.

1. Diagrama de dispersión. Se grafica el valor del índice i contra cada valor u_i de la muestra, para $i = 1, \ldots, n$. Se obtiene así una nube de puntos a la que se le denomina diagrama de dispersión. Véase la parte izquierda de la Figura 1.4. Cuando los puntos son realmente aleatorios, su distribución es homogénea en el eje vertical y la sucesión de puntos no presenta ninguna tendencia como función del índice i. La nube de puntos que se muestra en la figura mencionada parece cumplir con estas características.

2. Gráfico de u_i vs u_{i-1}. Como el nombre lo indica, esta gráfica se obtiene al indicar en el plano la posición de los $n-1$ puntos con coordenadas (u_{i-1}, u_i), para $i = 2, \ldots, n$. Observe que todos los puntos pertenecen al cuadrado unitario $(0,1) \times (0,1)$. La parte derecha de la Figura 1.4 muestra esta gráfica para el conjunto de puntos a partir de los cuales se conformó el diagrama de dispersión de la izquierda. Resulta revelador que la gráfica de puntos (u_{i-1}, u_i) muestra una perfecta dependencia entre u_i y u_{i-1}, a pesar de que el gráfico de dispersión parecía ser el de un conjunto de puntos al azar.

3. Histograma. Este un gráfico que ayuda a detectar posibles problemas con la uniformidad de los datos. Se construye tomando una partición

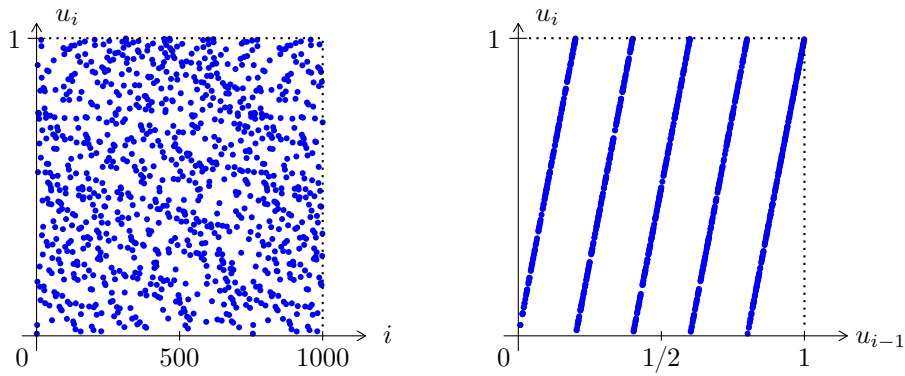


Figura 1.4: Gráfico de dispersión y gráfico de u_i vs u_{i-1}.

del intervalo $(0, 1)$ y dibujando una barra sobre cada subintervalo de la partición con altura el número de valores u_i que pertenecen a ese subintervalo. En la Figura 1.5 se muestra una de estas gráficas. Cuando los datos efectivamente se distribuyen de manera uniforme, las alturas de las barras tienden a ser semejantes. En ocasiones puede ser útil comparar el histograma con la función de densidad unif $(0, 1)$. Para esto se puede realizar un histograma de probabilidad (o histograma de frecuencias relativas), el cual se obtiene al reescalar la altura de las barrras del histograma original (histograma de frecuencias absolutas) para que la suma de sus áreas sea igual a uno.

4. Función de distribución empírica. Junto con el histograma, la gráfica de esta función ayuda a visualizar la posible distribución uniforme de un conjunto de datos numéricos. Su especificación aparece más abajo en la Definición 1.11. Cuando los datos efectivamente tienen distribución uniforme, la gráfica de la función de distribución empírica tiende a asemejarse a la línea recta de la función identidad en el intervalo $(0, 1)$. Véase la gráfica escalonada de la Figura 1.5.

5. Correlograma. Este gráfico es usado con frecuencia en el análisis de series de tiempo. Ayuda a detectar alguna posible dependencia lineal entre las observaciones. Este gráfico se calcula como aparece en la

Definición 1.12.

A continuación, se recuerdan las definiciones de función de distribución empírica y del correlograma.

Definición 1.11 *Sea $x_1, \ldots, x_n$ una colección de números. La función de distribución empírica es*

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(x_i \leqslant x)} = \frac{\#\{x_i \leqslant x\}}{n}, \quad x \in \mathbb{R}.$$

Es decir, para cada número real x, la función $F_n(x)$ representa el número de datos que quedan a la izquierda de x dividido entre el número total de datos. La expresión $\#A$ denota la cardinalidad del conjunto A. La gráfica de una función de distribución empírica es escalonada creciente como la que se muestra en la Figura 1.5 .

Definición 1.12 *La función de autocorrelación de una colección de observaciones $x_1, \ldots, x_n$ se denota por $r(\tau)$ y se define como aparece abajo para valores $\tau = 0, 1, \ldots, n-1$. A la gráfica de esta función se le llama correlograma.*

$$\tau \mapsto r(\tau) := \frac{\displaystyle\sum_{t=1}^{n-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x})}{\displaystyle\sum_{t=1}^{n} (x_t - \bar{x})^2}.$$

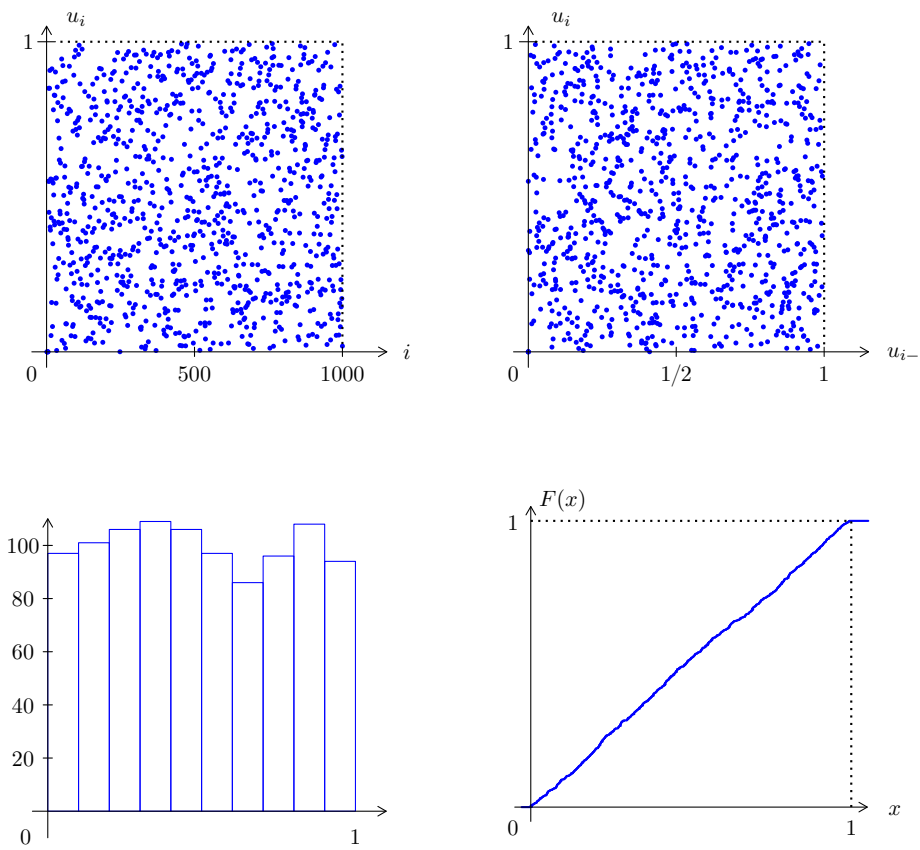


Figura 1.5: Criterios gráficos: Diagrama de dispersión, gráfico de u_i vs u_{i-1}, histograma y función de distribución empírica. Sucesión generada con la ecuación (1.26).

Debe advertirse que existen más gráficos que ayudan a determinar visualmente si un conjunto de números pseudoaleatorios cumple con algunas de las características deseables de un conjunto de números realmente aleatorios. Los que hemos presentando, sin embargo, son los de mayor uso y de fácil interpretación.

Ejemplo 1.13 *En la Figura 1.5 se pueden observar el gráfico de dispersión, el gráfico de u_i vs u_{i-1}, el histograma y la gráfica de la función de distribución empírica, de una colección de números obtenidos mediante el generador congruencial múltiple de orden* 5 *que aparece en la ecuación* (1.26)*, con las semillas iniciales $x_0 = 11$, $x_1 = 17$, $x_2 = 29$, $x_3 = 31$ y $x_4 = 37$.* ▪

1.6. Pruebas para números pseudoaleatorios

En esta sección se exponen distintas formas de evaluar la calidad de los números pseudoaleatorios. Existen pruebas de hipótesis para evaluar tanto la uniformidad de los números como la aleatoriedad o independencia. Para la uniformidad se utilizan pruebas de bondad de ajuste. En esta sección hablaremos sobre las pruebas de Kolmogorov-Smirnov y la prueba χ^2. Por otro lado, existen varios mecanismos para analizar si una muestra de números proviene de variables aleatorias independientes. Los mecanismos que en esta sección se presentan se enfocan en detectar "evidencia de dependencia". Si existe tal evidencia estadística, una técnica para intentar "generar independencia" consistirá en realizar una permutación de los valores de la muestra. Para detectar posible dependencia se expone la prueba estadística de Cox-Stuart, y para probar la aleatoriedad veremos la prueba de rachas.

Prueba de Kolmogorov-Smirnov

Esta prueba de hipótesis, estudiada en la estadística no paramétrica, puede ser utilizada para saber si los hipotéticos números aleatorios que se tengan provienen o no de una distribución uniforme. A continuación se muestran los elementos de la prueba.

Definición 1.13 *Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria numérica que proviene de una función de distribución continua $F(x)$ desconocida, y sea $F^*(x)$ una función de distribución propuesta. Sea $F_n(x)$ la función de distribución empírica de la muestra. Se dice que*

$$T = \sup_{1 \leqslant i \leqslant n} \{|F^*(x_i) - F_n(x_i)|\}, \tag{1.28}$$

es el estadístico de Kolmogorov para la prueba de hipótesis siguiente:

$$H_0 : F(x) = F^*(x) \text{ para todo valor de } x \in \mathbb{R}.$$
$$vs$$
$$H_a : F(x) \neq F^*(x) \text{ para al menos un valor de } x \in \mathbb{R}.$$

El criterio de decisión en este caso consiste en rechazar la hipótesis nula H_0 si el estadístico $T > q_{1-\alpha}$, con un nivel de significancia de $\alpha \in (0, 1)$. Los valores del cuantil $q_{1-\alpha}$ pueden ser consultados, por ejemplo, en la tabla A13 de W. J. Conover [5]. Otro criterio de decisión equivalente consiste en rechazar la hipótesis nula H_0 si el valor del *p-value* es menor a α, donde

$$p\text{-}value = 2 \left[T \sum_{j=0}^{\lfloor n(1-T) \rfloor} \binom{n}{j} (1 - T - j/n)^{n-j} (T + j/n)^{j-1} \right]. \tag{1.29}$$

Aplicación para números aleatorios

Sean $u_1, u_2, \ldots, u_n$ una muestra aleatoria numérica donde cada valor $u_i \in (0, 1)$. Se presume que estos números siguen una distribución uniforme en dicho intervalo, es decir,

$$F^*(x) = \begin{cases} 0 & \text{si } x \leqslant 0, \\ x & \text{si } 0 < x < 1, \\ 1 & \text{si } x \geqslant 1. \end{cases}$$

Entonces, en este caso el estadístico de prueba se calcula como

$$T = \sup_{1 \leqslant i \leqslant n} \{|u_i - F_n(u_i)|\}.$$

Así, la prueba tendrá como hipótesis a

$$H_0 : F(x) = F^*(x) \text{ para todo valor de } x \in \mathbb{R}.$$
$$\text{vs}$$
$$H_a : F(x) \neq F^*(x) \text{ para al menos un valor de } x \in \mathbb{R}.$$

Es decir, la hipótesis nula es que la muestra proviene de la distribución unif $(0, 1)$.

Ejemplo 1.14 *Supongamos que se tienen los siguientes valores* $x_1, \ldots, x_{20}$,

0.6078023	*0.9073220*	*0.5299141*	*0.8476512*
0.7802851	*0.9418568*	*0.3448128*	*0.4627547*
0.6645587	*0.3105249*	*0.3324742*	*0.8224943*
0.8121696	*0.5116627*	*0.1037860*	*0.6346132*
0.5954693	*0.8505040*	*0.4974272*	*0.8757268*

En la Figura 1.6 se observa la función de distribución empírica y su distancia más grande con respecto a la función de distribución unif $(0, 1)$,

$$T = \sup_{1 \leqslant i \leqslant n} \{|x_i - F_n(x_i)|\} = 0.2627547.$$

El p-value $= 0.1045$ *se calcula con la fórmula* (1.29). *En este caso, tomando una significancia de* $\alpha = 0.05$, *la decisión es no rechazar la hipótesis nula de que los valores de la muestra provienen de una distribución* unif $(0, 1)$. *Es decir, no hay evidencia estadística para rechazar los valores como una muestra uniforme en el intervalo* $(0, 1)$. ▪

Prueba ji-cuadrada

La llamada prueba ji-cuadrada (χ^2) puede ser utilizada como prueba de bondad de ajuste. Su uso permite observar si existe evidencia estadística de que los hipotéticos números aleatorios provengan de una distribución uniforme. A continuación se muestran los elementos de la prueba.

Supongamos que se tiene una muestra de n valores nominales u ordinales que pueden ser clasificados dentro de alguna de c clases. Si se sospecha que p_j^* es la probabilidad de que una observación de la muestra pertenezca a la clase j $(j = 1, \ldots, c)$, entonces en la prueba χ^2 se contrastan las hipótesis

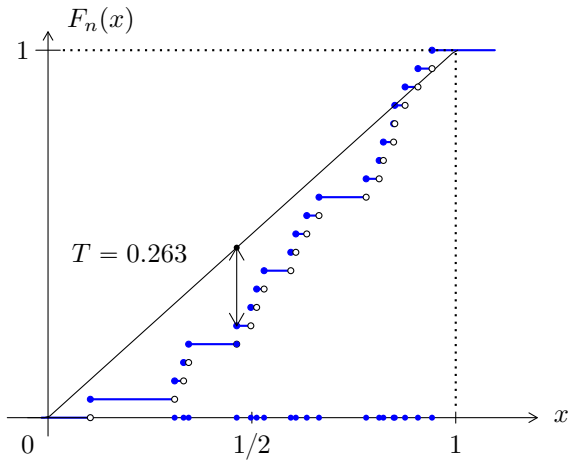


Figura 1.6: Funciones de distribución unif $(0, 1)$ y empírica.

$$H_0 : p_j = p_j^* \text{ para valores de } j = 1, \ldots, c,$$
$$\text{vs}$$
$$H_a : p_j \neq p_j^* \text{ para al menos un valor de } j,$$

donde p_j representa la verdadera probabilidad (desconocida) de que una observación de la muestra pertenezca a la clase j. Para entender el estadístico de prueba, primero se deben calcular los valores de $O_1, O_2, \ldots, O_c$, una sucesión de enteros no negativos tales que cada O_j es igual al número de elementos de la muestra en la clase j. Si H_0 es verdadera, entonces el número esperado de elementos en la clase j debe ser $E_j = n\, p_j^*$. Observemos que

- $\displaystyle\sum_{j=1}^{c} O_j = n.$
- $\displaystyle\sum_{j=1}^{c} E_j = \sum_{j=1}^{c} n\, p_j = n.$

Definición 1.14 *El estadístico de la prueba* χ^2 *se define como*

$$T = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}. \tag{1.30}$$

Se puede comprobar que para muestras de tamaño grande, la distribución nula de T bajo H_0, es decir, suponiendo que H_0 es cierta, es asintóticamente χ^2. En específico,

$$T \,|\, H_0 \sim \chi^2_{c-1}.$$

El criterio de decisión es rechazar la hipótesis nula H_0 si el estadístico de prueba cumple con la desigualdad $T > \chi^2_{(c-1),1-\alpha}$, donde $\chi^2_{(c-1),1-\alpha}$ representa el cuantil $1-\alpha$ de la distribución χ^2 con $c-1$ grados de libertad. De forma alternativa, haciendo uso del *p-value*, el criterio consiste en rechazar la hipótesis nula H_0 si se cumple que *p-value* $< \alpha$, donde *p-value* $= P(Y > T)$, suponiendo que $Y \sim \chi^2_{c-1}$.

Aplicación para números aleatorios

Sea $u_1, u_2, \ldots, u_n$ una muestra aleatoria donde cada $u_i \in (0,1)$. Se definen 10 clases $C_1, \ldots, C_{10}$ dadas por

$$C_j = \left(\frac{j-1}{10}, \frac{j}{10} \right].$$

Entonces se tienen los siguientes elementos,

$$\begin{aligned} O_j &= \#\left\{u_i \in C_j\right\}, \\ p_j^* &= \frac{1}{10}, \\ E_j &= \frac{n}{10}, \quad j = 1, \ldots, 10. \end{aligned}$$

Con las siguientes hipótesis,

$$H_0 : p_j = \frac{1}{10} \text{ para todo valor de } j$$
$$\text{vs}$$
$$H_a : p_j \neq \frac{1}{10} \text{ para al menos un valor de } j.$$

Es decir, la hipótesis nula establece que hay evidencia de uniformidad. El criterio de decisión será rechazar H_0 si se cumple que $T > \chi^2_{(9),1-\alpha}$, o bien, si *p-value* $< \alpha$, donde *p-value* $= P(Y > T)$ y $Y \sim \chi^2_{(9)}$.

Ejemplo 1.15 *Supongamos que se tienen los mismos valores que en el Ejemplo 1.14. En la Figura 1.7 se observa el histograma de frecuencia de los valores de la muestra en clases de longitud 0.1. Los valores de* $(O_1, \ldots, O_{10})$ *son* $(0, 1, 0, 3, 2, 3, 3, 1, 5, 2)$ *y* $E_j = 20/10 = 2$. *Entonces*

$$T = \sum_{j=1}^{10} \frac{(O_j - 2)^2}{2} = 11.$$

El p-value $= 0.2757089$ *es igual a* $P(Y > T)$, *donde* $Y \sim \chi^2_{(9)}$. *En este caso, tomando una significancia de* $\alpha = 0.05$, *la decisión de nuevo es no rechazar la hipótesis nula de que los valores de la muestra provienen de una distribución* unif $(0, 1)$. *Es decir, no hay evidencia estadística para rechazar los valores como una muestra uniforme en el intervalo* $(0, 1)$. ∎

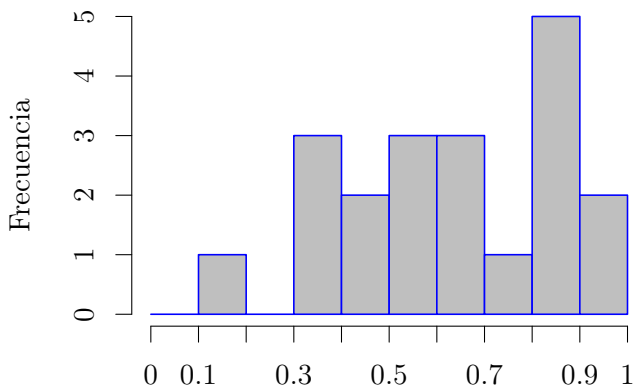


Figura 1.7: Histograma de la muestra del Ejemplo 1.15 .

Prueba de Cox-Stuart

La prueba de Cox-Stuart es útil para conocer si existe evidencia estadística de una tendencia en la muestra, lo cual no se adaptaría a la hipótesis de independencia que buscamos en los números pseudoaleatorios. Supongamos una muestra $u_1, \ldots, u_n$.

- Si n es par, sea $l = \dfrac{n}{2}$.
- Si n es impar, sea $l = \dfrac{n-1}{2}$.

Consideremos ahora las parejas que se forman con los conjuntos

- $u_1, u_2, \ldots, u_l$ y $u_{l+1}, u_{l+2}, \ldots, u_n$, siendo n número par, o bien,
- $u_1, u_2, \ldots, u_l$ y $u_{l+2}, u_{l+3}, \ldots, u_n$, cuando n es número impar.

Es decir,

- $(u_1, u_{l+1}), \ldots, (u_l, u_n)$, n par,
- $(u_1, u_{l+2}), \ldots, (u_l, u_n)$, n impar.

Ahora, a cada pareja $(u_i, u_{i+\tau})$, se le asigna una etiqueta o calificación de acuerdo a la siguiente regla:

- $'+'$ si $u_i < u_{i+\tau}$ (incremento).
- $'='$ si $u_i = u_{i+\tau}$ (empate).
- $'-'$ si $u_i > u_{i+\tau}$ (decremento).

La idea es descartar los empates y sólo considerar las parejas donde hubo algún cambio. Se define entonces

$$m = \text{número de parejas con etiqueta } '+' \text{ o } '-'.$$

Así, se obtiene una muestra $S_1, S_2, \ldots, S_m$ donde $S_i = 1$ cuando la pareja i tiene etiqueta $'+'$, o bien, $S_i = 0$ cuando la pareja i tiene etiqueta $'-'$. Las hipótesis que se contrastan son

$$H_0 : \frac{1}{m}\sum_{i=1}^{m} S_i = \frac{1}{2} \quad \text{vs} \quad H_a : \frac{1}{m}\sum_{i=1}^{m} S_i \neq \frac{1}{2}.$$

Tales hipótesis se interpretan como sigue.

H_0: No existe evidencia de tendencia.
vs
H_a: Existe evidencia de tendencia.

Esta prueba es un caso particular de las pruebas binomiales y el estadístico de prueba está dado por el número de signos positivos, es decir,

$$T = \sum_{i=1}^{m} S_i.$$

Sea Y una variable aleatoria con distribución bin $(m, 1/2)$. Definimos

$$p_1 = \sum_{x=0}^{T} P(Y = x),$$

$$p_2 = \sum_{x=T}^{m} P(Y = x),$$

y

$$p\text{-}value = 2 \cdot \text{mín}\left\{p_1, p_2\right\}.$$

El criterio de decisión consiste en rechazar la hipótesis nula si el valor del *p-value* es menor a α, donde α representa el nivel de significancia de la prueba.

Ejemplo 1.16 *Supongamos que se tienen los mismos valores que en los dos ejemplos anteriores. En la Figura 1.8 se observa el diagrama de dispersión de la muestra, justo después de la décima observación se encuentra una línea vertical que separa la primera mitad de puntos para realizar las comparaciones. No parece haber ni tendencia a la baja ni tendencia a la alza. Para aplicar la prueba de Cox-Stuart se tiene que* $n = 20$, $l = 10$ *y no hay empates al comparar la primera mitad de los datos con la segunda mitad. Por lo anterior,* $m = 10$ *y se obtiene que*

$$T = \sum_{i=1}^{10} S_i = 4.$$

El lector puede verificar que sólo en las parejas 3, 7, 8 y 10 los datos de la primera mitad fueron menores a los respectivos de la segunda mitad. Por otro lado, sea $Y \sim \text{bin}\,(10, 1/2)$ *y*

$$p_1 = \sum_{x=0}^{4} P(Y = x) = 0.377,$$

$$p_2 = \sum_{x=4}^{10} P(Y = x) = 0.828,$$

y

$$p\text{-}value = 2 \cdot \text{mín}\,\{p_1, p_2\} = 0.754.$$

En este caso, tomando una significancia de $\alpha = 0.05$*, la decisión es no rechazar la hipótesis nula debido a que el valor del p-value es mayor al valor de* α*. Por lo tanto, se puede decir que no existe evidencia de alguna tendencia.*

■

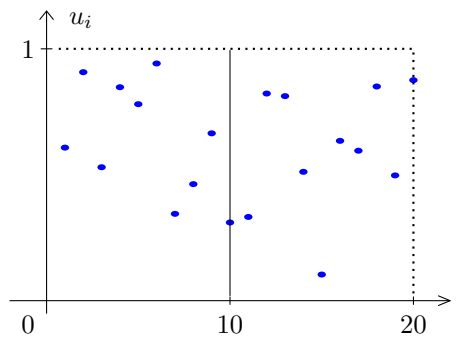


Figura 1.8: Diagrama de dispersión de la muestra del Ejemplo 1.16 .

Prueba de rachas

Esta prueba contrasta las siguientes hipótesis

H_0: existe aleatoriedad
vs
H_a: no existe aleatoriedad.

El criterio de decisión consiste en rechazar H_0 si sucede cualquiera de las dos condiciones siguientes

$$|Z| \geqslant q_{1-\alpha/2}, \text{ o bien, } \textit{p-value} < \alpha, \tag{1.31}$$

donde

- $Z = (R - E(R))/\sqrt{\text{Var}(R)}$.
- R representa el número de rachas. Si la variable q_2 representa la mediana de la muestra, entonces se considera que hubo una racha cuando para algún valor de $u_i \in [0, q_2]$ se tiene que $u_{i+1} \in (q_2, 1]$; o viceversa, cuando para algún valor de $u_i \in (q_2, 1]$ se tiene que $u_{i+1} \in [0, q_2]$.
- $E(R) = (2n_1n_2)/(n_1 + n_2) + 1$, donde n_1 representa el número de observaciones en $[0, q_2]$ y n_2 es el número de observaciones en $(q_2, 1]$.
- $\text{Var}(R) = (2n_1n_2(2n_1n_2 - n_1 - n_2))/\left[(n_1 + n_2)^2(n_1 + n_2 - 1)\right]$.
- $q_{1-\alpha/2}$ representa el cuantil $1 - \alpha/2$ de una distribución $\text{N}(0, 1)$.
- Finalmente, $\textit{p-value} = 2\,P(Y > Z)$, donde $Y \sim \text{N}(0, 1)$.

En caso de que los valores de n_1 y de n_2 sean de al menos once, se puede usar como distribución nula para el estadístico Z a la distribución normal estándar. Si n_1 y n_2 son menores o iguales a 10, entonces se puede consultar la distribución exacta de R, la cual se puede encontrar, por ejemplo, en la tabla 10 del texto de D. Wackerly, W. Mendenhall y R.L. Scheaffer [59]. Además, en las páginas 777 a 782 del libro citado, más detalles relacionados a la prueba pueden ser consultados.

Esta prueba viene implementada en el paquete `randtests` del programa estadístico `R` a través de la función `run.test()`, la cual devuelve el valor del estadístico de prueba y el *p-value*.

En lo que resta del libro se supondrá que se tiene un buen generador de números pseudoaleatorios y ahora el objetivo será producir valores de variables aleatorias con una distribución conocida.

1.7. Ejercicios

Elementos de probabilidad

1. Suponga que se realizan 100 lanzamientos de una moneda justa. En cada ocasión si en un lanzamiento se obtiene cruz, entonces se repite el lanzamiento y se registra el nuevo resultado. Sin embargo, si se obtiene cara se registra el resultado inmediatamente.

   a) Calcule la probabilidad de que en un lanzamiento cualquiera se registre cruz.

   b) Calcule el número esperado de ocasiones en que se registra cruz.

   c) Explique de qué forma están relacionados los dos valores que se obtienen en los primeros dos incisos.

2. Sea X una variable aleatoria con distribución bin $(5, 1/2)$ y sea Y otra variable aleatoria con distribución bin $(8, 1/2)$. Suponga que X y Y son independientes. Calcule:

   a) $E(Y - X)^2$.
   b) $\text{Var}(Y - X)$.
   c) $E(Y - 2X)^2$.
   d) $\text{Var}(Y - 2X)$.

3. En cierto consultorio médico, se tienen 60 registros del número de pacientes que fueron atendidos diariamente durante dos meses. En 30 ocasiones se atendieron 15 pacientes, en 10 ocasiones se atendieron 14 pacientes, en 20 ocasiones se atendieron menos de 14 pacientes. Con estos datos, ¿es posible calcular el valor esperado de pacientes en un día cualquiera? De ser afirmativa la respuesta, calcule tal valor esperado. De ser negativa, mencione por qué no es posible hacerlo.

4. Defina $\tau = \text{mín}\{n \geqslant 0 : X_n = 2\}$ como el tiempo del primer arribo al estado 2 en la cadena de Markov del ejemplo de esta sección. Entonces, el valor de $m_k := E(\tau \,|\, X_0 = k)$ es el tiempo esperado de arribo partiendo del estado k. Resuelva el siguiente sistema de ecuaciones para encontrar el valor de m_0, el cual es el valor exacto del número esperado

de lanzamientos de una moneda hasta obtener dos unos consecutivos.

$$\begin{aligned} m_0 &= 1 + (1/2)m_0 + (1/2)m_1 \\ m_1 &= 1 + (1/2)m_0. \end{aligned}$$

5. Exprese en términos de valores esperados las siguientes integrales completando alguna función de densidad adecuada.

   a) $\int_0^{10} e^{-x^2} dx.$ *c)* $\int_{-\infty}^{\infty} e^{-x^2} dx.$

   b) $\int_0^{\infty} e^{-x^2} dx.$ *d)* $\int_0^{1} e^{-4x^2} dx.$

6. Sea X una variable aleatoria discreta con valores $0, 1, \ldots$ y con esperanza finita. Demuestre que

$$E(X) = \sum_{k=0}^{\infty} P(X > k).$$

7. Demuestre que la covarianza es un operador bilineal, es decir, demuestre que

$$\text{Cov}\left(\sum_{i=1}^{n} \alpha_i X_i, \sum_{j=1}^{m} \beta_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j \text{Cov}\left(X_i, Y_j\right).$$

Generación de números pseudoaleatorios

8. Demuestre la Proposición 1.5, es decir, demuestre que la operación congruencia es reflexiva, simétrica y transitiva.

9. Sea $k \neq 0$ un número entero y sea c cualquier número real. Defina $X \equiv (kY + c) \mod 1$, en donde $Y \sim \text{unif}\,(0, 1)$. Demuestre que

$$X \sim \text{unif}\,(0, 1).$$

10. Considerando el conjunto de números enteros entre 1 y 100, realice el siguiente método de selección: Escoja un número primo, después un número par, después otro número primo distinto al inicial y finalmente un múltiplo de 13. Repita el procedimiento anterior en 4 ocasiones más. Siempre proponga números distintos a los que ya escribió. Al final se tendrá una muestra de 20 números. Divida los números de la muestra entre 100 y obtenga una muestra de 20 valores entre 0 y 1. Realice un gráfico de dispersión de los números obtenidos y comente si parecen ser valores independientes y que tengan distribución unif $(0, 1)$.

11. Coloque 10 fichas o monedas del mismo tamaño numeradas del 1 al 10 en una bolsa. Realice en 20 ocasiones el siguiente experimento:

    - Extraiga una ficha al azar, escriba el número que obtuvo y regrese la ficha a la bolsa.
    - Extraiga otra ficha al azar, escriba el número que obtuvo y regrese la ficha a la bolsa.
    - Realice la multiplicación de los números obtenidos y divida el resultado entre 100.
    - Registre el resultado como su número pseudoaleatorio.

    Realice un gráfico de dispersión de los números obtenidos y mencione si parecen tener distribución unif $(0, 1)$ y ser aleatorios. ¿Esta muestra tiene un mejor aspecto que la obtenida en el ejercicio anterior? Mencione las razones por las que esta muestra debe considerarse una mejor muestra de números pseudoaleatorios. ¿Puede considerarse una auténtica muestra de números aleatorios?

Generadores congruenciales

12. Escriba un programa de cómputo que obtenga los 20 primeros valores del generador congruencial que aparece abajo, en donde $a = 13$, $c = 14$, $M = 100$ y $x_0 = 1$. Compruebe que el periodo es $N = 20$.

$$x_i = a\, x_{i-1} + c \mod M.$$

13. Demuestre que $x_{i+k} = a^k\, x_i \mod M$, para $k \geqslant 1$, para la sucesión de Lehmer dada por la ecuación

$$x_i = a\, x_{i-1} \mod M.$$

14. Usando el método congruencial lineal simple (1.20), genere una muestra de 1000 números aleatorios. Encuentre los valores de x_0, a, c y M tales que el periodo sea mayor o igual a 1000. Usando un teorema de los que se han expuesto, compruebe que se cumplen las condiciones necesarias para obtener el periodo de la sucesión. Además, agregue el cálculo de la media y la varianza muestral, y los gráficos de dispersión e histograma.

15. Realice la siguiente transformación con la muestra del ejercicio anterior: $y_i = -\ln u_i$. Realice el diagrama de dispersión y el histograma (con al menos 10 clases) de la nueva sucesión $y_1, y_2, \ldots, y_{1000}$, y calcule la media y la varianza muestral. Viendo los nuevos gráficos y resultados, ¿tiene alguna conjetura de cómo se distribuye la variable aleatoria relacionada a esta nueva sucesión?

16. Considere el generador congruencial de Fibonacci

$$x_i = (x_{i-1} + x_{i-2}) \mod M,$$

en donde $x_0 = 0$ y $x_1 = 1$. Compruebe que el periodo es $N = 336$ en los casos $M = 98$, $M = 2 \times 98 = 196$, $M = 3 \times 98 = 294$ y $M = 4 \times 98 = 392$. Sin embargo, para $M = 5 \times 98 = 490$, el periodo es $N = 1680$.

17. Escriba un programa de cómputo que obtenga los primero 200 valores del generador congruencial de L'Ecuyer que aparece especificado en la ecuación (1.26). Escoja los valores iniciales $x_0, \ldots, x_4$ a su elección.

18. Con los valores obtenidos en el ejercicio anterior, genere los números aleatorios $u_1, \ldots, u_{200}$. Ahora genere los valores $y_1, \ldots, y_{200}$ de una variable aleatoria con distribución Ber $(1/2)$ usando el siguiente criterio: si $u_i \leqslant 1/2$, entonces $y_i = 0$ y en otro caso $y_i = 1$. Calcule finalmente la varianza muestral de $y_1, \ldots, y_{200}$ para verificar que es aproximadamente $1/4$.

Pruebas para números pseudoaleatorios

19. Usando cartas (naipes) que tengan numeración, genere una muestra numérica de tamaño 50 siguiendo la siguiente metodología. Mezcle las cartas, escoja una al azar y registre el valor obtenido. Regrese la carta al mazo aleatoriamente. Repita lo anterior hasta obtener 50 valores. Finalmente, divida la sucesión de valores obtenidos entre el número más grande de ellos para conseguir una muestra de valores entre el cero y uno. Ahora, realice el cálculo de la media y de la varianza muestral y también realice un diagrama de dispersión. Escriba un análisis de los valores y diga si podrían ser números aleatorios.

20. Usando un generador recursivo múltiple de segundo orden, genere una muestra de números pseudoaleatorios. Encuentre valores iniciales x_0, x_1 y coeficientes a_1 y a_2 y el valor de M para lograr que la sucesión tenga un periodo mayor a 1000.

    a) Realice un código en algún lenguaje de programación de su preferencia para generar la sucesión $x_0, x_1, x_2, \ldots$ que se menciona y para calcular su periodo.

    b) Divida la sucesión $x_1, x_2, \ldots, x_{1000}$ entre M para obtener la muestra de números pseudoaleatorios $u_1, u_2, \ldots, u_{1000}$. A continuación realice un histograma de probabilidad (cuando el área de las barras es igual a uno) para decidir si se parece a la función de densidad unif $(0, 1)$. Por ejemplo, en el lenguaje de programación `R`, este gráfico se obtiene con el comando `hist(x,freq=FALSE)`, en donde `x` representa el vector con los valores de la muestra.

    c) Realice el gráfico de dispersión y el gráfico de u_i *vs* u_{i+1} y realice un análisis mencionando si se detectan indicios de que no son números pseudoaleatorios.

    d) Realice el gráfico de la función de distribución empírica de la muestra $u_1, u_2, \ldots, u_{1000}$ y agregue la función de distribución unif $(0, 1)$ para observar la calidad del ajuste.

    e) Realice una prueba de Kolmogorov-Smirnov para saber si existe evidencia estadística de uniformidad en los números pseudoaleatorios, calculando el estadístico de prueba y el *p-value*. Concluya y explique si se rechaza o no se rechaza la hipótesis nula.

f) Realice una prueba χ^2 para conocer si existe evidencia estadística de uniformidad en los números pseudoaleatorios, calculando el estadístico de prueba y el *p-value*. De nuevo mencione si se rechaza o no se rechaza la hipótesis nula.

g) Realice un análisis del correlograma.

h) Aplique una prueba de Cox-Stuart a los valores u_i.

i) Aplique una prueba de rachas a los valores u_i.

Miscelánea

21. Sea X una variable aleatoria con función de densidad como aparece abajo. Calcule $E(X)$.

$$f(x) = \begin{cases} (3/7)\, e^{-x} + (8/7)\, e^{-2x} & \text{si } \ x > 0, \\ 0 & \text{si } \ x \leqslant 0. \end{cases}$$

22. *Aproximación del número* π. Con ayuda de un programa de cómputo, obtenga una muestra de 2000 números pseudoaleatorios con el algoritmo de su preferencia. Usando la muestra obtenida, construya 1000 parejas ordenadas

$$(x_1, y_1), \ldots, (x_{1000}, y_{1000}),$$

por ejemplo, emparejando los primeros 1000 valores obtenidos con los últimos 1000. Considerando el gráfico de los puntos (x_i, y_i) en el conjunto $[0, 1] \times [0, 1]$, defina a n como el número de casos donde la distancia de los puntos al origen es menor o igual a 1, es decir, cuando $x_i^2 + y_i^2 \leqslant 1$. Entonces, de acuerdo a la Proposición 1.4 el valor de $n/1000$ aproximará el valor de $\pi/4$ y, por lo tanto, el valor de $n/250$ aproximará el valor de π. Realice varias aproximaciones al número π. Después repita el ejercicio con 10000 parejas en vez de 1000 y observe qué tanto mejora la aproximación.

23. Suponga que usted tiene 3 manzanas, una de ellas está podrida y conoce cuál es.

a) Le pide a alguien que escoja una manzana y si elige la podrida registra el número 1. Si elige cualquier otra, registra el número 0. Repitiendo este procedimiento con 100 personas, al final tendrá 100 números registrados, cada uno tomando el valor cero o uno. ¿A qué número debe aproximarse el promedio de esos valores?

b) Repita el procedimiento, pero ahora con la siguiente modificación: después de que se realizó la primera elección, descarte esa manzana. De las 2 manzanas restantes retire aquella que usted sepa que no es la podrida. La manzana que sobra, puede estar podrida o puede estar saludable. Si está podrida registe el número 1 y si está saludable registre el número cero. Repitiendo este procedimiento con 100 personas al final tendrá 100 números registrados, cada uno tomando el valor cero o uno. ¿A qué número se aproxima el promedio de esos valores?

c) Investigue en qué consiste el problema de Monty Hall y analice el parecido con el experimento de las manzanas. ¿Qué probabilidad condicional aproximó en el inciso anterior?

24. Considere el generador multiplicativo dado por (1.21) con $M = 64$. Determine los valores de a que producen el máximo periodo posible.

25. Considere el generador $x_{n+1} = x_n^2 \mod M = 19261141$. Elaborando un programa de cómputo, calcule el periodo del generador que comienza con la semilla:

a) $x_0 = 196$.
b) $x_0 = 400$.
c) $x_0 = 200$.
d) $x_0 = 505$.

26. Realice en 20 ocasiones el procedimiento que aparece abajo para obtener números pseudoaleatorios $x_1, \ldots, x_{20}$. Realice un histograma y un diagrama de dispersión. Además realice una prueba de Cox-Stuart y una prueba de rachas para determinar la calidad de la muestra obtenida como números aleatorios en el intervalo $(0, 1)$.

a) Lance un dado y registre el valor obtenido.

b) Lance una moneda tantas veces como indicó el resultado del dado. En cada lanzamiento, si el resultado es cara registre un valor de 1, si el resultado es cruz registre un valor de 2.

c) Sume el total de los valores que registró en el inciso anterior. Al final obtendrá un valor entre 1 y 12, divida tal número entre 12 para obtener el número x_i.

Capítulo 2

Métodos para generar valores de variables aleatorias

En este capítulo estudiaremos algunos métodos para obtener valores de una variable aleatoria con una distribución de probabilidad dada. Consideraremos tanto el caso variables aleatorias univariadas como multivariadas, discreta y continuas.

2.1. Método de la función inversa

Este es un método para generar valores de una variable aleatoria y está basado en la función inversa generalizada de una función de distribución. Definiremos primero a esta función.

Definición 2.1 *La función inversa generalizada de una función de distribución $F(x)$ es*

$$F^{-1}(u) = \text{ínf}\,\{x : F(x) \geqslant u\}, \quad 0 < u < 1. \tag{2.1}$$

Es decir, $F^{-1}(u)$ representa el valor x más pequeño tal que su evaluación $F(x)$ alcanza o rebasa el nivel u. Por brevedad, nos referiremos a la función (2.1) simplemente como función inversa y, como se indica en la

definición anterior, se utiliza el mismo símbolo que se usa para las funciones inversas usuales. Además, para efectos de calcular su función inversa, nos interesará analizar una función de distribución $F(x)$ únicamente en aquellos valores x tales que $0 < F(x) < 1$.

En las Figuras 2.1, 2.3 y 2.2 se muestran varios ejemplos de funciones de distribución $F(x)$ y sus correspondientes funciones inversas $F^{-1}(u)$. Es conveniente recordar aquí que la expresión $f(x+)$ indica el límite por la derecha de la función f en el punto x. De manera análoga, la expresión $f(x-)$ indica el límite por la izquierda de la función f en el punto x. La afirmación de que la función f es continua por la derecha en el punto x se expresar mediante la igualdad $f(x) = f(x+)$, y la continuidad por la izquierda se escribe $f(x) = f(x-)$. Cuando la función es continua en x se cumplen las igualdades $f(x-) = f(x) = f(x+)$.

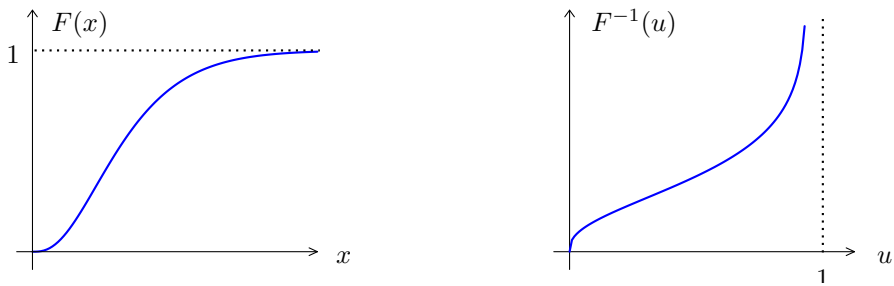


Figura 2.1: Un ejemplo de $F(x)$ y su función inversa $F^{-1}(u)$.

Ejemplo 2.1 (Cuando $F(x)$ tiene una discontinuidad)
En la parte izquierda de la Figura 2.2 se muestra una sección de una función de distribución $F(x)$ que presenta una discontinuidad en el punto $x = a$. Puede comprobarse que la función inversa se mantiene constante en el intervalo $(F(a-), F(a))$ en el valor $u = F(a)$.

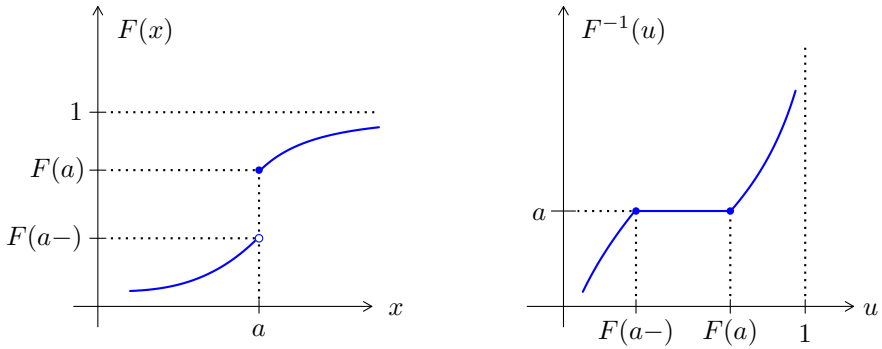


Figura 2.2: Un ejemplo de $F(x)$ y su función inversa $F^{-1}(u)$.

■

Ejemplo 2.2 (Cuando $F(x)$ es constante en intervalos)
En la parte izquierda de la Figura 2.3 muestra una sección de una función de distribución que no es invertible pues no es una función biunívoca: el intervalo completo (a, b) es llevado a la constante $F(a)$. En la parte derecha de la misma figura se muestra la función inversa $F^{-1}(u)$. Se ilustra el hecho de que cuando la función de distribución $F(x)$ es constante en un intervalo (a, b), la función inversa tiene una discontinuidad en el valor $u = F(a)$. Los límites laterales en este punto u son:

$$\begin{aligned} F^{-1}(F(a)-) &= a, \\ F^{-1}(F(a)+) &= b. \end{aligned}$$

El tamaño de la discontinuidad o salto es la longitud del intervalo en el que $F(x)$ es constante.

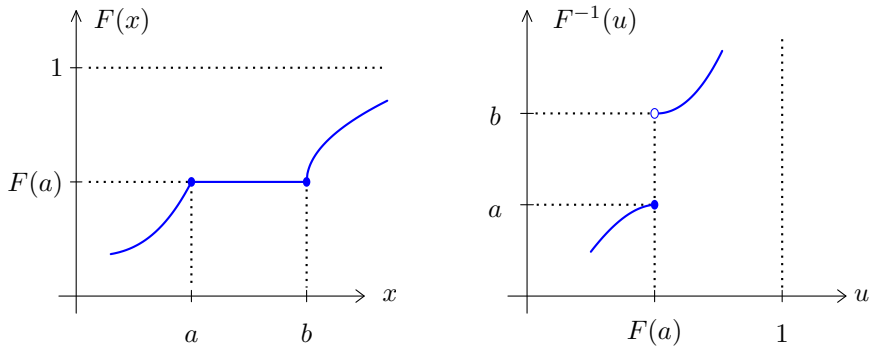


Figura 2.3: Un ejemplo de $F(x)$ y su función inversa $F^{-1}(u)$.

En el caso cuando $F(x)$ es estrictamente creciente, su función inversa existe y coincide con la función definida en (2.1). En tal caso se cumple que $F(F^{-1}(u)) = u$, para $0 < u < 1$, y se verifica también que $F^{-1}(F(x)) = x$, para $-\infty < x < \infty$.

En el caso general cuando $F(x)$ no necesariamente es invertible, se cumplen las propiedades que aparecen en la siguiente proposición. Dado que la demostración de estas propiedades es un tanto técnica, por el momento se recomienda al lector leer y entender sólo el enunciado y pasar después directamente a la Proposición 2.2, que fundamenta el método de la función inversa.

Proposición 2.1 *Sea $F^{-1}(u)$ la función inversa generalizada de una función de distribución $F(x)$. Entonces*

1. *$F^{-1}(u)$ es no decreciente.*
2. *$F^{-1}(u)$ es continua por la izquierda.*
3. *$F(F^{-1}(u)) \geqslant u$, para $0 < u < 1$.*
4. *$F^{-1}(F(x)) \leqslant x$, para x tal que $0 < F(x) < 1$.*

Demostración.

1. Sean u_1 y u_2 dos números tales que $0 < u_1 \leqslant u_2 < 1$. Entonces

$$\{x : F(x) \geqslant u_2\} \subseteq \{x : F(x) \geqslant u_1\}.$$

   Tomando el ínfimo se obtiene ínf $\{x : F(x) \geqslant u_1\} \leqslant$ ínf $\{x : F(x) \geqslant u_2\}$, pues el ínfimo del conjunto con mayor número de elementos puede ser más pequeño. Esto significa que $F^{-1}(u_1) \leqslant F^{-1}(u_2)$.

2. Sea $u \in (0, 1)$ y sea $\{u_n : n = 1, 2 \ldots\}$ una sucesión de números dentro del intervalo unitario $(0, 1)$ tal que $u_n \nearrow u$. Se cumple la descomposición

$$\{x : F(x) \geqslant u_n\} = \{x : F(x) \geqslant u\} \cup \{x : u_n \leqslant F(x) < u\},$$

   en donde los dos conjuntos del lado derecho son disjuntos. Tomando el ínfimo de estos conjuntos, tenemos que el ínfimo de la unión que aparece en el lado derecho de la igualdad se puede escribir como el mínimo de los dos ínfimos, es decir,

$$F^{-1}(u_n) = \text{mín}\,\{F^{-1}(u), \text{ínf}\,\{x : u_n \leqslant F(x) < u\}\}.$$

   Al tomar el límite cuando $n \to \infty$ y observar que el conjunto $\{x : u_n \leqslant F(x) < u\}$ converge decrecientemente al conjunto vacío y seguir la convención ínf $\varnothing = \infty$, tenemos que

$$\lim_{n\to\infty} F^{-1}(u_n) = \text{mín}\,\{F^{-1}(u), \infty\} = F^{-1}(u).$$

3. Sea u un número tal que $0 < u < 1$ y definamos

$$x^* := F^{-1}(u) = \text{ínf}\,\{x : F(x) \geqslant u\}.$$

Supongamos que $\{x_n : n = 1, 2, \ldots\}$ es una sucesión de números dentro del conjunto $\{x : F(x) \geqslant u\}$ tal que converge a x^* en forma decreciente, es decir, $x_n \searrow x^*$ cuando $n \to \infty$. Entonces

$$\begin{aligned} F(F^{-1}(u)) &= F(x^*) \\ &= F\left(\lim_{n\to\infty} x_n\right) \\ &= \lim_{n\to\infty} F(x_n) \\ &\geqslant \lim_{n\to\infty} u \\ &= u. \end{aligned}$$

Se ha usado el hecho de que toda función de distribución es continua por la derecha.

4. Todo número real x es un elemento del conjunto $\{u : F(u) \geqslant F(x)\}$. Por lo tanto,

$$F^{-1}(F(x)) = \text{ínf}\,\{u : F(u) \geqslant F(x)\} \leqslant x.$$

■

El siguiente resultado establece una manera de generar valores de una variable aleatoria a partir de la función inversa generalizada de su función de distribución.

Proposición 2.2 (Método de la función inversa) *Sea X una variable aleatoria con función de distribución $F(x)$, con inversa generalizada $F^{-1}(u)$. Sea $U \sim \text{unif}\,(0, 1)$. Entonces*

$$F^{-1}(U) \stackrel{d}{=} X. \tag{2.2}$$

Demostración. Primero supondremos válida la siguiente igualdad de eventos: para cualquier valor real de x,

$$(F^{-1}(U) \leqslant x) = (U \leqslant F(x)). \tag{2.3}$$

Aplicando la probabilidad a los eventos en (2.3), tenemos que la función de distribución de la variable $F^{-1}(U)$ es

$$P(F^{-1}(U) \leqslant x) = P(U \leqslant F(x)) = F(x).$$

Esta igualdad se cumple para cualquier valor de x y significa que las variables $F^{-1}(U)$ y X tienen la misma función de distribución. Sólo resta comprobar la identidad (2.3). Demostraremos que cada evento está contenido en el otro.

($\subseteq$) Sea $\omega \in (F^{-1}(U) \leqslant x)$. Entonces $F^{-1}(U(\omega)) \leqslant x$. Aplicando F, usando el inciso (3) de la Proposición 2.1 y la monotonía de F,

$$U(\omega) \leqslant F(F^{-1}(U(\omega))) \leqslant F(x).$$

Esto significa que $\omega \in (U \leqslant F(x))$.

($\supseteq$) Sea $\omega \in (U \leqslant F(x))$. Entonces $U(\omega) \leqslant F(x)$. Aplicando F^{-1}, usando la monotonía de F^{-1} y el inciso (4) de la Proposición 2.1,

$$F^{-1}(U(\omega)) \leqslant F^{-1}(F(x)) \leqslant x.$$

Esto significa que $\omega \in (F^{-1}(U) \leqslant x)$.

■

De esta manera, si se desea generar un valor de la variable aleatoria X con función de distribución $F(x)$, simplemente se genera un valor u de la distribución unif $(0, 1)$ y se evalúa $F^{-1}(u)$. Éste será un valor de X. Por supuesto, el método de la función inversa es aplicable cuando se cuente con una expresión para $F^{-1}(u)$, o bien una forma de calcularla. Veremos a continuación algunos ejemplos.

Ejemplo 2.3 *Sea X una variable aleatoria con función de densidad*

$$f(x) = \begin{cases} 2x & si \;\; 0 < x < 1, \\ 0 & en \; otro \; caso. \end{cases}$$

Puede comprobarse que

$$F(x) = \begin{cases} 0 & si \;\; x \leqslant 0, \\ x^2 & si \;\; 0 < x < 1, \\ 1 & si \;\; x \geqslant 1. \end{cases}$$

En este caso, $F(x)$ es invertible sobre el intervalo $(0,1)$ y su inversa es $F^{-1}(u) = \sqrt{u}$, para $0 < u < 1$. De este modo, si u es un valor de la distribución unif $(0,1)$*, entonces $F^{-1}(u) = \sqrt{u}$ es un valor de la variable aleatoria X con función de densidad $f(x)$.* ▪

Ejemplo 2.4 (Máximo) *Sea $X_1, \ldots, X_n$ una muestra aleatoria de la distribución $F(x)$. Suponga que deseamos obtener valores de la estadística (función de la muestra aleatoria):*

$$X_{(n)} := \text{máx}\,\{X_1, \ldots, X_n\}.$$

Un primer método consiste en generar n valores $x_1, \ldots, x_n$ de la distribución $F(x)$, tomar el máximo de estos valores y ése sería un valor de $X_{(n)}$. Sin embargo, se puede lograr una mayor eficiencia usando el método de la función inversa encontrando la función de distribución de $X_{(n)}$. Recordemos que

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leqslant x) \\ &= P(X_1 \leqslant x, \cdots, X_n \leqslant x) \\ &= [\, P(X_1 \leqslant x)\,]^n \\ &= [\, F(x)\,]^n . \end{aligned}$$

Por lo tanto, su función inversa es

$$F^{-1}_{X_{(n)}}(y) = F^{-1}(y^{1/n}), \quad 0 < y < 1.$$

De manera análoga se puede encontrar un procedimiento similar para generar valores de la variable aleatoria mínimo de una muestra aleatoria, es decir, de $X_{(1)} := \text{mín}\,\{X_1, \ldots, X_n\}$. Véase el Ejercicio 33.

▪

Ejemplo 2.5 *Este es otro ejemplo de una función de distribución mixta $F(x)$ para la cual no existe su función inversa usual, pero se puede encontrar una expresión para la función inversa generalizada y, por lo tanto, se puede aplicar el método de la función inversa.*

Sean $0 < x_1 < x_2 < x_3$ tres números y sean $0 < a < b < c < 1$ otros tres números. Considere la función de distribución $F(x)$ con soporte el intervalo $(0, x_3)$ dada por

$$F(x) = \begin{cases} \dfrac{a}{x_1}x & si \;\; 0 < x < x_1, \\ b & si \;\; x_1 \leqslant x < x_2, \\ 1 + \dfrac{1-c}{x_3 - x_2}(x - x_3) & si \;\; x_2 \leqslant x < x_3. \end{cases}$$

Esta función no es invertible en el sentido usual pues lleva el intervalo $[x_1, x_2)$ a la constante b. Véase una gráfica de esta función en la parte izquierda de la Figura 2.4. Puede comprobarse que su función inversa generalizada es

$$F^{-1}(u) = \begin{cases} \dfrac{x_1}{a}u & si \;\; 0 < u \leqslant a, \\ x_1 & si \;\; a < u \leqslant b, \\ x_2 & si \;\; b < u \leqslant c, \\ x_3 + \dfrac{x_3 - x_2}{1-c}(u-1) & si \;\; c < u < 1. \end{cases}$$

La gráfica de la función $F^{-1}(u)$ se muestra en la parte derecha de la Figura 2.4.

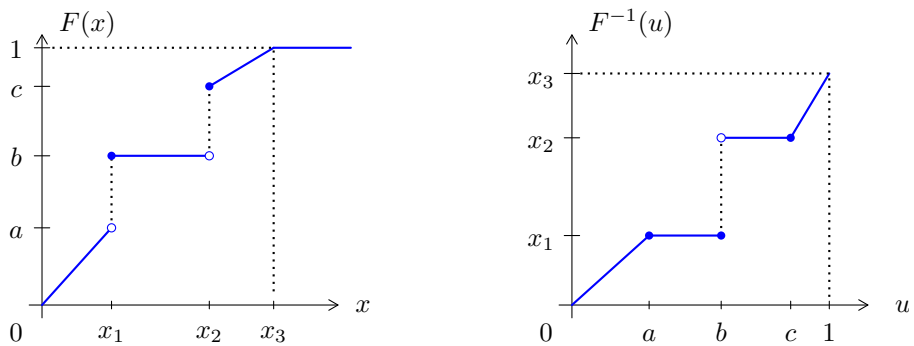


Figura 2.4: Gráficas del Ejemplo 2.5.

En conclusión, si u es un valor de la distribución unif $(0,1)$*, entonces* $F^{-1}(u)$ *es un valor de la distribución* $F(x)$ *arriba indicada.*

■

En la siguiente sección aplicaremos el método de la función inversa al caso particular de variables aleatorias discretas.

2.2. Generación de valores de variables aleatorias discretas

.

En esta breve sección se muestra la forma de generar valores de una variable aleatoria discreta aplicando el método de la función inversa.

Sea X una variable aleatoria discreta con valores $x_1 < x_2 < \cdots$ y probabilidades estrictamente positivas $p_1, p_2, \ldots$, respectivamente. La función de distribución de X es

$$F(x) = \begin{cases} 0 & \text{si } x < x_1, \\ p_1 & \text{si } x_1 \leqslant x < x_2, \\ p_1 + p_2 & \text{si } x_2 \leqslant x < x_3, \\ p_1 + p_2 + p_3 & \text{si } x_3 \leqslant x < x_4, \\ \vdots & \vdots \end{cases}$$

La función inversa es, para $0 < u < 1$,

$$F^{-1}(u) = \begin{cases} x_1 & \text{si } 0 < u < p_1, \\ x_2 & \text{si } p_1 \leqslant u < p_1 + p_2, \\ x_3 & \text{si } p_1 + p_2 \leqslant u < p_1 + p_2 + p_3, \\ \vdots & \vdots \end{cases}$$

De esta manera, si u es un valor de la distribución unif $(0,1)$, entonces $F^{-1}(u)$ es un valor de la variable aleatoria discreta X con la distribución dada. Este procedimiento puede implementarse en una computadora siguiendo el diagrama de flujo que aparece en la Figura 2.5.

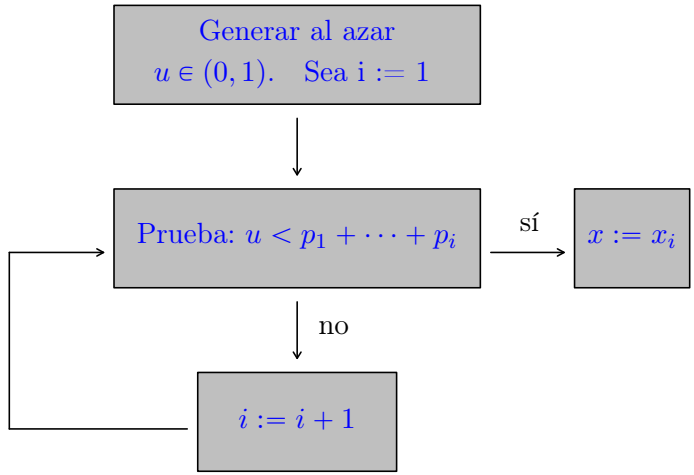


Figura 2.5: Diagrama de flujo.

Ejemplo 2.6 ***(Distribución Bernoulli)*** *Sea* $p \in (0,1)$ *y sea* u *un valor de la distribución* unif $(0,1)$. *Puede comprobarse que el número* x *definido abajo es un valor de la distribución* Ber (p).

$$x = 1_{(0,p]}(u) = \begin{cases} 1 & si \quad u \leqslant p, \\ 0 & si \quad u > p. \end{cases}$$

■

Sobre la eficiencia del algoritmo

Uno puede preguntarse por el número promedio de veces que se lleva a cabo la pregunta ¿$u \leqslant p_1 + \cdots + p_i$? en el diagrama de flujo de la Figura 2.5. Para variables aleatorias con un número infinito de valores, la respuesta es

$$\sum_{i=1}^{\infty} i\, p_i.$$

Este valor se obtiene debido a que se hace 1 pregunta cuando $u \in (0, p_1]$, se hacen 2 preguntas cuando $u \in (p_1, p_1 + p_2]$, se hacen 3 preguntas cuando $u \in (p_1 + p_2, p_1 + p_2 + p_3]$, etcétera.

Es claro que el algoritmo es más eficiente cuando las probabilidades están ordenadas de mayor a menor: si escribimos la distribución como la sucesión de parejas $(x_1, p_1), (x_2, p_2), \ldots$, entonces debemos considerar el orden $(y_1, p_{(1)}), (y_2, p_{(2)}), \ldots$, en donde $p_{(i)}$ es la i-ésima probabilidad mayor, es decir, $p_{(1)} \geqslant p_{(2)} \geqslant \cdots$ y y_i representa el valor asociado a la probabilidad $p_{(i)}$. Se garantiza así que el número promedio de preguntas en el algoritmo es mínimo y está dado por el lado izquierdo de la siguiente desigualdad

$$\sum_{i=1}^{\infty} i\, p_{(i)} \leqslant \sum_{i=1}^{\infty} i\, p_i.$$

Se puede definir la eficiencia del algoritmo como el cociente

$$\sum_{i=1}^{\infty} i\, p_{(i)} / \sum_{i=1}^{\infty} i\, p_i,$$

suponiendo que el denominador es finito. Así, la eficiencia depende del orden en el que se consideren las probabilidades $p_1, p_2, \ldots$ y es un número en el intervalo $(0, 1]$, en donde el valor 1 significa la eficiencia máxima. Véase el Ejercicio 40 para el caso de variables aleatorias con un número finito de valores.

§

El método de la función inversa no es el único método por el que pueden generarse valores de una variable aleatoria discreta. En las Secciones 2.3 y 2.4 veremos algunos métodos alternativos.

2.3. Método de von Neumann

Este es un procedimiento para generar valores de una variable aleatoria continua o discreta, cuya función de densidad o de probabilidad $f(x)$ satisface ciertas condiciones. En el caso continuo, se requiere que $f(x)$

sea acotada y que su soporte también sea acotado. En el caso discreto, la función $f(x)$ está acotada por el valor 1 pues se trata de una probabilidad, pero se pide que su soporte sea un conjunto con un número finito de valores. Expondremos el procedimiento para cada una de estas dos situaciones. El método puede aplicarse también al caso de vectores aleatorios discretos o continuos. Este caso multidimensional se expone en la Sección 2.8.

El método de von Neumann es un caso particular de un método más general que estudiaremos en la siguiente sección llamado método de aceptación y rechazo.

Caso continuo

Sea $f(x)$ una función de densidad con soporte el intervalo completo (a, b), en donde $-\infty < a < b < \infty$, y tal que para $x \in (a, b)$, se cumple que $0 \leqslant f(x) \leqslant M$ con $M > 0$ constante. Esta es nuestra función objetivo. Nos interesa generar valores de esta función de densidad. Observemos que

$$M(b-a) \geqslant \int_a^b f(x)\,dx = 1.$$

Sean U_1 y U_2 dos variables aleatorias independientes, ambas con distribución unif $(0, 1)$. Definamos las variables aleatorias independientes

$$\begin{aligned} X &:= a + (b-a)\,U_1, \\ U &:= M\,U_2. \end{aligned}$$

Es inmediato comprobar que $X \sim$ unif (a, b) y que $U \sim$ unif $(0, M)$.

El método de von Neumann está basado en el siguiente resultado. En él aparece la expresión $f(X)$, que es la composición de funciones $f \circ X$. Observe el uso de minúsculas y mayúsculas: $f(x)$ es una función de densidad, mientras que $f(X)$ es una variable aleatoria.

Proposición 2.3 (Método de von Neumann) *Sea $f(x)$ una función de densidad con soporte el intervalo (a, b), en donde $-\infty < a < b < \infty$, y tal que $0 \leqslant f(x) \leqslant M$, con $M > 0$ constante. Sean U_1 y U_2 dos variables aleatorias independientes con distribución* unif $(0, 1)$ *y defina*

$$\begin{aligned} X &:= a + (b-a)\,U_1, \\ U &:= M\,U_2. \end{aligned}$$

Entonces la distribución de X condicionada al evento $(U \leqslant f(X))$ tiene función de densidad $f(x)$.

Demostración. Para cualquier valor $x \in (a, b)$,

$$\begin{aligned} P(X \leqslant x \,|\, U \leqslant f(X)) &= \frac{P(X \leqslant x, U \leqslant f(X))}{P(U \leqslant f(X))} \\ &= \frac{\displaystyle\int_a^x \int_0^{f(v)} f_{X,U}(v,u)\,du\,dv}{\displaystyle\int_a^b \int_0^{f(v)} f_{X,U}(v,u)\,du\,dv} \\ &= \frac{\displaystyle\int_a^x \int_0^{f(v)} \frac{1}{b-a}\,\frac{1}{M}\,du\,dv}{\displaystyle\int_a^b \int_0^{f(v)} \frac{1}{b-a}\,\frac{1}{M}\,du\,dv} \\ &= \frac{\displaystyle\int_a^x f(v)\,dv}{\displaystyle\int_a^b f(v)\,dv} \\ &= \int_a^x f(v)\,dv. \end{aligned}$$

■

Observemos que, inicialmente, X tiene distribución unif (a, b), sin embargo, su distribución condicionada al evento $(U \leqslant f(X))$ tiene función de densidad la función objetivo $f(x)$.

En resumen, el método de von Neumann es el siguiente: si u_1 y u_2 son dos valores de la distribución unif $(0, 1)$, escogidos de manera independiente, y si se definen los números

$$\begin{aligned} x &:= a + (b - a)\, u_1, \\ u &:= M\, u_2, \end{aligned}$$

entonces, cuando el par (x, u) es tal que se cumple la condición $u \leqslant f(x)$, entonces x es un valor de la distribución con función de densidad $f(x)$. Véase la Figura 2.6. Si el punto (x, u) que queda por abajo de la curva, el valor x es aceptado. En cambio, si (x, u) queda por arriba de la curva, el valor x no es aceptado.

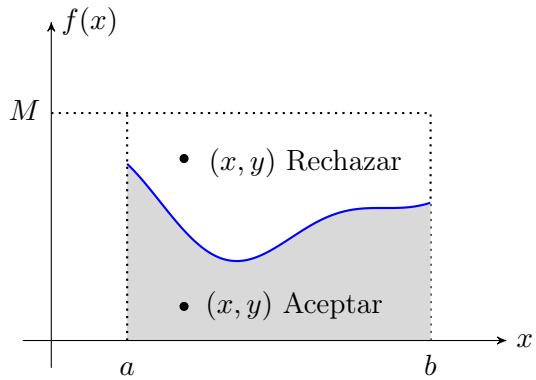


Figura 2.6: Ejemplo del método de von Neumann.

Mientras más pequeño sea el valor de la constante M, más frecuentemente se cumplirá la condición $u \leqslant f(x)$, y esto llevará a que x sea un valor aceptado y a que el método sea más eficiente. Esto produce la siguiente definición.

Definición 2.2 *Bajo el contexto del método de von Neumann que aparece en la Proposición 2.3, la eficiencia se define como la probabilidad del evento condicionante, es decir,* $P(U \leqslant f(X))$.

Esto es, la eficiencia es la probabilidad de que se cumpla la condición para

aceptar un valor x en un ensayo del método de von Neumann. Esta probabilidad es

$$\begin{aligned} P(U \leqslant f(X)) &= \int_a^b \int_0^{f(x)} f_{X,U}(x,u)\, du\, dx \\ &= \int_a^b \int_0^{f(x)} f_X(x)\, f_U(u)\, du\, dx \\ &= \int_a^b \int_0^{f(x)} \frac{1}{b-a}\, \frac{1}{M}\, du\, dx \\ &= \frac{1}{b-a}\, \frac{1}{M} \int_a^b f(x)\, dx \\ &= \frac{1}{b-a}\, \frac{1}{M}, \end{aligned}$$

en donde, como observamos antes, $(b-a)\, M \geqslant 1$.

Ejemplo 2.7 (Distribución beta) *Consideremos el caso de la distribución* $\text{beta}\,(a,b)$*, es decir, su función de densidad es*

$$f(x) = \begin{cases} \dfrac{1}{\mathrm{B}(a,b)}\, x^{a-1}\,(1-x)^{b-1} & \textit{si} \;\; 0 < x < 1, \\ 0 & \textit{en otro caso,} \end{cases}$$

en donde $a > 0$ *y* $b > 0$ *son dos parámetros. El soporte de esta distribución es el intervalo unitario* $(0,1)$ *y debe observarse que* a *y* b *son los parámetros de la distribución y no el intervalo* (a,b) *como se usa en el enunciado del método de von Neumann. Puede comprobarse que esta distribución tiene una única moda con valor*

$$\text{Moda} = x^* = \frac{a-1}{a+b-2},$$

cuando $a > 1$ *y* $b > 1$*. Bajo estas restricciones de los parámetros, el valor máximo de la función es* $f(x^*) = (x^*)^{a-1}\,(1-x^*)^{b-1}/\mathrm{B}(a,b)$*, el cual puede tomarse como la constante* M*. Siendo el soporte de la distribución acotado, se puede aplicar el método de von Neumann para generar valores de esta distribución. De esta manera, el mecanismo de simulación es el siguiente: sean* u_1 *y* u_2 *dos valores generados de manera independiente de la distribución* $\text{unif}(0,1)$*. Se definen* $x := u_1$ *y* $u := M u_2$*. Cuando ocurre que* $u \leqslant f(x)$*, se acepta a* x *como un valor de la distribución* $\text{beta}\,(a,b)$*.* ∎

El método de von Neumann es un procedimiento en el que es evidente que es de utilidad conocer la moda de una distribución. Tal moda, o un valor superior, puede tomarse como el valor de la cota M en el método de von Neumann.

Caso discreto

A continuación veremos el método de von Neumann en el caso de variables aleatorias discretas con un número finito de valores. El enunciado del resultado y su demostración son análogos al caso continuo.

Proposición 2.4 *Sea $f(x)$ una función de probabilidad con soporte el conjunto finito $\{x_1, \ldots, x_n\}$. Sean $X \sim \text{unif}\,\{x_1, \ldots, x_n\}$ y $U \sim \text{unif}\,(0,1)$. Entonces*

$$P(X = x \,|\, U \leqslant f(X)\,) = f(x), \quad \textit{para} \quad x = x_1, \ldots, x_n.$$

Demostración. Sea x cualquier elemento del conjunto $\{x_1, \ldots, x_n\}$. Observemos que $P(X = x) = 1/n$ y $P(U \leqslant f(X) \,|\, X = x) = f(x)$. Entonces

$$\begin{aligned} P(X = x \,|\, U \leqslant f(X)\,) &= \frac{P(X = x, U \leqslant f(X))}{P(U \leqslant f(X))} \\ &= \frac{P(U \leqslant f(X) \,|\, X = x) P(X = x)}{\sum_{i=1}^{n} P(U \leqslant f(X) \,|\, X = x_i)\, P(X = x_i)} \\ &= \frac{f(x)\,(1/n)}{\sum_{i=1}^{n} f(x_i)\,(1/n)} \\ &= f(x). \end{aligned}$$

∎

Observemos nuevamente que $f(x)$ es nuestra función objetivo. Además, la distribución inicial de X es uniforme, pero su distribución condicional es la función objetivo $f(x)$. Es decir, a partir de las distribuciones

unif $\{x_1, \ldots, x_n\}$ y unif $(0, 1)$, podemos generar valores de cualquier distribución sobre el conjunto $\{x_1, \ldots, x_n\}$.

Ejemplo 2.8 *Supongamos que se desean obtener valores de la distribución* bin (n, p)*, es decir, la función objetivo es*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

Sea x un valor de la distribución unif $\{0, 1, , \ldots, n\}$ *y sea u un valor de la distribución* unif $(0, 1)$*. Cuando ocurra la desigualdad que aparece abajo se acepta a x como un valor de la distribución binomial, en caso contrario se rechaza.*

$$u \leqslant \binom{n}{x} p^x (1-p)^{n-x}.$$

■

El método de von Newmann puede aplicarse también al caso de vectores aleatorios discretos o continuos. Se estudian estos procedimientos en la Sección 2.8.

2.4. Método de aceptación y rechazo

Este método es una generalización del método de von Neumann estudiado en la sección anterior. Fue creado por Stan Ulam y John von Neumann. Presentaremos primero el caso continuo y después veremos el caso discreto. El método puede aplicarse también al caso de vectores aleatorios discretos y continuos. El caso multidimensional se expone en la Sección 2.8.

Caso continuo

Supongamos que deseamos obtener valores de acuerdo a una función de densidad $f(x)$. Esta es nuestra función objetivo. Supongamos también que $f(x)$ tiene soporte el intervalo (a, b), en donde $-\infty \leqslant a < b \leqslant \infty$. Este intervalo no necesariamente es acotado. Por otro lado, sea X una variable aleatoria con función de densidad $g(x)$ tal que:

- El soporte de $g(x)$ es también el intervalo (a, b).

- Se conoce una forma de generar valores de la distribución con función de densidad $g(x)$.
- Existe una constante $c \geqslant 1$ tal que $c \cdot g(x) \geqslant f(x)$, para $x \in (a, b)$.

A la función $g(x)$ se le llama función propuesta (*proposal function*). Definamos también la variable aleatoria

$$U \sim \text{unif}\,(0, c \cdot g(X)).$$

Observemos que el lado derecho del intervalo de esta distribución es aleatorio y que, claramente, X y U no son independientes. El método de aceptación y rechazo está basado en el siguiente resultado.

Proposición 2.5 (Método de aceptación y rechazo) *Sea $f(x)$ una función con soporte el intervalo (a, b), en donde $-\infty \leqslant a < b \leqslant \infty$. Sea X una variable aleatoria con función de densidad $g(x)$ tal que el soporte de $g(x)$ es también (a, b). Suponga que se conoce una forma de generar valores de la distribución $g(x)$ y que existe una constante $c \geqslant 1$ tal que $c \cdot g(x) \geqslant f(x)$, para $x \in (a, b)$. Entonces la variable aleatoria X condicionada al evento $(U \leqslant f(X))$ tiene función de densidad $f(x)$, es decir,*

$$X \,|\, (U \leqslant f(X)) \sim f(x).$$

Demostración. Para $x \in (a, b)$,

$$\begin{aligned}
P(X \leqslant x \,|\, U \leqslant f(X)) &= \frac{P(X \leqslant x, U \leqslant f(X))}{P(U \leqslant f(X))} \\
&= \frac{\displaystyle\int_a^x \int_0^{f(v)} f_{X,U}(u,v)\, du\, dv}{\displaystyle\int_a^b \int_0^{f(v)} f_{X,U}(u,v)\, du\, dv} \\
&= \frac{\displaystyle\int_a^x \int_0^{f(v)} f_{U\,|\,X}(u\,|\,v)\, g(v)\, du\, dv}{\displaystyle\int_a^b \int_0^{f(v)} f_{U\,|\,X}(u\,|\,v)\, g(v)\, du\, dv} \\
&= \frac{\displaystyle\int_a^x \int_0^{f(v)} \frac{1}{c\, g(v)}\, g(v)\, du\, dv}{\displaystyle\int_a^b \int_0^{f(v)} \frac{1}{c\, g(v)}\, g(v)\, du\, dv} \\
&= \frac{\displaystyle\int_a^x f(v)\, dv}{\displaystyle\int_a^b f(v)\, dv} \\
&= \int_a^x f(v)\, dv.
\end{aligned}$$

■

Observemos que, inicialmente, X tiene función de densidad $g(x)$, pero su distribución condicionada al evento $(U \leqslant f(X))$ tiene la función de densidad objetivo $f(x)$. Por otro lado, debe observarse que no es necesario que los soportes de las funciones de densidad $f(x)$ y $g(x)$ coincidan. En realidad pueden utilizarse funciones de densidad $g(x)$ cuyo soporte contiene al soporte de la función objetivo $f(x)$, pues si x_0 es un valor generado según la función de densidad $g(x)$ pero que no pertenece al soporte de $f(x)$, entonces tenemos que $f(x_0) = 0$ y la condición de aceptación $u \leqslant f(x)$ se convierte en $u \leqslant 0$, lo cual nunca se cumple, de modo que el valor x_0 se rechazará siempre. Considerar una función de densidad $g(x)$ con soporte mayor al de $f(x)$ puede no ser eficiente pero puede ayudar en la facilidad

o comodidad de generar valores de una función de densidad $g(x)$ de esas características si ya se cuenta con un mecanismo automatizado o sencillo para ello. Por ejemplo, hemos visto que se pueden obtener valores de la función de densidad exponencial con facilidad, usando esos valores y el método de aceptación y rechazo se pueden generar valores de cualquier función de densidad acotada con soporte acotado contenido en el intervalo $(0, \infty)$.

En resumen, el método de aceptación y rechazo es el siguiente.

Método de aceptación y rechazo

Sea $f(x)$ una función de densidad con soporte (a, b), en donde $-\infty \leqslant a < b \leqslant \infty$. Sea X una variable aleatoria con función de densidad $g(x)$ y con el mismo soporte (a, b).

- Se genera un valor x de la variable X.
- Se genera un valor u de la distribución unif $(0, c\, g(x))$.

Cuando (x, u) es tal que $u \leqslant f(x)$, el número x se acepta como un valor de la distribución $f(x)$. En caso contrario, se rechaza.

Mientras más pequeño sea el valor de c, más frecuentemente se cumplirá la condición de aceptación $u \leqslant f(x)$, y eso llevará a que se acepte x y a que el método sea más eficiente. Esto produce la siguiente definición.

Definición 2.3 *Bajo el contexto del método de aceptación y rechazo que aparece en la Proposición 2.5, la eficiencia se define como la probabilidad del evento condicionante, es decir, como* $P(U \leqslant F(X))$.

Esta probabilidad es

$$\begin{aligned} P(U \leqslant F(X)) &= \int_a^b \int_0^{f(x)} f_{X,U}(u,v)\, du\, dx \\ &= \int_a^b \int_0^{f(x)} f_{U\,|\,X}(u\,|\,x)\, g(x)\, du\, dx \\ &= \int_a^b \int_0^{f(x)} \frac{1}{c\, g(x)}\, g(x)\, du\, dx \\ &= \frac{1}{c} \int_a^b f(x)\, dx \\ &= \frac{1}{c}, \quad \text{en donde} \;\; c \geqslant 1. \end{aligned}$$

A continuación haremos algunas observaciones generales del método de aceptación y rechazo.

1. Sea N la variable aleatoria que cuenta el número de ensayos, i.e. generaciones de valores del vector (X, U), hasta obtener un éxito en el método de aceptación y rechazo. Entonces $N \sim \text{geo}\,(p)$, con $p = 1/c$, es decir,

   $$P(N = n) = (1 - 1/c)^{n-1}\,(1/c), \quad \text{para} \;\; n = 1, 2, \ldots$$

   Esta es la distribución geométrica que inicia en el valor 1. Su valor promedio es $E(N) = 1/(1/c) = c$.

2. Cuando la función $g(x)$ y su dominio (a, b) son acotados, se puede tomar $M = \sup\{c\, g(x) : a < x < b\}$ para recuperar el método de von Neumann.

3. Otra manera equivalente de expresar la condición de aceptación es la siguiente: la variable $U \sim \text{unif}\,(0, c\, g(X))$ se puede expresar también como $c\, g(X)\, V$, en donde $V \sim \text{unif}\,(0, 1)$. Por lo tanto, la condición de aceptación es

   $$\begin{aligned} (U \leqslant f(X)) &= (c\, g(X)\, V \leqslant f(X)) \\ &= \left(V \leqslant \frac{f(X)}{c\, g(X)}\right). \end{aligned}$$

De esta manera, el resultado teórico que da origen al método de aceptación y rechazo se puede escribir como sigue: "La distribución condicional de X dado el evento $(V \leqslant f(X)/(c\,g(X)))$ tiene función de densidad $f(x)$, en donde $V \sim \text{unif}\,(0,1)$".

4. La dificultad del método consiste en encontrar una función de densidad $g(x)$ para la cual se conozca una manera de generar valores y que dicha función $g(x)$ cumpla la condición $c\,g(x) \geqslant f(x)$.

Ejemplo 2.9 *Supongamos que deseamos usar el método de aceptación y rechazo para producir valores de la distribución con función de densidad*

$$f(x) = 2x, \quad \textit{para} \;\; 0 < x < 1.$$

Podemos tomar la variable aleatoria X con distribución $\text{unif}\,(0,1)$ *y $c = 2$. Es decir, la función de densidad de X es $g(x) = 1$ para $0 < x < 1$, y se cumple el requisito $c\,g(x) \geqslant f(x)$ para $0 < x < 1$. Así, se genera un valor x de la variable aleatoria X y después se genera un valor u de la distribución* $\text{unif}\,(0,2)$*. Cuando ocurre que $u \leqslant f(x)$, se acepta el valor x. La eficiencia es $1/c = 1/2$, es decir, en promedio, uno de cada dos puntos generados (x,u) será útil. Observemos que para este ejemplo el método de la función inversa podría ser más sencillo, sin embargo nuestro objetivo aquí es ilustrar el método de aceptación y rechazo.* ∎

Ejemplo 2.10 *Consideremos la función de densidad*

$$f(x) = 8x, \quad \textit{para} \;\; 0 < x < 1/2.$$

Deseamos obtener valores de esta distribución usando el método de aceptación y rechazo. Veremos algunas alternativas para la función propuesta.

- *Como primer ejemplo podemos tomar $g(x) = 2$, para $0 < x < 1/2$. Esta es la distribución* $\text{unif}\,(0,1/2)$*. Puede comprobarse que se satisface la condición $c\,g(x) \geqslant f(x)$, para cualquier constante $c \geqslant 2$. El valor $c = 2$ es el más conveniente.*

- *Si la gráfica de la función propuesta $g(x)$ es similar a la gráfica de la función objetivo $f(x)$, entonces el método se vuelve más eficiente en el sentido de disminuir el número de rechazos. Supongamos que se elige como función propuesta a*

$$g(x) = 3\sqrt{2x}, \quad \textit{para} \;\; 0 < x < 1/2.$$

Esta función de densidad tiene una gráfica más cercana la función objetivo $f(x)$. Puede comprobarse que la desigualdad $c\,g(x) \geqslant f(x)$ se cumple con $c = 4/3$. La función de distribución asociada a $g(x)$ sobre su soporte es

$$G(x) = (2x)^{3/2}, \quad \textit{para} \;\; 0 < x < 1/2,$$

con inversa

$$G^{-1}(u) = \frac{1}{2}\,u^{2/3}, \quad \textit{para} \;\; 0 < u < 1. \tag{2.4}$$

Esto nos permite obtener valores de la distribución con función de densidad $g(x)$ usando el método de la función inversa. Por supuesto, para este caso sencillo, este último método puede ser aplicado directamente a la función objetivo. En efecto, la función de distribución asociada a $f(x)$ es $F(x) = 4x^2$, para $0 < x < 1/2$, con inversa $F^{-1}(u) = \frac{1}{2}\sqrt{u}$, para $0 < u < 1$. Sin embargo, nuestro ejemplo muestra la manera en la que dos métodos pueden ser utilizados en un mismo esquema de simulación.

- *Otra alternativa para generar valores de la distribución con función de densidad $f(x)$ se obtiene al observar que esta función es muy parecida a la función de densidad* $\text{beta}\,(2,1)$ *dada por*

$$g(x) = 2x, \quad \textit{para} \;\; 0 < x < 1.$$

Sin embargo, esta distribución tiene como soporte el intervalo $(0,1)$. Haciendo la transformación $Y = X/2$ para $X \sim \text{beta}\,(2,1)$, se puede comprobar que Y tiene función de densidad $f(x)$. De esta manera, valores de una distribución conocida, como lo es $\text{beta}\,(a,b)$*, pueden ser usados para generar valores de la función objetivo $f(x)$.*

■

Ejemplo 2.11 (Distribución normal) *Explicaremos una forma de producir valores de la distribución* $\mathrm{N}(\mu, \sigma^2)$ *usando el método de aceptación y rechazo. Seguiremos los siguientes pasos:*

- *Consideraremos la variable aleatoria* $Z \sim \mathrm{N}(0,1)$. *Puede comprobarse que la función de densidad de la variable* $|Z|$ *es la función* $f(x)$ *que aparece abajo. Esta será la función de densidad objetivo en el método de aceptación y rechazo.*

$$f(x) = \sqrt{\frac{2}{\pi}}\, e^{-x^2/2}, \quad \textit{para} \;\; x > 0.$$

- *Utilizaremos una variable aleatoria auxiliar* S *con distribución* $\mathrm{unif}\,\{+1, -1\}$. *La letra* S *proviene de la palabra "signo". Puede comprobarse que* $S \cdot |Z| \sim \mathrm{N}(0,1)$.

$$S = \begin{cases} +1 & \textit{con prob. } 1/2, \\ -1 & \textit{con prob. } 1/2. \end{cases}$$

- *Finalmente, tendremos que* $\mu + (S \cdot |Z|)\,\sigma \sim \mathrm{N}(\mu, \sigma^2)$.

En resumen, si z *es un valor de la variable* $|Z|$ *y* s *es un signo, entonces* $\mu + s\sigma z$ *es un valor de la distribución* $\mathrm{N}(\mu, \sigma^2)$.

Veamos entonces una forma de generar valores de la distribución con la función de densidad objetivo $f(x)$. *Consideraremos una variable aleatoria* X *con distribución* $\exp(\lambda)$ *con* $\lambda = 1$, *es decir, su función de densidad es*

$$g(x) = e^{-x}, \quad \textit{para} \;\; x > 0.$$

Tomando $c = \sqrt{2\,e/\pi}$ *se puede comprobar que* $c\,g(x) \geqslant f(x)$. *En efecto, para cualquier* $x > 0$,

$$\begin{aligned} c\,g(x) &= \sqrt{\frac{2\,e}{\pi}}\, e^{-x} \\ &= \sqrt{\frac{2}{\pi}}\, e^{-x+1/2} \\ &\geqslant \sqrt{\frac{2}{\pi}}\, e^{-x^2/2} \\ &= f(x). \end{aligned} \tag{2.5}$$

La desigualdad (2.5) *se obtiene a partir de* $-(x-1)^2/2 \leqslant 0$. *Al desarrollar este cuadrado se obtiene que* $-x^2/2 \leqslant -x+1/2$. *Tomando ahora exponencial se llega a* (2.5). *La condición de aceptación es*

$$V \leqslant \frac{f(X)}{c\, g(X)},$$

en donde $V \sim \text{unif}\,(0,1)$ *y* $g(x)$ *es la función de densidad de* $X \sim \exp(\lambda)$ *con* $\lambda = 1$. *A continuación simplificaremos la escritura de esta condición. Tenemos las siguientes igualdades de eventos,*

$$\begin{aligned}
\left(V \leqslant \frac{f(X)}{c\, g(X)}\right) &= \left(V \leqslant \frac{\sqrt{2/\pi}\, e^{-X^2/2}}{\sqrt{2\, e/\pi}\, e^{-X}}\right) \\
&= \left(V \leqslant \exp\left\{-X^2/2 + X - 1/2\right\}\right) \\
&= \left(V \leqslant \exp\left\{-(X-1)^2/2\right\}\right) \\
&= \left(-\ln V \geqslant (X-1)^2/2\right).
\end{aligned}$$

Es fácil comprobar que $-\ln V \sim \exp(\lambda)$ *con* $\lambda = 1$. *Por lo tanto, la condición de aceptación se puede escribir como*

$$(V_1 \geqslant (V_2 - 1)^2/2),$$

en donde V_1 *y* V_2 *son variables aleatorias independientes con idéntica distribución* $\exp(\lambda)$ *con* $\lambda = 1$. *Por lo demostrado antes, la eficiencia es* $1/c = (2\, e/\pi)^{-1/2} \leqslant 1$, *y de acuerdo a la versión simplificada de la condición de aceptación, este número debe coincidir con* $P(V_1 \geqslant (V_2-1)^2/2)$, *con* V_1 *y* V_2 *independientes y con distribución* $\exp(\lambda)$ *con* $\lambda = 1$. ▪

Caso discreto

Veremos ahora el método de aceptación y rechazo aplicado al caso de variables aleatorias discretas. No supondremos que las variables aleatorias toman un número finito de valores como en el método de von Neumann.

Supongamos que deseamos generar valores de una distribución con una función de probabilidad $f(x)$ con soporte el conjunto $\{x_1, x_2, \ldots\}$. Por otro lado, sea X una variable aleatoria discreta con función de probabilidad $g(x)$ tal que:

- El soporte de $g(x)$ es también el conjunto $\{x_1, x_2, \ldots\}$.
- Se conoce una forma de generar valores de la distribución $g(x)$.
- Existe una constante $c \geqslant 1$ tal que $c \cdot g(x) \geqslant f(x)$, para $x \in \{x_1, x_2, \ldots\}$.

Como antes, a la función $f(x)$ se le llama función objetivo, mientras que a $g(x)$ se le llama función propuesta (*proposal function*). Dado que $f(x)$ es una función de probabilidad, está acotada por el valor 1. Definamos también la variable aleatoria

$$U \sim \text{unif}\,(0, c\, g(X)).$$

El método de aceptación y rechazo para variables aleatorias discretas está basado en el siguiente resultado.

Proposición 2.6 *Sea $f(x)$ una función de probabilidad con soporte $\{x_1, x_2, \ldots\}$. Sea X una variable aleatoria discreta con función de probabilidad $g(x)$, cuyo soporte es también el conjunto $\{x_1, x_2, \ldots\}$, es tal que existe una constante $c \geqslant 1$ tal que $c\, g(x) \geqslant f(x)$ para $x = x_1, x_2, \ldots$ y se conoce una manera de generar valores de X. Sea $U \sim \text{unif}\,(0, c\, g(X))$. Entonces*

$$P(X = x \,|\, U \leqslant f(X)) = f(x), \quad \textit{para} \;\; x = x_1, x_2, \ldots$$

Demostración. Sea x cualquier elemento del conjunto $\{x_1, \ldots, x_n\}$. Observemos primero que

$$P(X = x) = g(x) > 0,$$
$$P(U \leqslant f(X) \,|\, X = x) = \frac{f(x)}{c\, g(x)}.$$

Entonces

$$\begin{aligned} P(X = x \,|\, U \leqslant f(X)) &= \frac{P(X = x, U \leqslant f(X))}{P(U \leqslant f(X))} \\ &= \frac{P(U \leqslant f(X) \,|\, X = x)P(X = x)}{\sum_{i=1}^{\infty} P(U \leqslant f(X) \,|\, X = x_i)P(X = x_i)} \\ &= \frac{\frac{f(x)}{c\,g(x)}\, g(x)}{\sum_{i=1}^{\infty} \frac{f(x_i)}{c\,g(x_i)}\, g(x_i)} \\ &= f(x). \end{aligned}$$

∎

Nuevamente, observemos que, inicialmente, X tiene función de probabilidad $g(x)$, pero su distribución condicionada al evento $(U \leqslant f(X))$ tiene función de probabilidad la función objetivo $f(x)$.

Ejemplo 2.12 *Consideremos el caso cuando la función objetivo $f(x)$ es una función de probabilidad con soporte finito $\{x_1, \ldots, x_n\}$. Supongamos que deseamos aplicar el método de aceptación y rechazo para generar valores de la distribución $f(x)$. Se puede tomar $c = n$ y $g(x)$ la función de probabilidad uniforme sobre $\{x_1, \ldots, x_n\}$, es decir, $g(x) = 1/n$ para $x = x_1, \ldots, x_n$. Tenemos entonces que $c\,g(x) = 1 \geqslant f(x)$. Puede comprobarse que el método de von Neumann (Proposición 2.3) y el método de aceptación y rechazo (Proposición 2.6) coinciden.* ∎

Como ocurre en el caso continuo, se debe observar que no es necesario que los soportes de las funciones de probabilidad $f(x)$ y $g(x)$ coincidan. En realidad pueden utilizarse funciones de probabilidad $g(x)$ cuyo soporte contiene al soporte de la función objetivo $f(x)$, pues si x_0 es un valor generado según la función de probabilidad $g(x)$ pero que no pertenece al soporte de $f(x)$, entonces tenemos que $f(x_0) = 0$ y la condición de aceptación $u \leqslant f(x)$ se convierte en $u \leqslant 0$, lo cual nunca se cumple, de modo que el valor x_0 se rechazará siempre. Una función de probabilidad

$g(x)$ con soporte mayor al de $f(x)$ puede no ser eficiente pero puede ayudar en la facilidad para generar valores de la función de probabilidad $g(x)$ de esas características si ya se cuenta con un mecanismo sencillo para ello. Por ejemplo, si la función objetivo $f(x)$ tiene soporte $\{0, 1, 2, \ldots, n\}$, como es el caso de la distribución binomial, se pueden usar valores de la distribución geométrica y el método de aceptación y rechazo para obtener valores con función de probabilidad $f(x)$. Esta situación se muestra en el siguiente ejemplo.

Ejemplo 2.13 *Supongamos que deseamos obtener valores que siguen la función de probabilidad* $\text{bin}\,(n, p)$ *a través de valores de la distribución geométrica, es decir, la función objetivo es*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad si \;\; x = 0, 1, \ldots, n,$$

la cual tiene soporte finito. Por otro lado, la función propuesta tiene soporte infinito y está dada por

$$g(x) = \theta(1-\theta)^x, \quad si \;\; x = 0, 1, \ldots$$

Para buscar una posible constante c tal que $c\,g(x) \geqslant f(x)$ usaremos la desigualdad

$$\binom{n}{x} \leqslant \frac{n!}{\lfloor n/2 \rfloor! \lceil n/2 \rceil!}, \quad para \;\; x = 0, 1, \ldots, n,$$

en donde $\lfloor x \rfloor$ denota el entero más grande menor o igual a x, y $\lceil x \rceil$ denota el entero más pequeño mayor o igual a x. Entonces

$$\begin{aligned} \binom{n}{x} p^x (1-p)^{n-x} &\leqslant \binom{n}{x} p^x \\ &\leqslant \frac{n!}{\lfloor n/2 \rfloor! \lceil n/2 \rceil!}\, p^x \\ &\leqslant \underbrace{\frac{n!}{\lfloor n/2 \rfloor! \lceil n/2 \rceil!} \left\lceil \frac{1}{1-p} \right\rceil}_{c} \underbrace{(1-p)p^x}_{g(x)}. \end{aligned}$$

Con la constante c indicada se cumple que $c\,g(x) \geqslant f(x)$ *para* $x = 0, 1, \ldots, n$, *en donde* $g(x)$ *es la función de probabilidad geométrica de parámetro* $\theta = 1 - p$. *Se puede así llevar a cabo el procedimiento de aceptación y rechazo. Aquellos valores propuestos* x *tales que* $x > n$ *son descartados pues no satisfacen la condición de aceptación.* ▪

Ejemplo 2.14 *Una variable aleatoria* X *es infinitamente divisible si se puede expresar como suma* n *de variables aleatorias independientes con la misma distribución que* X, *para cualquier* $n \geqslant 1$. *Por ejemplo, las distribuciones normal, ji-cuadrada, gama y Poisson satisfacen esa propiedad. Así, se puede generar, por ejemplo, el valor que una variable aleatoria Poisson cuando* $\lambda = 5$, *ya sea sumando 10 variables aleatorias independientes e idénticamente distribuidas, cada una con distribución* Poisson $(1/2)$, *o bien, sumando 5, pero cada una con distribución* Poisson (1). ▪

2.5. Método de Box y Muller

Este es un mecanismo para producir valores de la distribución $\mathrm{N}(0, 1)$. Sean X y Y dos variables aleatorias independientes, ambas con distribución normal estándar, es decir, su función de densidad conjunta es

$$f(x, y) = \frac{1}{2\pi} \exp\left\{-(x^2 + y^2)/2\right\}, \quad \text{para} \quad -\infty < x, y < \infty.$$

El método de Box y Muller está basado en la transformación de $\mathbb{R}^2$ que lleva las coordenadas cartesianas a las coordenadas polares:

$$(r, \theta) = \varphi(x, y) = \left(\sqrt{x^2 + y^2}, \tan^{-1}(y/x)\right), \quad x \neq 0.$$

Usando el teorema de cambio de variable para vectores aleatorios (véase el Apéndice A), puede comprobarse que la función de densidad del vector aleatorio $(R, \Theta) = \varphi(X, Y)$ es

$$f(r, \theta) = \frac{r}{2\pi} \exp\left\{-r^2/2\right\}, \quad \text{para} \quad r \geqslant 0, \quad \theta \in [0, 2\pi).$$

Está claro que generar un valor de la función de densidad $f(x, y)$ es equivalente a generar un valor de la función de densidad $f(r, \theta)$. Veremos una forma

de obtener valores (r, θ) de la función de densidad $f(r, \theta)$. De la expresión para $f(r, \theta)$, se desprenden las siguientes afirmaciones:

a) Las variables aleatorias R y Θ son independientes.

b) El ángulo Θ tiene función de densidad unif $[0, 2\pi)$.

c) El radio R tiene función de densidad $f(r) = r \exp\left\{-r^2/2\right\}$, para $r \geqslant 0$.

Estos resultados ayudan a obtener valores del vector (R, Θ). Para el ángulo Θ, podemos tomar una variable aleatoria V con distribución unif $(0, 1)$ y expresarla como

$$\Theta = 2\pi\, V.$$

Para el radio R, se puede usar el método de la función inversa para producir sus valores. Puede comprobarse que su función de distribución es

$$F(r) = 1 - \exp\left\{-r^2/2\right\}, \quad \text{para} \;\; r \geqslant 0,$$

con función inversa

$$F^{-1}(u) = \sqrt{-2\,\ln(1-u)}, \quad \text{para} \;\; 0 < u < 1.$$

Por lo tanto, si $U \sim$ unif $(0, 1)$, entonces R se puede expresar como

$$R = \sqrt{-2\,\ln U}.$$

Generado un valor de (R, θ) se puede obtener un valor de (X, Y) regresando a coordenadas cartesianas,

$$(X, Y) = \varphi^{-1}(R, \Theta) = (R\cos\Theta, R\,\mathrm{sen}\,\Theta).$$

De esta manera se llega al siguiente resultado.

Proposición 2.7 (Método de Box y Muller)
Sean U y V dos variables aleatorias independientes, cada una con distribución unif $(0, 1)$. *Las variables del vector aleatorio (X, Y) especificado abajo son independientes y tienen distribución* $\mathrm{N}(0, 1)$.

$$(X, Y) := \sqrt{-2\,\ln U}\,(\cos 2\pi\, V, \mathrm{sen}\, 2\pi\, V).$$

Finalmente, recordemos que si $X \sim \mathrm{N}(0, 1)$, entonces la transformación $\sigma\, X + \mu$ tiene distribución $\mathrm{N}(\mu, \sigma^2)$.

2.6. Método de Marsaglia

Este es otro procedimiento para generar valores de la distribución normal. Fue propuesto por G. Marsaglia y T. A. Bray [38] en 1964. El método está basado en el siguiente resultado.

Proposición 2.8 (Método de Marsaglia)
Sea (U, V) un vector aleatorio con distribución uniforme sobre el disco unitario $\{(u, v) : 0 < u^2 + v^2 < 1\}$, en donde se omite el origen. Sea $S = U^2 + V^2$. El vector (X, Y) especificado abajo está compuesto por variables aleatorias independientes, cada una con distribución $\mathrm{N}(0, 1)$.

$$(X, Y) := \left(\frac{U}{\sqrt{S}} \sqrt{-2 \ln S}, \frac{V}{\sqrt{S}} \sqrt{-2 \ln S} \right). \tag{2.6}$$

Demostración. Utilizaremos el teorema de cambio de variable para vectores aleatorios, el cual puede consultarse en el Apéndice A. La transformación que utilizaremos es

$$(u, v) \mapsto \varphi(u, v) = \left(\frac{u}{\sqrt{s}} \cdot \sqrt{-2 \ln s}, \frac{v}{\sqrt{s}} \cdot \sqrt{-2 \ln s} \right) = (x, y),$$

en donde $s = u^2 + v^2$. Puede comprobarse que $x^2 + y^2 = -2 \ln s$, de donde se obtiene que $s = \exp\left\{-(x^2 + y^2)/2\right\}$. La transformación inversa de φ es

$$\begin{aligned} (x, y) \mapsto \varphi^{-1}(x, y) &= (\varphi_1^{-1}(x, y), \varphi_2^{-1}(x, y)) \\ &= \left(\frac{x}{\sqrt{x^2 + y^2}} e^{-(x^2+y^2)/4}, \frac{y}{\sqrt{x^2 + y^2}} e^{-(x^2+y^2)/4} \right) \\ &= (u, v). \end{aligned}$$

Por lo tanto,

$$f_{X,Y}(x, y) = f_{U,V}(\varphi^{-1}(x, y)) \cdot |J(x, y)| = \frac{1}{\pi} \cdot |J(x, y)|,$$

en donde

$$J(x,y) = \begin{vmatrix} \dfrac{\partial}{\partial x}\,\varphi_1^{-1}(x,y) & \dfrac{\partial}{\partial y}\,\varphi_1^{-1}(x,y) \\ \dfrac{\partial}{\partial x}\,\varphi_2^{-1}(x,y) & \dfrac{\partial}{\partial y}\,\varphi_2^{-1}(x,y) \end{vmatrix}.$$

Puede comprobarse que

$$\begin{aligned} \frac{\partial}{\partial x}\,\varphi_1^{-1}(x,y) &= \frac{e^{-(x^2+y^2)/4}}{\sqrt{x^2+y^2}}\left(1 - \frac{x^2}{x^2+y^2} - \frac{1}{2}x^2\right), \\ \frac{\partial}{\partial y}\,\varphi_1^{-1}(x,y) &= \frac{e^{-(x^2+y^2)/4}}{\sqrt{x^2+y^2}}\left(-\frac{xy}{x^2+y^2} - \frac{1}{2}xy\right), \end{aligned}$$

en donde, por simetría, las otras dos derivadas requeridas tienen expresiones similares. Después de algunos cálculos algebraicos se puede demostrar que

$$J(x,y) = -\frac{1}{2}\,e^{-(x^2+y^2)/2}.$$

Se concluye que $f_{X,Y}(x,y) = (1/2\pi)\,e^{-(x^2+y^2)/2}$, lo que corresponde a una función de densidad conjunta de dos variables aleatorias independientes con distribución $\mathrm{N}(0,1)$.

■

Debe observarse que las variables U y V no son independientes, sin embargo, las variables X y Y lo son. En resumen, el método Marsaglia es el siguiente.

Método de Marsaglia

- Sea (u,v) un valor de la distribución unif $\{(u,v) : 0 < u^2 + v^2 < 1\}$.
- Sea $s = u^2 + v^2$. El par de números (x,y) definidos abajo son dos valores independientes de la distribución $\mathrm{N}(0,1)$.

$$(x,y) = \left(\frac{u}{\sqrt{s}}\sqrt{-2\ln s}, \frac{v}{\sqrt{s}}\sqrt{-2\ln s}\right).$$

Mediante la transformación usual, los valores de la distribución normal estándar se pueden llevar a valores de la distribución normal de parámetros arbitrarios.

2.7. El método del cociente de uniformes

En algunos casos, la distribución de una variable aleatoria continua se puede representar como la distribución que se obtiene al tomar el cociente de dos variables aleatorias con distribución uniforme. En esta sección explicaremos la manera en la que esta representación, junto con el método de aceptación y rechazo, producen un algoritmo para la simulación de ciertas variables aleatorias. Este método fue propuesto por A. J. Kinderman y J. F. Monahan [33] en 1977 y está basado en el siguiente resultado.

Proposición 2.9 (Método del cociente de uniformes)
Sea $h(x) \geqslant 0$ una función integrable, acotada y tal que:

a) $0 < \int_{-\infty}^{\infty} h(x)\,dx < \infty$.

b) $\sup_x |x|\sqrt{h(x)} < \infty$.

Defina la región

$$S = \left\{(u,v) : 0 < u < \sqrt{h(v/u)}\ ,\ v/u \in \operatorname{sop}(h)\right\} \subseteq \mathbb{R}^2,$$

en donde $\operatorname{sop}(h) = \{x : h(x) > 0\}$ *es el soporte de* $h(x)$. *Si* (U,V) *es un vector aleatorio con distribución uniforme sobre la región* S, *entonces el cociente* V/U *tiene función de densidad*

$$f(x) = \frac{h(x)}{\int_{-\infty}^{\infty} h(x)\,dx}. \tag{2.7}$$

Demostración. Comprobaremos primero que la región S tiene área finita.

Sea $(u, v) \in S$. En particular, tenemos que $u > 0$.

a) Para $v > 0$, $\quad u < \sqrt{h(v/u)} \;\Leftrightarrow\; v < (v/u)\sqrt{h(v/u)}$, en donde $v/u > 0$. Por lo tanto,

$$\begin{aligned} v &< \sup_{x>0} x\sqrt{h(x)} \\ &= \sup_{x>0} |x|\sqrt{h(x)} \\ &\leqslant \sup_{x\in\mathbb{R}} |x|\sqrt{h(x)}. \end{aligned} \tag{2.8}$$

b) Para $v < 0$, el análisis es similar, $\quad u < \sqrt{h(v/u)} \;\Leftrightarrow\; v > (v/u)\sqrt{h(v/u)}$, en donde $v/u < 0$. Por lo tanto,

$$\begin{aligned} v &> \inf_{x<0} x\sqrt{h(x)} \\ &= \inf_{x<0} -|x|\sqrt{h(x)} \\ &\geqslant -\sup_{x\in\mathbb{R}} |x|\sqrt{h(x)}. \end{aligned} \tag{2.9}$$

Incorporando los resultados (2.8) y (2.9) en la definición de la región S se obtiene que

$$S \subseteq \left[0, \sup_{x\in\mathbb{R}} \sqrt{h(x)}\right] \times \left[-\sup_{x\in\mathbb{R}} |x|\sqrt{h(x)}, \sup_{x\in\mathbb{R}} |x|\sqrt{h(x)}\right],$$

es decir, S está contenido en el rectángulo acotado indicado. En consecuencia, las hipótesis sobre la función $h(x)$ garantizan que la región S tiene área finita. Denotemos por $|S|$ al área de S. Entonces la función de densidad del vector (U, V) es

$$f_{U,V}(u, v) = \frac{1}{|S|}, \quad \text{para} \;\; (u, v) \in S.$$

Por la fórmula general para la función de densidad del cociente de dos variables aleatorias continuas (véase el apéndice A), tenemos que, para $-\infty <$

$x < \infty$,

$$\begin{aligned} f_{V/U}(x) &= \int_{-\infty}^{\infty} f_{U,V}(xu, u)\,|x|\,du \\ &= \int_{(u,xu)\in S} \frac{1}{|S|}\,|x|\,du \\ &= \frac{1}{|S|}\int_0^{\sqrt{h(x)}} u\,du \\ &= \frac{1}{2\,|S|}\,h(x). \end{aligned}$$

Como esta es una función de densidad, integrando se obtiene que

$$\int_{-\infty}^{\infty} h(x)\,dx = 2\,|S|. \tag{2.10}$$

Por lo tanto, la función de densidad $f_{V/U}(x)$ se puede escribir como aparece en (2.7). ∎

En conclusión, a partir de una función $h(x)$ con las características indicadas, se puede construir la función de densidad $f(x)$ dada por (2.7). Valores de esta distribución se pueden obtener a través del cociente V/U de dos variables aleatorias con distribución uniforme sobre la región S. A su vez, valores de estas variables aleatorias uniforme pueden obtenerse mediante una transformación simple de la distribución unif $(0, 1)$. Observemos que, de la identidad (2.10) se obtiene que el área de la región S es

$$|S| = \frac{1}{2} \cdot \int_{-\infty}^{\infty} h(x)\,dx.$$

En el caso cuando $h(x)$ sea una función de densidad, tenemos que $|S| = 1/2$, y la proposición recién demostrada establece que el cociente V/U tiene función de densidad $h(x)$. Veremos a continuación algunos ejemplos.

Ejemplo 2.15 (Distribución exponencial)
Sabemos que se pueden obtener valores de la distribución exponencial con suma facilidad a través del método de la función inversa, sin embargo, usaremos esa misma distribución en este ejemplo para ilustrar la forma en

la que se aplica el método del cociente de uniformes. Recordemos que es suficiente generar valores de X *con distribución* $\exp(1)$ *pues* $X/\lambda \sim \exp(\lambda)$.

Supongamos entonces que X *tiene función de densidad* $f(x) = e^{-x}$, *cuyo soporte es* $(0, \infty)$. *Tomaremos a* $h(x)$ *como la misma función* $f(x)$. *No es difícil verificar que se cumplen las condiciones*

$$0 < \int_{-\infty}^{\infty} h(x)\, dx < \infty \qquad y \qquad \sup_x \ |x| \sqrt{h(x)} < \infty.$$

Siguiendo el procedimiento explicado antes, se debe considerar la región

$$\begin{aligned} S &= \left\{(u,v) : 0 < u < \sqrt{h(v/u)}\ ,\ v/u \in \operatorname{sop}(h)\right\} \\ &= \left\{(u,v) : 0 < u < \sqrt{\exp\{-v/u\}}\ ,\ v/u \in (0,\infty)\right\}. \end{aligned}$$

Puesto que $u > 0$, *tenemos que también* $v > 0$ *pues se debe satisfacer que* $v/u \in (0, \infty)$. *Colocando la raíz cuadrada dentro de la exponencial,*

$$S = \{(u,v) : 0 < u < \exp\{-v/(2u)\}\ ,\ v > 0\}.$$

Ahora observamos que $-v/(2u)$ *es negativo y, por lo tanto, el valor de la exponencial tiene a* 1 *como cota superior. Además, resolviendo para* v *en la desigualdad que involucra a* u *y* v *conjuntamente, se encuentra finalmente la siguiente expresión de la región* S,

$$S = \{(u,v) : 0 < u < 1\ ,\ 0 < v < -2u \ln u\}.$$

Gráficamente, esta región se muestra en la Figura 2.7.

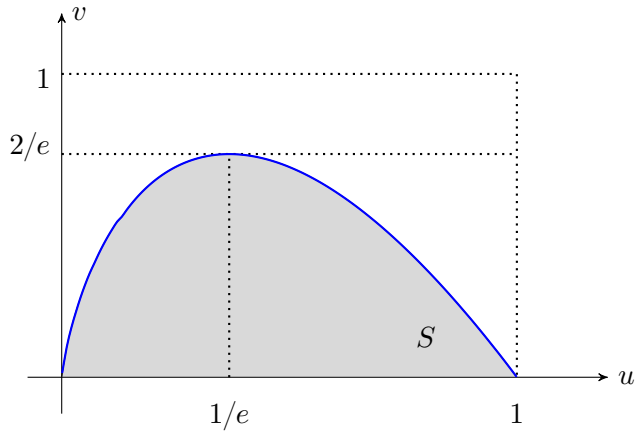


Figura 2.7: Región de aceptación S.

Puede comprobarse para este caso particular que el área de la región S es

$$\begin{aligned} |S| &= \int_0^1 -2u \ln u \, du \\ &= -\int_0^1 \left(\frac{d}{du} u^2\right) \ln u \, du \\ &= 1/2. \end{aligned}$$

Para obtener valores del vector (U, V) con distribución uniforme en la región S utilizaremos el método de aceptación y rechazo. Para ello generaremos valores de la distribución unif $(0, 1) \times (0, 1)$. *Observemos que este rectángulo contiene a la región S y que un muestreo óptimo (muestro de eficiencia máxima) se obtiene cuando $(U, V) \sim$* unif $(0, 1) \times (0, 2/e)$. *El procedimiento se muestra en el siguiente recuadro.*

Método del cociente de uniformes para generar valores $\exp(\lambda)$

- *Sea* (u, v) *un valor de la distribución* $\text{unif}\,(0,1) \times (0,1)$.
- *Cuando* (u, v) *es tal que* $v < -2u \ln u$, *se acepta a* (u, v) *como un valor de la distribución* $\text{unif}\,(S)$.
- *El cociente* $x = v/u$ *produce un valor de la distribución* $\exp(1)$.
- *Finalmente, la operación* x/λ *produce un valor de la distribución* $\exp(\lambda)$.

■

Ejemplo 2.16 (Distribución normal)
Aplicaremos el método del cociente de uniformes para generar valores de la distribución $\text{N}(\mu, \sigma^2)$. *Recordemos que es suficiente obtener valores de* X *con distribución* $\text{N}(0,1)$ *pues* $\mu + \sigma X \sim \text{N}(\mu, \sigma^2)$.

Sea X *con función de densidad* $f(x) = (2\pi)^{-1/2} \exp\left\{-x^2/2\right\}$, *con soporte* $(-\infty, \infty)$. *Podemos considerar a* $h(x)$ *como* $f(x)$, *pero por simplicidad en la determinación de la región* S *tomaremos* $h(x) = \exp\left\{-x^2/2\right\}$. *No es difícil verificar que se cumplen las condiciones*

$$0 < \int_{-\infty}^{\infty} h(x)\,dx < \infty \qquad y \qquad \sup_x\ |x|\sqrt{h(x)} < \infty.$$

Siguiendo el procedimiento explicado antes, se debe considerar la región

$$\begin{aligned} S &= \left\{(u,v) : 0 < u < \sqrt{h(v/u)}\ ,\ v/u \in \text{sop}\,(h)\right\} \\ &= \left\{(u,v) : 0 < u < \sqrt{\exp\left\{-v^2/(2u^2)\right\}}\ ,\ v/u \in (-\infty,\infty)\right\} \\ &= \left\{(u,v) : 0 < u < \exp\left\{-v^2/(4u^2)\right\}\ ,\ v \in (-\infty,\infty)\right\} \\ &= \left\{(u,v) : 0 < u < 1\ ,\ v^2 < -4u^2 \ln u\right\}. \end{aligned}$$

Gráficamente, esta región se muestra en la Figura 2.8.

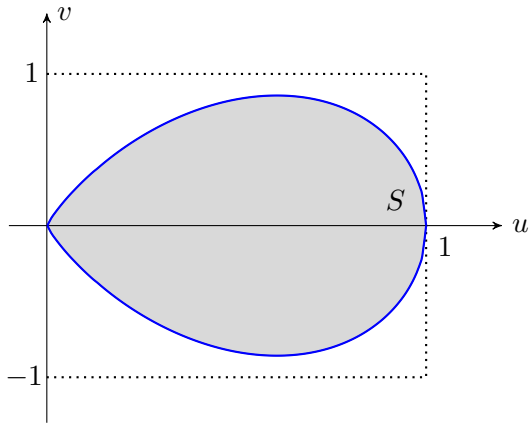


Figura 2.8: Región de aceptación S.

Puede comprobarse para este caso particular que el área de la región S es

$$\begin{aligned}
|S| &= 2\int_0^1 \sqrt{-4u^2\,\ln u}\,du \\
&= 2\int_0^1 2u\sqrt{\ln(1/u)}\,du \\
&= 8\int_0^\infty v^2\,e^{-2v^2}\,dv, \qquad v^2 := \ln(1/u) \\
&= 4\sqrt{2\pi\sigma^2}\,E(W^2), \qquad W \sim \mathrm{N}(0,\sigma^2),\ \textit{con}\ \ \sigma^2 = 1/4 \\
&= \frac{1}{2}\sqrt{2\pi} \\
&= \frac{1}{2}\int_{-\infty}^{\infty} h(x)\,dx.
\end{aligned}$$

Para producir valores del vector (U,V) con distribución uniforme en la región S utilizaremos nuevamente el método de aceptación y rechazo. Para ello generaremos valores de la distribución unif $(0,1)\times(-1,1)$. *Observemos que el rectángulo de esta distribución contiene a la región S y que un rectángulo más reducido es posible incrementando la eficiencia del muestreo. El procedimiento se muestra en el siguiente recuadro.*

Método del cociente de uniformes para generar valores $\mathrm{N}(\mu, \sigma^2)$

- *Sea* (u, v) *un valor de la distribución* $\text{unif}\,(0,1) \times (-1,1)$.
- *Cuando* (u, v) *es tal que* $v^2 < -4u^2 \ln u$, *se acepta a* (u, v) *como un valor de la distribución* $\text{unif}(S)$.
- *El cociente* $x = v/u$ *produce un valor de la distribución* $\mathrm{N}(0,1)$.
- *Finalmente, la operación* $\mu + \sigma x$ *produce un valor de la distribución* $\mathrm{N}(\mu, \sigma^2)$.

■

Otros ejemplos en donde se aplica el método del cociente de uniformes se encuentran en la sección de ejercicios.

2.8. Generación de valores de variables aleatorias multivariadas

En esta sección estudiaremos los métodos de von Neumann, y de aceptación y rechazo, para generar valores de vectores aleatorios, tanto en el caso discreto como continuo. Estos procedimientos son similares al caso unidimensional.

Método de von Neumann: caso discreto multivariado

El siguiente enunciado contiene la versión del método de von Neumann para vectores aleatorios discretos. Es una extensión de la Proposición 2.4. Por simplicidad en la escritura, sólo consideraremos el caso bidimensional. El resultado puede extenderse fácilmente a cualquier dimensión.

Proposición 2.10 *Sea $f(x,y)$ una función de probabilidad bivariada con soporte el conjunto finito $S = \{x_1,\ldots,x_n\} \times \{y_1,\ldots,y_m\}$. Sea (X,Y) con distribución uniforme sobre S e independiente de $U \sim \text{unif}\,(0,1)$. Entonces*

$$P(X = x, Y = y \,|\, U \leqslant f(X,Y)\,) = f(x,y), \quad \textit{para} \;\; (x,y) \in S.$$

Demostración. Sea $(x,y) \in S$. Observemos primero que

$$\begin{aligned} &P(X = x, Y = y) = \frac{1}{nm}, \\ &P(U \leqslant f(X,Y) \,|\, X = x, Y = y) = P(U \leqslant f(x,y)) = f(x,y). \end{aligned}$$

Por definición de probabilidad condicional,

$$\begin{aligned} P(X = x, Y = y \,|\, U \leqslant f(X,Y)\,) &= \frac{P(X = x, Y = y, U \leqslant f(X,Y)\,)}{P(U \leqslant f(X,Y)\,)} \\ &= \frac{P(U \leqslant f(X,Y) \,|\, X = x, Y = y)\, P(X = x, Y = y)}{\displaystyle\sum_{(u,v)\in S} P(U \leqslant f(X,Y) \,|\, X = u, Y = v)\, P(X = u, Y = v)} \\ &= \frac{f(x,y)/(nm)}{\displaystyle\sum_{(u,v)\in S} f(u,v)/(nm)} \\ &= f(x,y). \end{aligned}$$

■

Método de von Neumann: caso continuo multivariado

El siguiente enunciado es una extensión al caso multidimensional del método de von Neumann unidimensional estudiado en la Proposición 2.3. Por brevedad, se enuncia y demuestra en el caso bidimensional pero es evidente la forma en la que puede expresarse el resultado al caso de cualquier dimensión mayor.

Consideraremos que nuestra función objetivo es una función de densidad dada $f(x, y)$ que satisface ciertas condiciones. La demostración del método es análoga al caso unidimensional.

Proposición 2.11 *Sea $f(x, y)$ una función de densidad bivariada con soporte el rectángulo acotado $(a, b) \times (c, d)$ y suponga que es acotada, es decir, existe una constante $M > 0$ tal que $0 \leqslant f(x, y) \leqslant M$. Sean U_1, U_2 y U_3 variables aleatorias independientes con distribución* unif $(0, 1)$ *y defina*

$$\begin{aligned} X &= a + (b - a)U_1, \\ Y &= c + (d - c)U_2, \\ U &= MU_3. \end{aligned}$$

Entonces la distribución de (X, Y) condicionada al evento $(U \leqslant f(X, Y))$ es $f(x, y)$.

Demostración. Sean $x \in (a, b)$ y $y \in (c, d)$. Por definición de probabilidad condicional,

$$P(X \leqslant x, Y \leqslant y \,|\, U \leqslant f(X, Y)) = \frac{P(X \leqslant x, Y \leqslant y, U \leqslant f(X, Y))}{P(U \leqslant f(X, Y))},$$

en donde el numerador es

$$\begin{aligned} &\int_a^x \int_c^y \int_0^{f(a,b)} f_{X,Y,U}(a, b, u)\, du\, db\, da \\ &= \int_a^x \int_c^y \int_0^{f(a,b)} f_{U|X,Y}(u|a, b)\, f(a, b)\, du\, db\, da \\ &= \int_a^x \int_c^y \int_0^{f(a,b)} \frac{1}{M}\, \frac{1}{b-a}\, \frac{1}{d-c}\, du\, db\, da \\ &= \frac{1}{M}\, \frac{1}{b-a}\, \frac{1}{d-c} \int_a^x \int_c^y f(a, b)\, db\, da. \end{aligned}$$

El cálculo del denominador es similar y resulta ser

$$\begin{aligned}
&\int_a^b \int_c^d \int_0^{f(a,b)} f_{X,Y,U}(a,b,u)\, du\, db\, da \\
&= \frac{1}{M}\,\frac{1}{b-a}\,\frac{1}{d-c} \underbrace{\int_a^b \int_c^d f(a,b)\, db\, da}_{=1} \\
&= \frac{1}{M}\,\frac{1}{b-a}\,\frac{1}{d-c}.
\end{aligned}$$

Al obtener el cociente se obtiene

$$P(X \leqslant x, Y \leqslant y \,|\, U \leqslant f(X,Y)) = \int_a^x \int_c^y f(a,b)\, db\, da,$$

lo cual significa que $f(x,y)$ es la función de densidad de $(X,Y)|(U \leqslant f(X,Y))$.

■

Observamos nuevamente que la distribución no condicional del vector (X,Y) es uniforme, pero su distribucion condicionada al evento $(U \leqslant f(X,Y))$ es la función de densidad objetivo $f(x,y)$.

Es interesante observar que el procedimiento permanece válido en un caso más general cuando el soporte de la función objetivo $f(x,y)$ es una región $S \subseteq (a,b) \times (c,d)$, cuya área es positiva. Cuando un punto (x,y) es generado de la distribución unif $(a,b) \times (c,d)$ y se encuentra fuera de S, $f(x,y) = 0$ y la condición $(u \leqslant f(x,y))$ nunca se cumple, de modo que el punto (x,y) es rechazado como un valor de la función de densidad $f(x,y)$.

Método de aceptación y rechazo: caso continuo multivariado

Utilizaremos el método de aceptación y rechazo para generar valores de un vector aleatorio continuo con una función de densidad dada. Este procedimiento es similar al caso unidimensional. Reescribiremos el procedimiento como un recordatorio.

Sea $f(x_1, \ldots, x_n)$ una función de densidad que tiene como soporte un conjunto A contenido en $\mathbb{R}^n$. Esta es nuestra función objetivo. Deseamos

obtener vectores que sigan esta distribución. El método de aceptación y rechazo está basado en el siguiente resultado. Su demostración es idéntica al caso unidimensional.

Proposición 2.12 *Sea $f(x_1, \ldots, x_n)$ una función de densidad con soporte $A \subseteq \mathbb{R}^n$. Sea $(X_1, \ldots, X_n)$ un vector aleatorio continuo con función de densidad $g(x_1, \ldots, x_n)$ tal que:*

1. *El soporte de $g(x_1, \ldots, x_n)$ contiene al soporte A.*
2. *Se conoce una forma de generar valores de la distribución $g(x_1, \ldots, x_n)$.*
3. *Existe una constante $c \geqslant 1$ tal que $c \cdot g(x_1, \ldots, x_n) \geqslant f(x_1, \ldots, x_n)$.*

Sea $U \sim \text{unif}\,(0, 1)$ independiente del vector $(X_1, \ldots, X_n)$. Entonces la distribución de $(X_1, \ldots, X_n)$ condicionada al evento

$$\left(U \leqslant \frac{f(X_1, \ldots, X_n)}{c \cdot g(X_1, \ldots, X_n)} \right).$$

tiene función de densidad $f(x_1, \ldots, x_n)$.

Demostración. Para hacer la escritura más corta consideraremos únicamente el caso bidimensional. El análisis puede extenderse sin dificultad al

caso de dimensión n. Para cualquier $(x_1, x_2) \in A \subset \mathbb{R}^2$,

$$
\begin{aligned}
&P\left(X_1 \leqslant x_1, X_2 \leqslant x_2 \,\middle|\, U \leqslant \frac{f(X_1, X_2)}{c \cdot g(X_1, X_2)}\right) \\
&= \frac{P(X_1 \leqslant x_1, X_2 \leqslant x_2, U \leqslant f(X_1, X_2)/[c \cdot g(X_1, X_2)])}{P(U \leqslant f(X_1, X_2)/[c \cdot g(X_1, X_2)])} \\
&= \frac{\displaystyle\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\int_{0}^{f(\xi_1,\xi_2)/[c\cdot g(\xi_1,\xi_2)]} f_{X_1,X_2,U}(\xi_1,\xi_2,u)\,du\,d\xi_2\,d\xi_1}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{f(\xi_1,\xi_2)/[c\cdot g(\xi_1,\xi_2)]} f_{X_1,X_2,U}(\xi_1,\xi_2,u)\,du\,d\xi_2\,d\xi_1} \\
&= \frac{\displaystyle\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\int_{0}^{f(\xi_1,\xi_2)/[c\cdot g(\xi_1,\xi_2)]} f_{U|X_1,X_2}(u|\xi_1,\xi_2)\,g(\xi_1,\xi_2)\,du\,d\xi_2\,d\xi_1}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{f(\xi_1,\xi_2)/[c\cdot g(\xi_1,\xi_2)]} f_{U|X_1,X_2}(u|\xi_1,\xi_2)\,g(\xi_1,\xi_2)\,du\,d\xi_2\,d\xi_1},
\end{aligned}
$$

usando que U es independiente de X_1 y X_2,

$$
\begin{aligned}
&= \frac{\displaystyle\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\int_{0}^{f(\xi_1,\xi_2)/[c\cdot g(\xi_1,\xi_2)]} f_U(u)\,g(\xi_1,\xi_2)\,du\,d\xi_2\,d\xi_1}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{f(\xi_1,\xi_2)/[c\cdot g(\xi_1,\xi_2)]} f_U(u)\,g(\xi_1,\xi_2)\,du\,d\xi_2\,d\xi_1} \\
&= \frac{\displaystyle\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \frac{f(\xi_1,\xi_2)}{c\cdot g(\xi_1,\xi_2)}\,g(\xi_1,\xi_2)\,d\xi_2\,d\xi_1}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{f(\xi_1,\xi_2)}{c\cdot g(\xi_1,\xi_2)}\,g(\xi_1,\xi_2)\,d\xi_2\,d\xi_1} \\
&= \frac{\displaystyle\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(\xi_1,\xi_2)\,d\xi_2\,d\xi_1}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi_1,\xi_2)\,d\xi_2\,d\xi_1} \\
&= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(\xi_1,\xi_2)\,d\xi_2\,d\xi_1.
\end{aligned}
$$

■

Ejemplo 2.17 *Sea $f(x_1, x_2)$ la función de densidad uniforme sobre una región A del plano cartesiano $\mathbb{R}^2$, posiblemente irregular pero que es acotada y tal que su área $|A|$ está bien definida. Entonces*

$$f(x_1, x_2) = \begin{cases} \dfrac{1}{|A|} & si \;\; (x_1, x_2) \in A, \\ 0 & en\ otro\ caso. \end{cases}$$

Sea (X_1, X_2) un vector aleatorio con función de densidad uniforme sobre el rectángulo $R = (a_1, b_1) \times (a_2, b_2)$ suficientemente amplio tal que $A \subseteq (a_1, b_1) \times (a_2, b_2)$.

$$g(x_1, \ldots, x_n) = \begin{cases} \dfrac{1}{(b_1 - a_1)(b_2 - a_2)} & si \;\; (x_1, x_2) \in R, \\ 0 & en\ otro\ caso. \end{cases}$$

Claramente, $|R| \geqslant |A|$. Tomando $c = |R|/|A| \geqslant 1$ se verifica la condición

$$c \cdot g(x_1, x_2) \geqslant f(x_1, x_2).$$

Se genera entonces un valor (x_1, x_2) dentro del rectángulo R con distribución uniforme y al mismo tiempo, de manera independiente, se genera un valor u de la distribución unif $(0, 1)$. *El valor (x_1, x_2) se acepta como valor de la función de densidad f cuando*

$$u \leqslant \frac{f(x_1, x_2)}{c \cdot g(x_1, x_2)}.$$

Este ejemplo se puede extender al caso de funciones de densidad n dimensionales con soporte una región $A \subset \mathbb{R}^n$, acotada y con área bien definida, por ejemplo, una bola en $\mathbb{R}^n$ de radio $r > 0$ y centrada en el origen, i.e.

$$B(r) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 \leqslant r^2\}.$$

■

Método de aceptación y rechazo: caso discreto multivariado

Concluimos esta sección con la versión multivariada del método de aceptación y rechazo en el caso discreto. El enunciado es el siguiente y es una extensión de la Proposición 2.6. Como antes, nos concentraremos sólo en el caso bidimensional.

Proposición 2.13 *Sea $f(x,y)$ una función de probabilidad bivariada con soporte $S = \{x_1, x_2, \ldots\} \times \{y_1, y_2, \ldots\}$. Sea (X,Y) un vector aleatorio discreto con función de probabilidad $g(x,y)$ que satisface las siguientes propiedades: (a) El soporte de $g(x,y)$ es también el conjunto finito S. (b) Existe una constante $c \geqslant 1$ tal que $c\,g(x) \geqslant f(x)$ para $(x,y) \in S$ y (c) Se conoce una manera de generar valores de (X,Y). Sea $U \sim \text{unif}\,(0, c\,g(X,Y))$. Entonces*

$$P(X = x, Y = y \,|\, U \leqslant f(X,Y)) = f(x,y), \quad \textit{para} \;\; (x,y) \in S.$$

Demostración. Sea (x,y) cualquier elemento del conjunto S. Observemos primero que

$$\begin{aligned} P(X = x, Y = y) &= g(x,y) > 0, \\ P(U \leqslant f(X,Y) \,|\, X = x, Y = y) &= \frac{f(x,y)}{c\,g(x,y)}. \end{aligned}$$

Entonces

$$\begin{aligned} P(X = x, Y = y \,|\, U \leqslant f(X,Y)) & \\ &= \frac{P(X = x, Y = y, U \leqslant f(X,Y))}{P(U \leqslant f(X,Y))} \\ &= \frac{P(U \leqslant f(X) \,|\, X = x, Y = y) P(X = x, Y = y)}{\sum_{i,j=1}^{\infty} P(U \leqslant f(X,Y) \,|\, X = x_i, Y = y_j) P(X = x_i, Y = y_j)} \\ &= \frac{\frac{f(x,y)}{c\, g(x,y)}\, g(x,y)}{\sum_{i,j=1}^{\infty} \frac{f(x_i,y_j)}{c\, g(x_i,y_j)}\, g(x_i,y_j)} \\ &= f(x,y). \end{aligned}$$

■

Ejemplo 2.18 *Sea $f(x,y)$ cualquier función de probabilidad bivariada con soporte finito $S = \{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_m\}$. Se puede tomar como $g(x,y)$ a la función de probabilidad uniforme sobre S, es decir, $g(x,y) = 1/(nm)$ para $(x,y) \in S$. Es claro que se pueden obtener valores de esta distribución uniforme con facilidad y que existe una constante c tal que $c\, g(x,y) \geqslant f(x,y)$. En conclusión, a través de la distribución uniforme discreta y el método de aceptación y rechazo, se pueden generar valores de cualquier distribución bivariada con soporte finito. El mismo método puede usarse para distribuciones discretas de cualquier dimensión con soporte finito.* ■

2.9. Ejercicios

Método de la función inversa

27. *Distribución uniforme.* Sea X una variable aleatoria con distribución unif (a,b). Use el método de la función inversa para demostrar que un valor x de esta distribución se puede generar a partir de la expresión que aparece abajo, en donde u es un valor de la distribución unif $(0,1)$.

$$x = a + (b-a)u.$$

28. *Distribución discreta.* Sea X una variable aleatoria discreta con función de probabilidad

$$f(x) = \begin{cases} 0.4 & \text{si } x = 1, \\ 0.3 & \text{si } x = 2, \\ 0.2 & \text{si } x = 3, \\ 0.1 & \text{si } x = 4. \end{cases}$$

Se busca aplicar el método de la función inversa para producir valores de esta distribución.

   a) Grafique $f(x)$.

   b) Encuentre y grafique $F(x)$.

   c) Encuentre y grafique $F^{-1}(x)$.

   d) Elabore un programa de cómputo que utilice el método de la función inversa para generar $n = 500$ valores $x_1, \ldots, x_n$ de X.

   e) Elabore un histograma de probabilidad (en donde el área de las barras suma uno) de los números obtenidos y compare con la gráfica de la función de densidad.

29. *Distribución exponencial.* Sea X una variable aleatoria con distribución $\exp(\lambda)$. Use el método de la función inversa para demostrar que un valor x de esta distribución se puede obtener a partir de la expresión que aparece abajo, en donde u es un valor de la distribución unif $(0, 1)$.

$$x = -\frac{1}{\lambda} \ln u.$$

30. *Distribución Weibull.* Sea X una variable aleatoria con distribución Weibull (r, λ). Use el método de la función inversa para demostrar que un valor x de la variable X se puede obtener a partir de la expresión que aparece abajo, en donde u es un valor de la distribución unif $(0, 1)$.

$$x = \frac{1}{\lambda} \exp\left\{ \frac{1}{r} \ln\left(\ln\left(1/u\right) \right) \right\}.$$

31. *Distribución Cauchy.* Sea X una variable aleatoria con distribución Cauchy (a, b). Use el método de la función inversa para demostrar que

un valor x de esta distribución se puede obtener a partir de la expresión que aparece abajo, en donde u es un valor de la distribución unif $(0, 1)$.

$$x = a + b \tan[(u - 1/2)\pi].$$

32. *Distribución Pareto Tipo I.* Sea X una variable aleatoria con distribución Pareto (a, b) tipo I. Use el método de la función inversa para demostrar que un valor x de esta distribución se puede obtener a partir de la expresión que aparece abajo, en donde u es un valor de la distribución unif $(0, 1)$.
$$x = b\, u^{-1/a}.$$

33. *Mínimo.* Encuentre la función de distribución, y su inversa, de la variable aleatoria
$$X_{(1)} := \text{mín}\,\{X_1, \ldots, X_n\},$$
en términos de la función de distribución $F(x)$ de una muestra aleatoria $X_1, \ldots, X_n$. Use el método de la función inversa para explicar un método para generar valores de la variable aleatoria $X_{(1)}$.

34. *Distribución mixta.* Sea Y una variable aleatoria con distribución exp (λ) y sea $M > 0$ una constante. Defina la variable aleatoria $X = \text{mín}\,\{Y, M\}$ y denote por $F(x)$ su función de distribución.

    a) Encuentre y grafique $F(x)$.

    b) Encuentre y grafique $F^{-1}(u)$.

    c) Compruebe que $F(F^{-1}(u)) \geqslant u$, para $0 < u < 1$.

    d) Compruebe que $F^{-1}(F(x)) \leqslant x$, para x tal que $0 < F(x) < 1$.

    e) Explique la manera en la que pueden obtenerse valores de la variable aleatoria X por el método de la función inversa.

35. *Distribución mixta.* Sea Y una variable aleatoria con distribución exp (λ) y sea $M > 0$ una constante. Defina la variable aleatoria $X = \text{máx}\,\{Y, M\}$ y denote por $F(x)$ su función de distribución.

    a) Encuentre y grafique $F(x)$.

    b) Encuentre y grafique $F^{-1}(u)$.

c) Compruebe que $F(F^{-1}(u)) \geqslant u$, para $0 < u < 1$.

d) Compruebe que $F^{-1}(F(x)) \leqslant x$, para x tal que $0 < F(x) < 1$.

e) Explique la manera en la que pueden generarse valores de la variable aleatoria X por el método de la función inversa.

36. Sea X una variable aleatoria con función de distribución

$$F(x) = \begin{cases} 0 & \text{si } x \leqslant -2, \\ (x+2)/2 & \text{si } -2 < x < -1, \\ 1/2 & \text{si } -1 \leqslant x < 1, \\ x/2 & \text{si } 1 \leqslant x < 2, \\ 1 & \text{si } x \geqslant 2. \end{cases}$$

a) Grafique $F(x)$.

b) Encuentre y grafique $F^{-1}(u)$.

c) Compruebe que $F(F^{-1}(u)) \geqslant u$, para $0 < u < 1$.

d) Compruebe que $F^{-1}(F(x)) \leqslant x$, para x tal que $0 < F(x) < 1$.

e) Explique la manera en la que pueden generarse valores de la variable aleatoria X por el método de la función inversa.

f) Encuentre y grafique la función de densidad $f(x)$ de la variable X.

g) Usando una computadora y el método de la función inversa, genere $n = 500$ valores $x_1, \ldots, x_n$ de la variable X, elabore un histograma de probabilidad y compare con $f(x)$.

h) Demuestre que $E(X) = 0$.

i) Compruebe que $\bar{x} = (x_1 + \cdots + x_n)/n \approx 0$.

Generación de valores de variables aleatorias discretas

37. *Distribución Bernoulli.* Sea $U \sim \text{unif}\,(0,1)$ y defina la variable aleatoria X como aparece abajo, en donde $0 < p < 1$. Demuestre que X tiene distribución $\text{Ber}\,(p)$.

$$X = 1_{(0,p]}(U) = \begin{cases} 1 & \text{si } U \leqslant p, \\ 0 & \text{si } U > p. \end{cases}$$

38. Considere una variable aleatoria discreta X con distribución

$$f(x) = \begin{cases} 0.3 & \text{si } x = 1, \\ 0.5 & \text{si } x = 2, \\ 0.2 & \text{si } x = 3, \\ 0 & \text{en otro caso.} \end{cases}$$

a) Grafique $f(x)$

b) Encuentre y grafique $F(x)$.

c) Encuentre y grafique $F^{-1}(u)$.

d) Encuentre $E(X)$.

e) Encuentre $\text{Var}(X)$.

f) Elabore un programa de cómputo que use el método de la función inversa para generar $n = 200$ valores $x_1, \ldots, x_n$ de X.

g) Elabore un histograma de probabilidad de los valores obtenidos $x_1, \ldots, x_n$ y compare con la gráfica de $f(x)$.

h) Compruebe que $E(X) \approx \bar{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$.

i) Compruebe que $\text{Var}(X) \approx \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2$.

39. Elabore un programa de cómputo que use el método de la función inversa para producir $n = 100$ valores de una variable aleatoria discreta X con distribución:

a) $\text{bin}\,(m, p)$, con $m = 10$ y $p = 1/3$.

b) $\text{geo}\,(p)$, con $p = 3/4$.

c) $\text{Poisson}\,(\lambda)$, con $\lambda = 2$.

40. *Número finito de valores.* Sea X una variable aleatoria con $n \geqslant 2$ posibles valores $x_1 < x_2 < \cdots < x_n$ y probabilidades $p_1, p_2, \ldots, p_n$, respectivamente. Argumente la razón por la que el número promedio de preguntas que se llevan a cabo en el diagrama de flujo de la Figura 2.5 es

$$\sum_{i=1}^{n-1} i\, p_i + (n-1)\, p_n.$$

Método de von Neumann

41. Elabore un programa de cómputo para generar 500 valores de las siguientes distribuciones usando el método de von Neumann. Use como valor de M el valor máximo de $f(x)$ en cada caso. Calcule además la eficiencia.

$$a)\ f(x) = \begin{cases} x & \text{si } 0 < x < 1, \\ 2 - x & \text{si } 1 \leqslant x < 2, \\ 0 & \text{en otro caso.} \end{cases}$$

$$b)\ f(x) = \begin{cases} 1 - x & \text{si } 0 < x < 1, \\ x - 1 & \text{si } 1 \leqslant x < 2, \\ 0 & \text{en otro caso.} \end{cases}$$

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

$$d)\ f(x) = \begin{cases} x + 1 & \text{si } -1 < x < 0, \\ x & \text{si } 0 \leqslant x < 1, \\ 0 & \text{en otro caso.} \end{cases}$$

Método de aceptación y rechazo

42. *Distribución normal.* En el contexto de la aplicación del método de aceptación y rechazo para generar valores de la distribución $\mathrm{N}(\mu, \sigma^2)$ expuesto en el Ejemplo 2.11, demuestre directamente los siguientes resultados:

    a) Si Z tiene distribución $\mathrm{N}(0, 1)$, entonces $|Z|$ tiene función de densidad
    $$f(x) = \sqrt{\frac{2}{\pi}}\, e^{-x^2/2}, \quad \text{para } x > 0.$$

    b) Si S tiene distribución unif $\{+1, -1\}$, entonces $S\,|Z|$ tiene distribución $\mathrm{N}(0, 1)$, en donde $|Z|$ es como en el inciso anterior.

    c) Si $X \sim \exp(\lambda)$ con $\lambda = 1$, y $U \sim \text{unif}\,(0, 1)$ son independientes, entonces la distribución de X condicionada al evento

$(U \leqslant \exp\left\{-(X-1)^2/2\right\})$ tiene la función de densidad que aparece abajo. Esta es la función de densidad del valor absoluto de $Z \sim \mathrm{N}(0,1)$.

$$f(x) = \sqrt{\frac{2}{\pi}}\, e^{-x^2/2}, \quad \text{para} \;\; x > 0.$$

d) Si V_1 y V_2 son variables aleatorias independientes con idéntica distribución $\exp(\lambda)$ con $\lambda = 1$, entonces

$$P(V_1 \geqslant (V_2 - 1)^2/2) = \sqrt{\frac{\pi}{2\,e}}.$$

Método de Box y Muller

43. Demuestre el siguiente resultado que fue utilizado en la derivación del método de Box y Muller. Si (X, Y) es un vector aleatorio continuo con función de densidad

$$f(x,y) = \frac{1}{2\pi} \exp\left\{-(x^2+y^2)/2\right\}, \quad \text{para} \quad -\infty < x, y < \infty,$$

entonces el vector $(R, \Theta) = (\sqrt{X^2+Y^2}, \tan^{-1}(Y/X))$, con $X \neq 0$ tiene función de densidad

$$f(r,\theta) = \frac{r}{2\pi} \exp\left\{-r^2/2\right\}, \quad \text{para} \;\; (r,\theta) \in (0,\infty) \times (0, 2\pi].$$

44. Demuestre que si $W \sim \exp(\lambda)$ con $\lambda = 1/2$, entonces la variable $R = \sqrt{W}$ tiene la función de densidad $f(r)$ que aparece abajo. Esta es la función de densidad de la variable aleatoria R que aparece en el método de Box y Muller.

$$f(r) = r \exp\left\{-r^2/2\right\}, \;\; \text{para} \;\; r \geqslant 0.$$

Por el método de la función inversa, un valor de W se puede expresar como $-2\ln u$, en donde u es un valor de la distribución unif $(0,1)$, de modo que un valor de R es $\sqrt{-2\ln u}$. Esta es una expresión equivalente a la indicada en el método de Box y Muller.

45. Utilizando el método de Box y Muller, elabore un programa de cómputo para generar 1000 valores de una variable aleatoria con distribución $\mathrm{N}(\mu, \sigma^2)$, con valores para μ y σ^2 de su elección. Elabore un histograma con los valores obtenidos.

El método del cociente de uniformes

46. Considere el método del cociente de uniformes para generar valores de una variable aleatoria con función de densidad

$$h(x) = \begin{cases} 1 - x/2 & \text{si } 0 < x < 2, \\ 0 & \text{en otro caso.} \end{cases}$$

a) Determine analíticamente y muestre en un plano la región asociada $S \subseteq \mathbb{R}^2$. Compruebe que $|S| = 1/2$.

b) Elabore un programa de cómputo para generar 500 valores de la densidad $h(x)$.

c) Elabore un histograma (frecuencias relativas) de los datos obtenidos y en la misma gráfica dibuje la función $h(x)$.

47. *Distribución Cauchy estándar.* Aplique el método del cociente de uniformes para generar valores de una variable aleatoria con distribución Cauchy estándar, cuya función de densidad es

$$h(x) = \frac{1}{\pi\,(1 + x^2)}, \quad -\infty < x < \infty.$$

a) Determine analíticamente y muestre en un plano la región asociada $S \subseteq \mathbb{R}^2$. Compruebe que $|S| = 1/2$.

b) Elabore un programa de cómputo para generar 500 valores de la densidad $h(x)$.

c) Elabore un histograma (frecuencias relativas) de los datos obtenidos y en la misma gráfica dibuje la función $h(x)$.

48. *Distribución doble exponencial.* Aplique el método del cociente de uniformes para generar valores de una variable aleatoria con distribución doble exponencial dada por la expresión que aparece abajo, en donde $\lambda > 0$ es un parámetro.

$$h(x) = \frac{1}{2}\,\lambda\, e^{-\lambda|x|}, \quad -\infty < x < \infty.$$

a) Determine analíticamente y muestre en un plano la región asociada $S \subseteq \mathbb{R}^2$. Compruebe que $|S| = 1/2$.

b) Elabore un programa de cómputo para generar 500 valores de la densidad $h(x)$ para un valor λ de su elección.

c) Elabore un histograma (frecuencias relativas) de los datos obtenidos y en la misma gráfica dibuje la función $h(x)$.

49. *Distribución gama.* Siga los pasos que aparecen abajo para implementar el método del cociente de uniformes para producir valores de la distribución gama (α, γ). Recordemos que esta distribución tiene función de densidad

$$f(x) = \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}\, \lambda\, e^{-\lambda x}, \quad \text{para} \;\; x > 0.$$

a) Sea $\lambda > 0$. Demuestre que si $X \sim$ gama $(\alpha, 1)$, entonces $X/\lambda \sim$ gama (α, λ). Este resultado establece que es suficiente generar valores de la distribución gama $(\alpha, 1)$, cuya función de densidad es

$$f(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}\, e^{-x}, \quad \text{para} \;\; x > 0.$$

Escoja un valor particular de $\alpha > 1$. Se demostró antes la forma en la que el caso $0 < \alpha \leqslant 1$ se puede obtener a partir del caso $\alpha > 1$.

b) Considere la función $h(x) = x^{\alpha-1}\, e^{-x}$, para $x > 0$, y determine la expresión analítica que define a la región S para el valor escogido $\alpha > 1$.

c) Grafique en una computadora la región S. Aquí es necesario encontrar la gráfica de una función implícita.

d) Demuestre analíticamente, o por integración Monte Carlo (método de aceptación y rechazo), que el área de S es

$$|S| = \frac{1}{2} \int_{-\infty}^{\infty} h(x)\, dx = \frac{1}{2}\, \Gamma(\alpha).$$

e) Genere 500 valores de la función de densidad gama $(\alpha, 1)$ usando el método del cociente de uniformes con los elementos anteriores. Elabore un histograma de probabilidad con los datos obtenidos y en la misma gráfica dibuje la función de densidad gama $(\alpha, 1)$, para el valor $\alpha > 1$ escogido.

Generación de valores de variables aleatorias multivariadas

50. Usando el método de aceptación y rechazo, elabore un programa de cómputo para generar 200 valores (x, y) de la distribución especificada abajo. En un plano cartesiano dibuje el círculo unitario y marque con un punto cada uno de los datos obtenidos.

$$f(x,y) = \begin{cases} 1/\pi & \text{si } \ x^2+y^2 < 1, \\ 0 & \text{en otro caso.} \end{cases}$$

51. Grafique las siguientes funciones de densidad bivariadas y proporcione los detalles de cada una de ellas para generar valores de estas distribuciones usando el método de aceptación y rechazo.

$$a)\ f(x,y) = \begin{cases} 6x^2y & \text{si } \ 0<x<1,\ 0<y<1, \\ 0 & \text{en otro caso.} \end{cases}$$

$$b)\ f(x,y) = \begin{cases} 3x & \text{si } \ 0<y<x<1, \\ 0 & \text{en otro caso.} \end{cases}$$

52. *Método secuencial.* Sea $(X_1, \ldots, X_n)$ un vector aleatorio con función de densidad o de probabilidad $f(x_1, \ldots, x_n)$. Demuestre la identidad

$$f(x_1, \ldots, x_n) = f(x_1)\, f(x_2 \,|\, x_1) \cdots f(x_n \,|\, x_{n-1}, \ldots, x_1).$$

Esta fórmula sugiere un método secuencial para producir valores del vector $(X_1, \ldots, X_n)$ de la siguiente manera:

- Se genera un valor x_1 de X_1 con función de densidad $f(x_1)$.
- Dado $X_1 = x_1$, se genera un valor x_2 de X_2 con función de densidad $f(x_2 \,|\, x_1)$.
- Etcétera.

Por supuesto, el método requiere que se conozca una forma de generar valores de las funciones de densidad condicionales indicadas en la fórmula.

Miscelánea

53. Demuestre que $U \sim \text{unif}\,(0,1)$ si y sólo si $V = 1 - U \sim \text{unif}\,(0,1)$.

54. *Distribución geométrica a través de la distribución exponencial.* El siguiente procedimiento muestra la manera en la que pueden generarse valores de la distribución geométrica a partir de valores de la distribución exponencial. Se trata de una discretización de la distribución exponencial para obtener una distribución geométrica.

    a) Sea X una variable aleatoria con distribución $\text{geo}\,(p)$, $0 < p < 1$. Es decir, $P(X = x) = p\,(1-p)^x$ para $x = 0, 1, 2, \ldots$ Sea Y una variable aleatoria con distribución $\exp\,(\lambda)$. Calcule el valor de λ en función de p para que se cumpla que

    $$P(x \leqslant Y < x + 1) = P(X = x) \quad \text{para} \quad x = 0, 1, \ldots$$

    b) Defina $Z = \ln\,(1-U)/\ln(1-p)$, en donde $U \sim \text{unif}\,(0,1)$. Demuestre que $\lfloor Z \rfloor$ tiene distribución $\text{geo}\,(p)$, en donde $\lfloor x \rfloor$ indica la parte entera del número $x \geqslant 0$.

55. *Distribución log normal.* Aplique algún método de simulación para obtener $n = 200$ observaciones de una variable aleatoria log normal con media e^2 y varianza $e^4(e^2 - 1)$. Calcule la media y la varianza muestrales de estas observaciones y compare con los valores teóricos.

56. *Aplicación a finanzas.* Suponga que una acción financiera tiene un precio S_t al tiempo t. Dado que $S_0 = 10$, el precio de la acción S_1 puede subir a 12, con probabilidad $p = 0.48$, o caer hasta 8 con probabilidad $1 - p$. Sea $R = (S_1 - S_0)/S_0$ la tasa de retorno de la acción al final del periodo 1.

    a) Realice 1000 simulaciones del valor de S_1 para aproximar $E(R)$ y $\sqrt{\text{Var}(R)}$ usando la media y la varianza muestral de R.

    b) Suponga que se adquiere una cobertura financiera que otorga un beneficio de $C = \text{máx}\,\{S_1 - 10, 0\}$ al final del periodo. La persona que adquiere la cobertura debe pagar una prima P que teóricamente se calcula como $P = E(C)\,e^{-0.05}$. Usando las simulaciones del inciso anterior calcule una aproximación a la prima P.

57. Sea $m \geqslant 2$ un entero. Defina $W = (X + Y) \mod m$, en donde X es una variable aleatoria uniforme en los enteros $\{0, \ldots, m-1\}$ y Y es una variable aleatoria discreta independiente de X. Demuestre que W tiene distribución uniforme en los enteros $\{0, \ldots, m-1\}$.

58. Mediante algún método de simulación obtenga valores de una variable aleatoria X con la distribución abajo indicada. Obtenga los histogramas de probabilidad y compare con las funciones de densidad.

    a) Cauchy $(0, 1/10)$.

    b) unif $(-1, 2)$.

    c) N $(1, 4)$.

    d) $f(x) = \dfrac{3}{16}\sqrt{x}, \quad 0 < x < 4.$

    e) $P(X = x) = (7/10)^x(3/10), \quad x = 0, 1, \ldots$

    f) $f(x) = c\,x^2 e^{-x}, \quad 0 < x < 1, \quad$ en donde $c = 1/(2 - 5e^{-1})$.

    g)

x	0	1	2	3
$P(X = x)$	0.1	0.2	0.3	0.4

59. Utilice el método de la función inversa para generar 1000 valores de una variable aleatoria X con función de probabilidad como aparece abajo. En una misma gráfica compare el histograma de probabilidad de los datos generados y la función de probabilidad.

$$P(X = x) = \frac{1}{x(x+1)}, \quad x = 1, 2, \ldots$$

60. Utilice primero el método de la función inversa y después el método de aceptación y rechazo para generar 1000 valores de una variable aleatoria X con función de densidad como aparece abajo. ¿Cuál de los dos métodos tuvo menor varianza muestral?

$$f(x) = x^2 + \frac{2}{3}, \qquad 0 < x < 1.$$

61. Programe dos funciones que obtengan muestras aleatorias con el método que usted elija para las siguientes distribuciones. Obtenga muestras con distintos parámetros y compruebe que los histogramas de probabilidad se parecen a las funciones de densidad respectivas.

a) Distribución Rayleigh ($\sigma^2 > 0$ es un parámetro),

$$f(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad x > 0.$$

b) Distribución triangular ($a > 0$ es un parámetro),

$$f(x) = \frac{2}{a}\left(1 - \frac{x}{a}\right), \quad 0 < x < a.$$

62. Aplique el método del cociente de uniformes para generar valores de una variable aleatoria con la función de densidad que aparece abajo, en donde $n > 0$ es un parámetro. Muestre la región S asociada.

$$h(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad -\infty < x < \infty.$$

63. (G.S. Fishman [12]) Suponga que a, b y c son tres parámetros tales que $0 < a \leqslant b$ y $-\infty < c < \infty$. Considere la función de densidad trapezoidal dada por

$$f(x) = \begin{cases} (x-c)/(ab) & \text{si } \ c \leqslant x \leqslant a+c, \\ 1/b & \text{si } \ a+c \leqslant x \leqslant b+c, \\ (a+b+c-x)/(ab) & \text{si } \ b+c \leqslant x \leqslant a+b+c. \end{cases}$$

a) Usando el método de la función inversa, programe una función que obtenga muestras aleatorias de esta distribución con $a = 1$, $b = 2$ y $c = -1.5$.

b) Suponga que U_1 y U_2 son variables aleatorias independientes cada una con distribución unif $(0, 1)$. Demuestre que para a_1, a_2 y a_3 elegidos de forma adecuada, $Z := a_1U_1 + a_2U_2 + a_3$ sigue una distribución trapezoidal. Vuelva a generar una muestra usando la transformación Z y compare los histogramas de esta muestra y la obtenida en el inciso anterior.

64. (G.S. Fishman [12]) Sean U_1 y U_2 variables aleatorias independientes, ambas con distribución unif $(-1/2, 1/2)$ y sea U con distribución unif $(0, 1)$.

a) Suponga que $U_1^2 + U_2^2 \leqslant 1/4$,. Demuestre que $Y = U_1/\sqrt{U_1^2 + U_2^2}$ tiene la misma distribución que $\cos(2\pi U)$.

b) Suponga que $U_1^2 + U_2^2 \leqslant 1/4$,. Demuestre que $Z = U_2/\sqrt{U_1^2 + U_2^2}$ tiene la misma distribución que $\operatorname{sen}(2\pi U)$.

c) Suponga que $U_1^2 + U_2^2 \leqslant 1$. Calcule la distribución de $Y = U_1/U_2$.

65. Genere una muestra aleatoria numérica de 1000 valores de la función de densidad que se muestra abajo. Con el promedio de estos valores aproxime $E(X)$, el cual se calculó en el Ejercicio 21.

$$f(x) = \begin{cases} (3/7)\, e^{-x} + (8/7)\, e^{-2x} & \text{si } \ x > 0, \\ 0 & \text{si } \ x \leqslant 0. \end{cases}$$

66. Sea X una variable aleatoria con función de densidad $f(x)$ como aparece abajo. Aproxime $E(1/X)$.

$$f(x) = \begin{cases} |x|/10 & \text{si } \ -2 \leqslant x \leqslant 4, \\ 0 & \text{en otro caso.} \end{cases}$$

67. Suponga que los tiempos de llegada entre autobuses consecutivos se pueden modelar mediante variables aleatorias independientes e idénticamente distribuidas $T_1, T_2, \ldots$, cada una con distribución $\text{unif}(0, 1)$. Usando simulación y la Proposición 1.4 aproxime el valor de $P(T_1 + T_2 + \cdots + T_N \leqslant 3)$, en donde $N \sim \text{Poisson}(7)$ es una variable aleatoria independiente de $T_1, T_2, \ldots$

68. Repita el ejercicio anterior pero ahora suponiendo que los tiempos de llegada entre autobuses consecutivos son variables aleatorias independientes exponenciales con media 2.

69. Un examen de opción múltiple tiene 10 preguntas, y cada pregunta tiene 4 opciones de respuesta, pero sólo una respuesta es la correcta. Un estudiante selecciona aleatoriamente su respuesta en cada una de las preguntas. Usando simulación y la Proposición 1.4 aproxime la probabilidad de que el estudiante obtenga al menos 7 respuestas correctas.

70. *Aplicación a la industria.* La probabilidad de que una máquina de helados se descomponga en un día cualquiera es 0.03 y es independiente de las descomposturas en cualquier otro día. La máquina sólo puede descomponerse una vez por día. Usando simulación y la Proposición 1.4 aproxime la probabilidad de que la máquina se descomponga en menos de 5 ocasiones en un mes (30 días).

71. *Aplicación a seguros.* Suponga que una aseguradora calcula una prima para un seguro de autos considerando lo siguiente: En cualquier mes, puede ocurrir máximo un siniestro. La probabilidad de que ocurra un siniestro en cualquier mes es de 0.007 y existe independencia entre los siniestro de un mes con los siniestros que ocurren en otro mes. Usando simulación y la Proposición 1.4 aproxime la probabilidad de que ocurra más de un siniestro en un año.

72. *Aplicación a la industria.* Un hospital recibe anualmente la mitad de medicamentos de cierta compañía A, una cuarta parte de la compañía B y el resto de la compañía C. El control de calidad del hospital detecta que el 2 % de los empaques (cajas y frascos de pastillas) de medicamentos de la compañía A vienen con algún defecto. Para las compañías B y C, los porcentajes de empaques defectuosos son del 1 % y 4 %, respectivamente. Del gran total de medicamentos que han llegado se elije aleatoriamente un empaque y resulta defectuoso. Usando simulación y la Proposición 1.4 aproxime la probabilidad de que dicho empaque provenga de la compañía C.

73. *Aplicación a logística.* Un biólogo desea capturar 7 mariposas monarca vivas para realizar un estudio de su comportamiento. El biólogo captura al azar una mariposa a la vez, pero tiene una probabilidad de 2/3 de que la mariposa le sea útil debido a que deben tener cierto tamaño. Sólo conserva las mariposas útiles, el resto las libera por ser muy pequeñas. Usando simulación y la Proposición 1.4 aproxime la probabilidad de que el biólogo capture 8 mariposas muy pequeñas (no necesariamente consecutivas) antes de capturar las 7 mariposas que necesita.

74. *Aplicación a finanzas.* El precio de dólar en el día n se modela con una variable aleatoria X_n. Después de un estudio estadístico se determina

que $X_n \mid (X_{n-1} = x) \sim \mathrm{N}(x, 0.05)$. Si se tiene que $X_0 = 18$, usando simulación y la Proposición 1.4 aproxime la probabilidad de que $X_{10} > 19$.

75. *Aplicación a logística.* El tiempo de espera en años para que un conductor tenga un siniestro se modela con una variable aleatoria $T \sim \exp(1/5)$. Por otro lado, el tiempo de espera en años para que el mismo conductor deba realizar un cambio de batería a su auto se modela con una variable aleatoria $Y \sim \mathrm{unif}(2, 4)$. Usando simulación y la Proposición 1.4 aproxime la probabilidad de que el conductor tenga un siniestro antes de tener que hacer un cambio de batería.

76. *Aplicación a telecomunicaciones.* Una empresa de telecomunicaciones tiene en funcionamiento simultáneo 5 satélites para distribuir señal de internet en una ciudad. Cuando hay alguna falla en alguno de los satélites, el tiempo en minutos que pasa para que se resuelva la falla tiene distribución gama $(13, 7)$ y este tiempo es independiente de lo que suceda en los otros satélites. En cierto momento ocurre un fenómeno solar por el cual los 5 satélites dejan de funcionar al mismo tiempo. Usando simulación y la LGN aproxime el tiempo esperado que pasará antes de que se resuelva la falla en al menos un satélite.

77. *Aplicación a series de tiempo.* Para la temperatura promedio X_t en el mes $t = 0, 1, \ldots$ en una cierta población, se ha ajustado el siguiente modelo autoregresivo de orden 6,

$$X_t = 20 + 0.13X_{t-1} - 0.05X_{t-2} - 0.11X_{t-4} + 0.02X_{t-6} + Z_t, \quad t \geqslant 6,$$

en donde $Z_1, Z_2, \ldots$ son variables aleatorias independientes con idéntica distribución $\mathrm{N}(0, 1/10)$. Suponga que se conocen los valores iniciales $X_0, \ldots, X_5 = 18, 19, 20, 21, 24, 25$. Genere pronósticos para $X_7, X_8, \ldots, X_{12}$.

Capítulo 3

Integración Monte Carlo

Nuestro interés ahora es calcular valores aproximados de integrales en una y varias dimensiones a través de la generación de valores de variables aleatorias. Existen varios métodos para llevar a cabo estas aproximaciones. Éstos se denominan métodos de integración Monte Carlo y revisaremos dos de ellos en las siguientes secciones. Los métodos pueden aplicarse tanto a integrales unidimensionales como a integrales múltiples. Para integrales de una o dos dimensiones existen métodos numéricos deterministas que son más aconsejables. En cambio, cuando se tiene un problema complejo en donde es necesario calcular una integral de muchas dimensiones, la integración Monte Carlo es particularmente más conveniente.

3.1. Método clásico

Sea $g(x)$ una función integrable en el intervalo (a, b) y tal que satisface la condición $0 \leqslant g(x) \leqslant c$, para alguna constante $c > 0$. Supondremos que el intervalo (a, b) es acotado pues consideraremos una distribución uniforme sobre él. Nos interesa calcular la integral

$$\theta = \int_a^b g(x)\, dx. \tag{3.1}$$

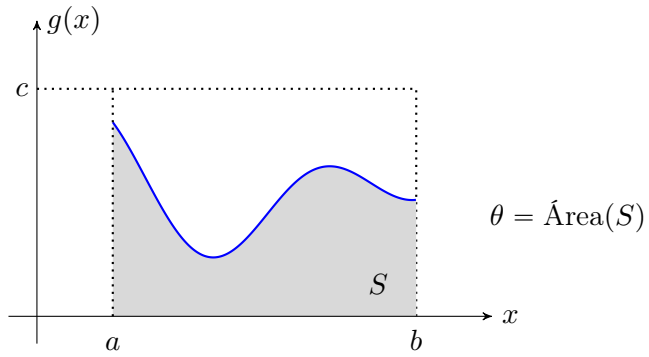


Figura 3.1: Región a estimar.

Sea (X, Y) un vector aleatorio con distribución unif $(a, b) \times (0, c)$. Esto implica que las coordenadas del vector son independientes y cada una tiene distribución uniforme en el intervalo respectivo. La probabilidad de que el punto (X, Y) se encuentre bajo la curva de $g(x)$ es el valor desconocido

$$\begin{aligned} p &= P(Y \leqslant g(X)) \\ &= \frac{\text{Área favorable}}{\text{Área total}} \\ &= \frac{1}{c(b-a)} \int_a^b g(x)\, dx \\ &= \frac{1}{c(b-a)}\, \theta. \end{aligned}$$

Es decir,

$$\theta = c(b-a)p. \tag{3.2}$$

Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una sucesión de n copias independientes del vector (X, Y). Esta es una muestra aleatoria y representa el experimento de tomar n puntos al azar dentro del rectángulo $(a, b) \times (0, c)$. Sea N el número de veces que se cumple la condición $Y_i \leqslant g(X_i)$, para $i = 1, 2, \ldots, n$. Claramente, $N \sim \text{bin}\,(n, p)$. Se propone como estimador para p a la variable aleatoria $\hat{p} = N/n$. Por lo tanto, el estimador para la integral θ es el siguiente.

Definición 3.1 *Sea $g(x)$ una función integrable en el intervalo acotado (a,b) y tal que $0 \leqslant g(x) \leqslant c$, para alguna constante $c > 0$. Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra aleatoria de la distribución* unif $(a,b) \times (0,c)$. *Sea N el número de veces que se cumple la condición $Y_i \leqslant g(X_i)$ para $i = 1, \ldots, n$. Se define el estimador Monte Carlo clásico para la integral (3.1) como*

$$\hat{\theta} := c(b-a)\,\frac{N}{n}. \tag{3.3}$$

Este estimador representa la fórmula para aproximar la integral θ por el método clásico de Monte Carlo: se generan al azar de manera uniforme n puntos dentro del rectángulo $(a,b) \times (0,c)$, se cuenta el número de puntos que quedan por debajo de la curva y se divide entre n, finalmente se multiplica por el área del rectángulo $(a,b) \times (0,c)$. El procedimiento se resume en el siguiente recuadro.

Integración Monte Carlo: método clásico

- Sea $g(x)$ una función integrable en el intervalo acotado (a,b).
- Suponga que $0 \leqslant g(x) \leqslant c$, para alguna constante $c > 0$.
- Sean $(x_1, y_1), \ldots, (x_n, y_n)$ valores de la distribución unif $(a,b) \times (0,c)$.
- Sea n_0 el número de valores tales que $y_i \leqslant g(x_i)$.

Entonces

$$\int_a^b g(x)\,dx \approx c(b-a)\,\frac{n_0}{n}.$$

Mediante los resultados que encontraremos más adelante, comprobaremos que es conveniente tomar el mínimo valor posible para la cota superior c de la función a integrar, se puede tomar $c = \sup\{g(x) : a < x < b\} < \infty$.

Observemos que $\hat{\theta}$ definida en (3.3) es una variable aleatoria con distribución binomial multiplicada por una constante. Esto permite encontrar algunas

de sus características numéricas con facilidad. A continuación calcularemos la media y la varianza. Cuando se desea enfatizar la dependencia de un estimador $\hat{\theta}$ del tamaño de la muestra n se le escribe también como $\hat{\theta}_n$.

Proposición 3.1 *El estimador $\hat{\theta}$ dado por (3.3) satisface las siguientes propiedades:*

1. *Es insesgado, es decir, $E(\hat{\theta}) = \theta$.*
2. *Su varianza es* $\mathrm{Var}(\hat{\theta}) = \dfrac{\theta}{n}\,[c(b-a) - \theta]$.
3. *Es consistente, es decir,* $\lim_{n\to\infty} \hat{\theta}_n = \theta$.

Demostración. Como N tiene distribución $\mathrm{bin}(n, p)$, se cumple que $E(N) = np$ y $\mathrm{Var}(N) = np(1-p)$. Por lo tanto,

1. Tenemos que

$$\begin{aligned} E(\hat{\theta}) &= c(b-a)\frac{1}{n}E(N) \\ &= c(b-a)\frac{1}{n}np \\ &= c(b-a)\,p \\ &= \theta. \end{aligned}$$

2. Se usan las propiedades de la varianza,

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(c(b-a)N/n) \\ &= \frac{c^2(b-a)^2}{n^2}\,np(1-p) \\ &= \frac{1}{n}\,c(b-a)\,[c(b-a)p - c(b-a)p^2] \\ &= \frac{\theta}{n}\,[c(b-a) - c(b-a)p] \\ &= \frac{\theta}{n}\,[c(b-a) - \theta]. \end{aligned}$$

3. Esto es consecuencia del inciso anterior pues cuando el tamaño n de la muestra crece a infinito, la varianza se hace cero. Por otro lado, una variable aleatoria con varianza cero es constante. En este caso, esa constante es $E(\hat{\theta})$, que es θ. Alternativamente, definiendo los eventos $A_i = (Y_i \leqslant g(X_i))$, para $i = 1, 2, \ldots, n$ y $A = (Y \leqslant g(X))$ el evento genérico, por la ley de los grandes números tenemos que

$$\begin{aligned} \lim_{n\to\infty} \hat{\theta}_n &= \lim_{n\to\infty} c(b-a)\,\frac{1}{n}\sum_{i=1}^{n} 1_{A_i} \\ &= c(b-a)\,E(1_A) \\ &= c(b-a)\,P((X,Y)\in S) \\ &= c(b-a)\,p \\ &= \theta. \end{aligned}$$

∎

Se observa que la varianza decrece conforme el valor de la cota superior c es más pequeña, de modo que es conveniente tomar c con el valor más pequeño posible.

Tamaño de muestra

El siguiente resultado permite calcular el tamaño de muestra n para que el valor del estimador (3.3) satisfaga cierto nivel de confianza.

Proposición 3.2 *Sean $\epsilon > 0$ y $0 < \alpha < 1$ dos cantidades dadas. Cuando el tamaño de muestra n es tal que $n \geqslant c^2(b-a)^2/(4\alpha\epsilon^2)$, se cumple*

$$P(|\hat{\theta} - \theta| < \epsilon) \geqslant 1 - \alpha.$$

Demostración. Consideremos la función $h(\theta) = \theta[c(b-a) - \theta]$, para $-\infty < \theta < \infty$. Se puede comprobar que $h(\theta) \leqslant c^2(b-a)^2/4$, para $-\infty < \theta < \infty$. Usando este resultado, por la desigualdad de Chebyshev y después

usando la expresión de la varianza que aparece en la Proposición 3.1,

$$\begin{aligned} P(|\hat{\theta} - \theta| \geqslant \epsilon) &\leqslant \frac{1}{\epsilon^2}\operatorname{Var}(\hat{\theta}) \\ &= \frac{1}{n\epsilon^2}\,\theta(c(b-a)-\theta) \\ &\leqslant \frac{1}{4n\epsilon^2}\,c^2(b-a)^2 \\ &\leqslant \alpha. \end{aligned}$$

La última desigualdad se cumple cuando n es suficientemente grande como indica la hipótesis. Por lo tanto, tomando complementos, tenemos que

$$P(|\hat{\theta} - \theta| < \epsilon) \geqslant 1 - \alpha.$$

■

El parámetro ϵ determina la cercanía entre $\hat{\theta}$ y θ. El número $1-\alpha$ representa el nivel de confianza. Nuevamente, se observa que el tamaño de muestra es más pequeño cuando se toman valores de c menores.

Intervalo de confianza

Explicaremos a continuación dos formas de encontrar un intervalo de confianza para el valor desconocido θ definido en (3.1).

- Por la Proposición 3.2, cuando el tamaño de muestra n es tal que

$$n \geqslant \frac{c^2(b-a)^2}{2\alpha\epsilon^2}, \tag{3.4}$$

  tenemos el intervalo de confianza

$$P(\hat{\theta} - \epsilon < \theta < \hat{\theta} + \epsilon) \geqslant 1 - \alpha, \tag{3.5}$$

  en donde $\epsilon > 0$ y $0 < \alpha < 1$ son dos cantidades dadas. De la ecuación (3.4) se obtiene $\epsilon^2 = c^2(b-a)^2/(2\alpha n)$, de modo que el intervalo de confianza (3.5) toma la forma

$$P\left(\hat{\theta} - \frac{1}{\sqrt{\alpha}}\frac{c(b-a)}{\sqrt{2n}} < \theta < \hat{\theta} + \frac{1}{\sqrt{\alpha}}\frac{c(b-a)}{\sqrt{2n}}\right) \geqslant 1 - \alpha. \tag{3.6}$$

Este es un primer intervalo de confianza para θ. Observemos que la longitud del intervalo es menor cuando se toman valores más pequeños para la cota superior c de la función a integrar. A continuación encontraremos un mejor intervalo en el sentido de que su longitud es menor que la del intervalo (3.6).

- La variable aleatoria N que cuenta el número de veces que el punto al azar (X, Y) queda bajo la curva de la función $g(x)$ se puede escribir como

$$N = \sum_{i=1}^{n} 1_{A_i},$$

en donde los sumandos son las funciones indicadoras de los eventos $A_i = (Y_i \leqslant g(X_i)),\ i = 1, \ldots, n$. Entonces el estimador (3.3) para θ se puede escribir como

$$\hat{\theta} = c(b-a)\,\frac{1}{n}\sum_{i=1}^{n} 1_{A_i}.$$

Por el teorema central del límite, de manera aproximada,

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim \text{N}(0,1).$$

Por lo tanto, se puede encontrar un valor $z_{\alpha/2} > 0$ tal que

$$P\left(-z_{\alpha/2} < (\hat{\theta} - \theta)/\sqrt{\text{Var}(\hat{\theta})} < z_{\alpha/2}\right) = 1 - \alpha,$$

con $0 < \alpha < 1$. Es decir,

$$P\left(\ \hat{\theta} - z_{\alpha/2} \cdot \sqrt{\text{Var}(\hat{\theta})} < \theta < \hat{\theta} + z_{\alpha/2} \cdot \sqrt{\text{Var}(\hat{\theta})}\ \right) = 1 - \alpha.$$

En palabras, con una confianza del $(1-\alpha)100\,\%$ el intervalo aleatorio simétrico

$$\left(\ \hat{\theta} - z_{\alpha/2} \cdot \sqrt{\text{Var}(\hat{\theta})}, \hat{\theta} + z_{\alpha/2} \cdot \sqrt{\text{Var}(\hat{\theta})}\ \right)$$

contiene al valor de θ. Sin embargo, el término

$$\mathrm{Var}(\hat{\theta}) = (\theta/n)(c(b-a) - \theta)$$

es desconocido. Aquí se puede usar nuevamente la estimación $h(\theta) = \theta(c(b-a)-\theta) \leqslant c^2(b-a)^2/2$ y obtener el intervalo aleatorio ampliado

$$P\left(\hat{\theta} - z_{\alpha/2}\frac{c(b-a)}{\sqrt{2n}} < \theta < \hat{\theta} + z_{\alpha/2}\frac{c(b-a)}{\sqrt{2n}} \right) \approx 1-\alpha. \qquad (3.7)$$

Es claro que los intervalos de confianza (3.6) y (3.7) difieren sólo debido a que el primero tiene el factor $1/\sqrt{\alpha}$ y el segundo tiene el factor $z_{\alpha/2}$. Puede comprobarse, por lo menos numéricamente, que

$$z_{\alpha/2} < \frac{1}{\sqrt{\alpha}}, \quad \text{para} \;\; 0 < \alpha < 1,$$

de modo que el intervalo (3.7) tiene longitud menor.

Caso multidimensional

El caso de una dimensión estudiado en la sección anterior se puede extender sin mucha dificultad al caso de dimensiones mayores. Por simplicidad en la escritura consideraremos sólo el caso bidimensional.

Sea $g(x, y)$ una función integrable en el rectángulo acotado $(a_1, b_1) \times (a_2, b_2)$ y tal que $0 \leqslant g(x, y) \leqslant c$, para alguna constante $c > 0$. La integral a calcular es

$$\theta = \int_{a_1}^{b_1}\int_{a_2}^{b_2} g(x, y)\, dy\, dx. \qquad (3.8)$$

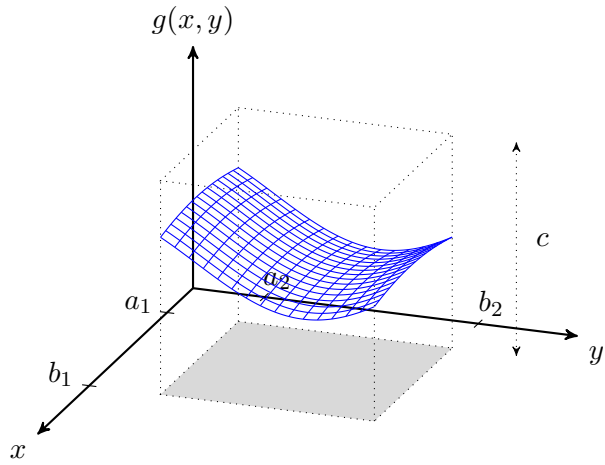


Figura 3.2: Volumen a estimar.

Sea (X, Y, Z) un vector aleatorio con distribución unif $(a_1, b_1) \times (a_2, b_2) \times (0, c)$. La probabilidad de que el punto al azar (X, Y, Z) se encuentre bajo la superficie dada por $g(x, y)$ es el valor desconocido

$$\begin{aligned} p &= P(Z \leqslant G(X,Y)) \\ &= \frac{\text{Volumen favorable}}{\text{Volumen total}} \\ &= \frac{1}{c(b_1-a_1)(b_2-a_2)} \int_{a_1}^{b_1}\int_{a_2}^{b_2} g(x,y)\,dy\,dx \\ &= \frac{1}{c(b_1-a_1)(b_2-a_2)}\,\theta. \end{aligned}$$

Es decir, $\theta = c(b_1 - a_1)(b_2 - a_2)\,p$, en donde p se estima mediante el cociente $\hat{p} = n_0/n$ con n un cierto número de valores del vector (X, Y, Z) y n_0 es el total de aquellos valores (x, y, z) que satisfacen la condición $z \leqslant g(x, y)$. El procedimiento se resume a continuación.

Integración Monte Carlo: método clásico (dimensión 2)

- Sea $g(x, y)$ una función integrable en el rectángulo acotado $(a_1, b_1) \times (a_2, b_2)$.
- Suponga que $0 \leqslant g(x, y) \leqslant c$ para alguna constante $c > 0$.
- Sean $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ puntos aleatorios de la distribución unif $(a_1, b_1) \times (a_2, b_2) \times (0, c)$.
- Sea n_0 el número de puntos aleatorios tales que $z_i \leqslant g(x_i, y_i)$.

Entonces

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} g(x, y)\, dy\, dx \approx c(b_1 - a_1)(b_2 - a_2)\, \frac{n_0}{n}.$$

Se puede definir el estimador para la integral doble dada por (3.8) de la misma manera que se hizo en el caso de la integral de dimensión 1. Reproducimos esto a continuación. Sea $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ una muestra aleatoria de la distribución unif $(a_1, b_1) \times (a_2, b_2) \times (0, c)$ y sea N el número de veces que se cumple la condición $Z_i \leqslant g(X_i, Y_i)$, para $i = 1, \ldots, n$. Entonces N tiene distribución bin (n, p), en donde p es desconocido. Se propone nuevamente $\hat{p} = N/n$ y se genera así un estimador para la integral dada por (3.8), cuyas propiedades son similares al caso unidimensional,

$$\hat{\theta} = c(b_1 - a_1)(b_2 - a_2)\, \frac{N}{n}.$$

3.2. Método de la media muestral

Como antes, si $g(x)$ es una función integrable, se busca encontrar el valor de

$$\theta = \int_a^b g(x)\, dx, \tag{3.9}$$

en donde ahora consideraremos que la función $g(x)$ no necesariamente es positiva ni acotada superiormente, como se supuso en el caso del método clásico de integración Monte Carlo. Más aún, el intervalo de integración (a, b) no necesariamente será acotado, puede ser el intervalo completo $(-\infty, \infty)$, por ejemplo.

Sea X cualquier variable aleatoria con una función de densidad $f(x)$, cuyo soporte es el intervalo (a, b). Entonces la integral θ dada por (3.9) se puede escribir como el siguiente valor esperado

$$\theta = \int_a^b \frac{g(x)}{f(x)} f(x)\, dx = E\left(\frac{g(X)}{f(X)}\right).$$

Observe que se puede considerar que el cociente $g(X)/f(X)$ es una variable aleatoria y que un estimador para su media es la media muestral, esto es, si $X_1, \ldots, X_n$ es una muestra aleatoria de la función de densidad $f(x)$, entonces se puede definir el siguiente estimador.

Definición 3.2 *Sea $g(x)$ una función integrable sobre el intervalo (a, b). La función y el intervalo no son necesariamente acotados. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una densidad $f(x)$ con soporte (a, b). Se define el estimador Monte Carlo de la media muestral para la integral (3.9) como*

$$\hat{\theta} := \frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}. \tag{3.10}$$

Este es el método de integración Monte Carlo usando la media muestral. Observe que el estimador se puede construir utilizando cualquier función de densidad $f(x)$ que tenga soporte el intervalo (a, b). El procedimiento se resume en el siguiente recuadro.

Integración Monte Carlo: método de la media muestral

- Sea $g(x)$ una función a integrar en el intervalo (a, b).
- Sea $f(x)$ cualquier función de densidad con soporte (a, b).
- Sean $x_1, \ldots, x_n$ valores de la distribución $f(x)$.

Entonces

$$\int_a^b g(x)\, dx \approx \frac{1}{n} \sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)}.$$

Dado que la media muestral es un estimador insesgado para la media de una distribución, tenemos que el estimador (3.10) es insesgado. En efecto,

$$E\left(\hat{\theta}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}\right) = \frac{1}{n}\sum_{i=1}^{n}E\left(\frac{g(X_i)}{f(X_i)}\right) = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta.$$

Otras características numéricas del estimador (3.10) no son fáciles de encontrar pues su cálculo involucra la esperanza de una función de una variable aleatoria.

Ejemplo 3.1 *Supongamos que se desea integrar una función $g(x)$ en el intervalo $(0,\infty)$. Puede usarse la función de densidad exponencial $f(x) = \lambda e^{-\lambda x}$, cuyo soporte es justamente ese intervalo. Por simplicidad tomaremos $\lambda = 1$. Entonces el método de la media muestral establece que*

$$\int_0^\infty g(x)\,dx = \int_0^\infty \frac{g(x)}{f(x)}f(x)\,dx = E\left(\frac{g(X)}{f(X)}\right) = E(e^X g(X)),$$

en donde el término $e^X g(X)$ es una nueva variable aleatoria. Si $x_1,\ldots,x_n$ son valores de la distribución exponencial indicada, entonces

$$\int_0^\infty g(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n} e^{x_i}g(x_i).$$

■

Ejemplo 3.2 *Consideremos ahora que tenemos una función $g(x)$ definida en el intervalo $(-\infty,\infty)$, que es integrable y cuya integral se desea aproximar por el método de la media muestral usando la distribución* $\mathrm{N}(0,1)$*. Si $\phi(x)$ denota la función de densidad* $\mathrm{N}(0,1)$*, entonces*

$$\int_{-\infty}^\infty g(x)\,dx = \int_0^\infty \frac{g(x)}{\phi(x)}\phi(x)\,dx = E\left(\frac{g(X)}{\phi(X)}\right) = \sqrt{2\pi}\,E(e^{X^2/2}g(X)),$$

en donde $X \sim \mathrm{N}(0,1)$. Si $x_1,\ldots,x_n$ son valores de la distribución normal estándar, entonces

$$\int_{-\infty}^\infty g(x)\,dx \approx \frac{\sqrt{2\pi}}{n}\sum_{i=1}^{n} e^{x_i^2/2}g(x_i).$$

■

Caso uniforme

Un caso sencillo y conveniente se obtiene cuando se toma a $f(x)$ como la función de densidad unif (a, b), suponiendo que este intervalo es acotado. En esta situación, la esperanza a calcular es

$$\theta = E(g(X)/f(X)) = (b-a)\, E(g(X)). \tag{3.11}$$

Así, tomando una muestra aleatoria $X_1, \ldots, X_n$ de la distribución unif (a, b), tenemos el estimador

$$\hat{\theta}_u := (b-a)\, \frac{1}{n} \sum_{i=1}^{n} g(X_i),$$

que sabemos que es insesgado. Usando (3.11), su varianza se puede escribir de la siguiente forma

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_u) &= \mathrm{Var}\left((b-a)\, \frac{1}{n} \sum_{i=1}^{n} g(X_i) \right) \\
&= (b-a)^2\, \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(g(X_i)) \\
&= (b-a)^2\, \frac{1}{n} \left[E(g^2(X)) - E^2(g(X)) \right] \\
&= (b-a)^2\, \frac{1}{n} \left[\frac{1}{b-a} \int_a^b g^2(x)\, dx - \frac{\theta^2}{(b-a)^2} \right] \\
&= \frac{1}{n} \left[(b-a) \int_a^b g^2(x)\, dx - \theta^2 \right].
\end{aligned}$$

En esta expresión no conocemos el valor de θ y posiblemente tampoco el valor de la integral de $g^2(x)$.

Caso uniforme: comparación de estimadores

En el caso cuando la función a integrar $g(x)$ satisface la condición $0 \leqslant g(x) \leqslant c$, para alguna constante $c > 0$, y el intervalo (a, b) es acotado, podemos

aplicar tanto el método clásico como el método de la media muestral en su versión uniforme para aproximar la integral de $g(x)$. Los correspondientes estimadores son:

$$\hat{\theta}_1 = c(b-a)\,\frac{N}{n} \qquad \text{Método clásico}$$

$$\hat{\theta}_2 = (b-a)\,\frac{1}{n}\sum_{i=1}^{n} g(X_i) \qquad \begin{array}{l}\text{Método de la media muestral con}\\ f(x) \text{ la función de densidad unif } (a,b)\end{array}$$

Ambos estimadores son insesgados pero nos gustaría comparar sus varianzas. El resultado es el siguiente.

Proposición 3.3 *Para los estimadores θ_1 y θ_2, se cumple*

$$\mathrm{Var}(\hat{\theta}_2) \leqslant \mathrm{Var}(\hat{\theta}_1).$$

Demostración. Por las hipótesis y los cálculos anteriores,

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_2) &= \frac{1}{n}\left[\,(b-a)\int_a^b g(x)\cdot g(x)\,dx - \theta^2\,\right]\\
&\leqslant \frac{1}{n}\left[\,(b-a)\,c\underbrace{\int_a^b g(x)\,dx}_{=\theta} - \theta^2\,\right]\\
&= \frac{\theta}{n}\left[\,c(b-a)-\theta\,\right]\\
&= \mathrm{Var}(\hat{\theta}_1).
\end{aligned}$$

∎

Así, en el sentido de varianza menor, el estimador $\hat{\theta}_2$ es mejor que el estimador $\hat{\theta}_1$. Recordemos que para el estimador $\hat{\theta}_2$ se tomó la distribución uniforme. Surge entonces la pregunta: ¿existe una función de densidad $f(x)$ que produzca un estimador para la integral θ que tenga varianza mínima? Trataremos este problema en el siguiente capítulo, el cual está dedicado

a estudiar algunos procedimientos para la reducción de la varianza de un estimador.

Caso multidimensional

El método de la media muestral en una dimensión puede extenderse fácilmente al caso de dimensiones mayores. En el siguiente recuadro se explica el procedimiento en el caso bidimensional.

Integración Monte Carlo:
método de la media muestral (dimensión 2)

- Sea $g(x, y)$ la función a integrar en el rectángulo $(a_1, b_1) \times (a_2, b_2)$.
- Sea $f(x, y)$ una función de densidad con soporte $(a_1, b_1) \times (a_2, b_2)$.
- Sean $(x_1, y_1), \ldots, (x_n, y_n)$ puntos al azar de acuerdo a la función de densidad $f(x, y)$.

Entonces

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} g(x, y)\, dy\, dx \approx \frac{1}{n} \sum_{i=1}^{n} \frac{g(x_i, y_i)}{f(x_i, y_i)}.$$

En particular, cuando el rectángulo $(a_1, b_1) \times (a_2, b_2)$ es acotado y se toma a $f(x, y)$ como la función de densidad uniforme, se tiene la aproximación

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} g(x, y)\, dy\, dx \approx (b_1 - a_1)(b_2 - a_2)\, \frac{1}{n} \sum_{i=1}^{n} g(x_i, y_i).$$

3.3. Ejercicios

Integración Monte Carlo: método clásico

78. Elabore un programa de cómputo que utilice el método clásico de Monte Carlo para encontrar un valor aproximado de la integral

$$\theta = \int_0^1 \sqrt{1 - x^2}\, dx.$$

Como sabemos, el valor exacto de la integral es $\pi/4$ pues se trata del área de un cuarto del disco unitario.

a) Comparando directamente los valores $4\hat{\theta}$ y π, encuentre el tamaño de muestra empírico n tal que estos dos valores coincidan en los primeros 3 dígitos decimales.

b) Suponiendo que no se conoce el valor de π, encuentre el tamaño de muestra n tal que $4\hat{\theta}$ y π coincidan, por lo menos, en los primeros 3 dígitos decimales con una confianza del 95 %.

79. Sea $\phi(x)$ la función de densidad $\mathrm{N}(0,1)$. El cuantil al 95 % de esta distribución es $c_p = 1.65$, es decir, el valor c_p es tal que

$$\int_{-\infty}^{c_p} \phi(x)\, dx = 0.95\,.$$

Elabore un programa de cómputo que use el método clásico de Monte Carlo para corroborar este resultado. Considere que el intervalo sobre el que se lleva a cabo la integración es el intervalo acotado $(-4, 1.65)$.

80. Escriba el algoritmo para llevar a cabo la integración Monte Carlo en su procedimiento clásico cuando la función a integrar es:

a) $g(x) = c$ (constante), para $a < x < b$.

b) $g(x) = 2x$ para $0 < x < 1$, cuando se toma $c = 1$.

81. Elabore un programa de cómputo que utilice el método clásico de integración Monte Carlo para calcular el volumen de una esfera de radio 1. Para corroborar su resultado, recuerde que el volumen de una esfera de radio r es $(4/3)\pi r^3$.

Integración Monte Carlo: método de la media muestral

82. Elabore un programa de cómputo que utilice el método de la media muestral de Monte Carlo para encontrar un valor aproximado de la integral θ que aparece abajo. Utilice una función de densidad $f(x)$ de su elección.

$$\theta = \int_0^1 \exp\left\{-x^2/2\right\}\, dx.$$

83. Elabore un programa de cómputo que utilice el método de la media muestral de Monte Carlo para encontrar un valor aproximado de la integral θ que aparece abajo. Utilice una función de densidad $f(x, y)$ de su elección.

$$\theta = \int_0^1 \int_0^1 \exp\left\{-(x+y)^2/2\right\}\, dx\, dy.$$

Miscelánea

84. Aproxime el valor de las siguientes integrales usando algún método de simulación:

$a)$ $\displaystyle\int_0^1 (1-x^2)^{3/2} dx.$ $\qquad$ $c)$ $\displaystyle\int_{-\infty}^{\infty} e^{-x^2} dx.$

$b)$ $\displaystyle\int_0^2 x^{3/2}(4-x)^{1/2} dx.$ $\qquad$ $d)$ $\displaystyle\int_0^1 \int_0^1 e^{(x+y)^2} dx dy.$

85. Exprese como integral la siguiente covarianza relacionada a una variable aleatoria $U \sim \text{unif}\,(0, 1)$,

$$\text{Cov}(U, e^U).$$

Resuelva analíticamente la integral y después realice una aproximación utilizando simulación.

86. Usando simulación aproxime el área de la región

$$\left\{(x, y) : -1 < x < 1,\ y > 0,\ \sqrt{1-2x^2} < y < \sqrt{1-2x^4}\right\}.$$

87. Encuentre un intervalo con 95 % de confianza para

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(x^2 + (y-1)^2 - \frac{x(y-1)}{10}\right)\right\} dx dy.$$

88. Utilice simulación para estimar cada una de las siguientes integrales:

$a)$ $\displaystyle\int_0^{\infty} x^2 \text{sen}(\pi x) e^{-\frac{x}{2}} dx.$

b) $\displaystyle\int_{-\infty}^{\infty}\int_{0}^{2}\int_{0}^{\infty} y\cos(\pi(x+y+z))e^{-x^2}e^{-z}dz\,dy\,dx.$

c) $\displaystyle\int_{-\infty}^{\infty}\int_{0}^{2}\int_{0}^{\infty} y\cos(\pi y)e^{-x^2y}e^{-yz}\,dz\,dy\,dx.$

89. *Aplicación en seguros.* Suponga que se tienen 3 tipos de riesgos en una aseguradora. Cada riesgo puede provocar una reclamación cuya cantidad en miles de pesos se modela con una variable aleatoria X_i, para $i = 1, 2, 3$. Los tres riesgos se consideran independientes y en cada caso $X_i \sim \text{unif}\,(0, 15)$. La aseguradora desea calcular el valor x tal que

$$P(X_1 + X_2 + X_3 \leqslant x) = 0.95\,.$$

Exprese la probabilidad anterior como la integral triple que corresponde. Aproxime el valor de x realizando simulaciones de X_1, X_2, X_3 y probando distintos valores de x.

90. *Aplicación en física.* Se tiene el siguiente modelo para el comportamiento de una partícula al tiempo t:

$$X_t = 3\sqrt{t} + B_t, \quad t > 0,$$

en donde B_t es una variable aleatoria con distribución $\text{N}\,(0.2t, 0.32t)$. Exprese la probabilidad siguiente como una integral y después aproxime usando simulación.

$$P(X_{10} > 12).$$

91. *Aplicación en seguros.* Suponga que una aseguradora debe pagar una cantidad inicial X cuando un asegurado tiene un choque en su auto. Después de una semana se realiza un ajuste y en total se paga $T = X + Y$, donde Y es un pago adicional por gastos administrativos. Suponga que $X \sim \text{unif}\,(2, 8)$ y que $Y \sim \text{gama}\,(2, 2)$. Además, suponga que X y Y son variables aleatorias independientes.

   a) Calcule una aproximación de $P(T > 5)$ solamente simulando parejas (X, Y) y contando en cuántas ocasiones se cumplió que $T > 5$, vea Proposición 1.4.

b) Calcule una aproximación de $P(T > 5)$ planteando la integral doble que resulta para hacer el cálculo exacto y posteriormente estimando tal integral.

c) Compare las varianzas muestrales de los valores generados para realizar las aproximaciones.

92. *Aplicación en teoría de riesgo.* Dentro de la teoría del riesgo se tiene el siguiente modelo para el capital C_t de una compañía aseguradora al tiempo $t \geqslant 0$,

$$C_t = u + ct - \sum_{j=0}^{N_t} Y_j, \quad t \geqslant 0,$$

en donde $u \geqslant 0$ y $c > 0$ son constantes, $N_t \sim \text{Poisson}\,(\lambda t)$ con $\lambda > 0$ y $Y_1, Y_2, \ldots$ son variables aleatorias independientes entre ellas e independientes de N_t. Suponga que $Y_0 := 0$ y que $Y_j \sim \exp(1)$ para $j \geqslant 1$. Suponga también que $u = 1$, $c = 3$, $\lambda = 2$. Usando simulación aproxime:

a) $P(C_5 < 0)$.

b) $P(C_5 < 0 \,|\, Y_1 < 1)$.

93. *Aplicación a ingeniería.* Suponga que se necesita calcular el volumen de un cerro sobre un terreno de forma irregular. El cerro se modela con la función $f(x, y) = 157 \exp\left\{-0.4x^2 - 0.35y^2\right\}$ y el terreno estaría delimitado por las curvas $y = x-4$ y $y^2 = 3x+7$. Exprese este volumen como una integral doble y después aproxime usando simulación.

Capítulo 4

Métodos de reducción de varianza

En este capítulo revisaremos algunos procedimientos que permiten reducir la varianza de un estimador construido a partir de valores de una variable o vector aleatorio.

4.1. Muestreo por importancia

Supongamos nuevamente que el problema es calcular una integral de la forma

$$\theta = \int_a^b g(x)\,dx.$$

Usando el método de la media muestral, si X es una variable aleatoria con función de función de densidad $f(x)$, cuyo soporte contiene al intervalo (a, b), entonces la integral θ se puede expresar como la siguiente esperanza

$$\theta = \int_a^b \frac{g(x)}{f(x)}\, f(x)\,dx = E\left(\frac{g(X)}{f(X)}\right). \tag{4.1}$$

El método Monte Carlo de la media muestral establece que si $X_1, \ldots, X_n$ es una muestra aleatoria de la función de densidad $f(x)$, entonces se propone como estimador para θ a la variable aleatoria

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}. \tag{4.2}$$

Usando (4.1), es fácil verificar que $\hat{\theta}$ es un estimador insesgado para el valor desconocido θ. Esta afirmación es válida para cualquier función de densidad $f(x)$ que se utilice. El objetivo del muestreo por importancia es utilizar la función de densidad $f(x)$ tal que los puntos que se generen de esta función tengan mayor o menor frecuencia según lo indique la función a integrar $g(x)$. Esta función de densidad existe y es aquella que hace que la varianza del estimador $\hat{\theta}$ sea mínima. Demostraremos esto a continuación.

Teorema 4.1 *La varianza del estimador $\hat{\theta}$ definido mediante el método de la media muestral (4.2) satisface*

$$\mathrm{Var}(\hat{\theta}) \geqslant \frac{1}{n}\left[\left(\int_a^b |g(x)|\, dx \right)^2 - \theta^2 \right].$$

Además, el valor mínimo se alcanza cuando se toma

$$f(x) = \frac{|g(x)|}{\displaystyle\int_a^b |g(x)|\, dx}, \quad a < x < b. \tag{4.3}$$

Demostración. Recordemos que la desigualdad de Cauchy-Schwarz para integrales establece que

$$\left(\int_a^b f(x) \cdot g(x)\, dx \right)^2 \leqslant \left(\int_a^b f^2(x)\, dx \right) \cdot \left(\int_a^b g^2(x)\, dx \right),$$

con igualdad $\Leftrightarrow f(x)/g(x)$ es constante. Por lo tanto, usando (4.2),

$$\begin{aligned} \mathrm{Var}(\hat{\theta}) &= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)} \right) \\ &= \frac{1}{n} \mathrm{Var}\left(\frac{g(X)}{f(X)} \right). \end{aligned}$$

La varianza del cociente g/f puede expresarse como

$$\begin{aligned}
\text{Var}\left(\frac{g(X)}{f(X)}\right) &= E\left(\frac{g^2(X)}{f^2(X)}\right) - E^2\left(\frac{g(X)}{f(X)}\right) \\
&= \int_a^b \frac{g^2(x)}{f^2(x)} \cdot f(x)\,dx - \theta^2 \qquad (4.4) \\
&= \left(\int_a^b \left(\frac{|g(x)|}{f^{1/2}(x)}\right)^2 dx\right) \cdot \underbrace{\int_a^b (f^{1/2}(x))^2\,dx}_{=1} - \theta^2 \\
&\geqslant \left(\int_a^b |g(x)|\,dx\right)^2 - \theta^2.
\end{aligned}$$

Cuando se toma la función de densidad $f(x) = |g(x)|/\int_a^b |g(x)|\,dx$ en la expresión (4.4), la varianza toma el valor

$$\begin{aligned}
\text{Var}(\hat{\theta}) &= \frac{1}{n}\int_a^b \frac{g^2(x)}{|g(x)|/\int_a^b |g(x)|\,dx}\,dx - \frac{\theta^2}{n} \\
&= \frac{1}{n}\left(\int_a^b |g(x)|\,dx\right)^2 - \frac{\theta^2}{n}.
\end{aligned}$$

■

De esta manera, en el sentido de varianza mínima, la mejor función de densidad $f(x)$ que uno puede tomar es la función definida como el valor absoluto de $g(x)$ dividido entre la integral del valor absoluto. Esto es una operación sobre $g(x)$ que la convierte en una función de función de densidad. Para que esto sea posible se necesita la hipótesis adicional de que $|g(x)|$ sea integrable en el intervalo (a, b). Por otro lado, el procedimiento de aproximación requiere que se conozca una manera de generar valores de la distribución $f(x)$.

Los valores $x_1, \ldots, x_n$ que se generen de esta función de densidad óptima $f(x)$ tenderán a concentrarse en las regiones donde $|g(x)|$ toma valores grandes, dándole mayor importancia a esas regiones. En cambio, la función otorga menor importancia a las regiones del intervalo (a, b) en donde $|g(x)|$ toma, comparativamente, valores pequeños.

Debe observarse que encontrar la función de densidad óptima $f(x)$ para aproximar la integral θ requiere conocer el valor de la integral

$$\theta = \int_a^b |g(x)|\, dx,$$

lo cual es similar al problema original de encontrar θ. De hecho, el problema es exactamente el mismo en el caso cuando $g(x) \geqslant 0$. Sin embargo, el teorema anterior sugiere tomar como $f(x)$ a aquella función de densidad que tenga un comportamiento parecido a $|g(x)|$ y de la cual conozcamos una manera de generar valores.

A continuación veremos una manera de encontrar una función, aproximada a la forma (4.3), que es factible de utilizarse.

Aproximación discreta

Como una aproximación a la función de densidad óptima $f(x)$ dada en (4.3), se puede dividir el intervalo (a, b) en m partes:

$$a < u_1 < u_2 < \cdots < u_{m-1} < b,$$

y definir la función constante por pedazos

$$\bar{g}(x) = \begin{cases} g_1 & \text{si } a < x < u_1, \\ g_2 & \text{si } u_1 < x < u_2, \\ \vdots & \vdots \\ g_m & \text{si } u_{m-1} < x < b, \end{cases}$$

en donde $g_1, g_2, \ldots g_m$ son ciertos valores de la función $g(x)$ en cada uno de los m subintervalos, respectivamente. Cada valor constante g_i puede ser cualquiera de los valores que toma la función $g(x)$ en el intervalo (u_{i-1}, u_i), con $u_0 = a$ y $u_m = b$.

Tomando valor absoluto de $\bar{g}(x)$ y dividiendo entre la suma de estos valores absolutos, la función $\bar{g}(x)$ se puede transformar en la función de probabilidad

$f_d(x)$ de una variable aleatoria discreta y de la cual sabemos una manera de obtener sus valores. Si $x_1, \ldots, x_n$ son valores de esta variable aleatoria discreta, entonces la aproximación es

$$\int_a^b g(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n}\frac{g(x_i)}{f_d(x_i)}.$$

Caso multidimensional

El análisis llevado a cabo en esta sección permanece válido cuando la función a integrar es función de varias variables. Por ejemplo, en el caso de 2 variables tenemos la integral

$$\theta = \int_{a_1}^{b_1}\int_{a_2}^{b_2} g(x,y)\,dy\;dx.$$

Como antes, cuando $f(x,y)$ es una función de densidad bivariada, el estimador por el método de la media muestral es

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i,Y_i)}{f(X_i,Y_i)}.$$

Por el teorema demostrado, la varianza de este estimador satisface

$$\mathrm{Var}(\hat{\theta}) \geqslant \frac{1}{n}\left[\left(\int_{a_1}^{b_1}\int_{a_2}^{b_2}|g(x,y)|\,dy\,dx\right)^2 - \theta^2\right],$$

en donde la cota inferior se alcanza cuando se toma

$$f(x,y) = \frac{|g(x,y)|}{\displaystyle\int_{a_1}^{b_1}\int_{a_2}^{b_2}|g(x,y)|\,dy\,dx}, \quad a_1 < x < b_1, \quad a_2 < y < b_2.$$

4.2. Muestreo condicional

Este es un procedimiento que puede producir una reducción en la varianza de un estimador insesgado. Está basado en el condicionamiento. Revisaremos primero los conceptos de esperanza y varianza condicional cuando el condicionando es una variable aleatoria, y después explicaremos la forma en la que estos conceptos pueden aplicarse al caso de la estimación de una integral por el método Monte Carlo.

Esperanza condicional

La esperanza condicional de una variable aleatoria X dado otra variable aleatoria Y se denota por el símbolo $E(X \,|\, Y)$. A pesar de su nombre y escritura, no se trata de un número sino de una variable aleatoria. Uno de sus posibles valores es el número $E(X \,|\, Y = y)$, en donde y es cualquier valor de Y. A esta última cantidad también se le llama esperanza condicional y se estudia en los cursos básicos de probabilidad.

La esperanza condicional $E(X \,|\, Y)$ se puede interpretar como el valor esperado de X a la luz de la información de Y. Parte de la dificultad para entender el término $E(X \,|\, Y)$ es que su definición se expresa en términos de tres propiedades que identifican de manera única a esta esperanza condicional. La definición es la siguiente.

Definición 4.1 *Sea X una variable aleatoria con esperanza finita y sea Y otra variable aleatoria. La esperanza condicional $E(X \,|\, Y)$ es una variable aleatoria que satisface las siguientes tres propiedades:*

1. *Es $\sigma(Y)$-medible.*
2. *Tiene esperanza finita.*
3. *Para cualquier evento $G \in \sigma(Y)$,*

$$E(X \cdot 1_G) = E(\, E(X \,|\, Y) \cdot 1_G \,).$$

Puede comprobarse que las tres propiedades indicadas en el recuadro anterior determinan de manera única (en el sentido casi seguro) a la variable $E(X \,|\, Y)$, es decir, si $\tilde{E}(X \,|\, Y)$ es otra variable aleatoria que satisface las propiedades anteriores, entonces $\tilde{E}(X \,|\, Y) = E(X \,|\, Y)$, casi seguramente. En la primera propiedad de la definición anterior, $\sigma(Y)$ es la mínima σ-álgebra respecto de la cual Y es variable aleatoria. En la tercera propiedad el símbolo 1_G indica la función indicadora del evento G.

En general, no es fácil encontrar $E(X \,|\, Y)$, o su distribución, o sus caracte-

rísticas. Su manejo se lleva a cabo a través de las propiedades que satisface, adicionales a las que aparecen en la definición. Algunas de estas propiedades se enuncian a continuación:

- $E(c\,|\,Y) = c$, $\quad c$ constante.
- $E(cX\,|\,Y) = c\,E(X\,|\,Y)$, $\quad c$ constante.
- Si $X \geqslant 0$ entonces $E(X\,|\,Y) \geqslant 0$.
- $E(X_1 + X_2\,|\,Y) = E(X_1\,|\,Y) + E(X_2\,|\,Y)$.
- Las variables aleatorias X y $E(X\,|\,Y)$ tienen la misma esperanza, es decir,
$$E(E(X\,|\,Y)) = E(X).$$
- Si X y Y son independientes entonces $E(X\,|\,Y) = E(X)$.
- Si Z es $\sigma(Y)$-medible, entonces $E(XZ\,|\,Y) = Z\,E(X\,|\,Y)$.

Las igualdades anteriores se cumplen en el sentido casi seguro, aunque esta condición no se indique. Se puede encontrar mayor información sobre $E(X\,|\,Y)$ en los libros de A. Gut [18] ó en D. Williams [61]. La esperanza condicional $E(X\,|\,Y)$ también se denota por el símbolo $E(X\,|\,\sigma(Y))$, es decir, la esperanza condicional de X dada la σ-álgebra $\sigma(Y)$. Este es un caso particular del término $E(X\,|\,\mathscr{G})$, el cual es la esperanza condicional de X dada una σ-álgebra $\mathscr{G}$. Este concepto general se estudia en las referencias indicadas.

Varianza condicional

También se puede definir la varianza condicional de X respecto de Y de la siguiente manera.

Definición 4.2 *Sea X una variable aleatoria con segundo momento finito y sea Y otra variable aleatoria. La varianza condicional de X dado Y es la variable aleatoria*

$$\mathrm{Var}(X\,|\,Y) = E(\,(X - E(X\,|\,Y))^2\,|\,Y).$$

Esta nueva variable aleatoria posee las siguientes propiedades.

Proposición 4.1 *La varianza condicional satisface las siguientes propiedades:*

1. $\mathrm{Var}(X \mid Y) \geqslant 0$.
2. $\mathrm{Var}(X \mid Y) = E(X^2 \mid Y) - E^2(X \mid Y)$.
3. $\mathrm{Var}(X) = E(\,\mathrm{Var}(X \mid Y)\,) + \mathrm{Var}(\,E(X \mid Y)\,)$.

Demostración.

1. Como $(X - E(X \mid Y))^2 \geqslant 0$, la esperanza condicional respecto de Y es no negativa. Es decir, $E(\,(X - E(X \mid Y))^2 \mid Y) \geqslant 0$.

2. Se desarrolla el cuadrado y se usa la propiedad de linealidad de la esperanza condicional, i.e.

$$\begin{aligned}
\mathrm{Var}(X \mid Y) &= E(\,(X - E(X \mid Y))^2 \mid Y) \\
&= E(X^2 - 2XE(X \mid Y) + E^2(X \mid Y) \mid Y) \\
&= E(X^2 \mid Y) - 2E(X \mid Y)\,E(X \mid Y) + E^2(X \mid Y) \\
&= E(X^2 \mid Y) - E^2(X \mid Y).
\end{aligned}$$

3. Por la propiedad anterior,

$$\begin{aligned}
E(\,\mathrm{Var}(X \mid Y)) &= E(\,E(X^2 \mid Y) - E^2(X \mid Y)\,) \\
&= E(X^2) - E(\,E^2(X \mid Y)\,) \\
&= E(X^2) - [\mathrm{Var}(\,E(X \mid Y)\,) + E^2(\,E(X \mid Y)\,)] \\
&= E(X^2) - E^2(X) - \mathrm{Var}(\,E(X \mid Y)\,) \\
&= \mathrm{Var}(X) - \mathrm{Var}(\,E(X \mid Y)\,).
\end{aligned}$$

∎

La segunda fórmula del enunciado anterior puede considerarse como una definición alternativa para la varianza condicional que recuerda la expresión

de la varianza usual de una variable aleatoria: $\text{Var}(X) = E(X^2) - E^2(X)$. Por otro lado, observe que cuando X y Y son independientes, se cumple que

$$\text{Var}(X \,|\, Y) = E(X^2) - E^2(X) = \text{Var}(X),$$

es decir, bajo la hipótesis de independencia, la variable aleatoria $\text{Var}(X \,|\, Y)$ es constante igual a $\text{Var}(X)$.

A partir de la tercera fórmula del enunciado anterior, se puede obtener el siguiente resultado que es de utilidad para proponer estimadores insesgados de varianza menor.

Proposición 4.2 *Para cualquier variable aleatoria integrable X,*

$$\text{Var}(\, E(X \,|\, Y)\,) \leqslant \text{Var}(X). \tag{4.5}$$

Demostración. Como $\text{Var}(X \,|\, Y) \geqslant 0$, se cumple que $E(\text{Var}(X \,|\, Y)) \geqslant 0$. Por la tercera fórmula de proposición previa,

$$\begin{aligned} \text{Var}(X) &= E(\, \text{Var}(X \,|\, Y)\,) + \text{Var}(\, E(X \,|\, Y)\,) \\ &\geqslant \text{Var}(\, E(X \,|\, Y)\,). \end{aligned}$$

∎

Cuando X y Y son independientes, el lado izquierdo de (4.5) es cero y la desigualdad no es informativa. Hemos mencionado antes que la esperanza condicional $E(X \,|\, Y)$ se puede considerar como un estimador insesgado para la cantidad $E(X)$ pues $E(\, E(X \,|\, Y)\,) = E(X)$, y tiene varianza posiblemente más pequeña que la varianza de X debido a (4.5). A la aplicación de este resultado al problema de estimar una integral usando el método de la media muestral de Monte Carlo se le llama muestreo condicional. Explicaremos esto en las siguientes secciones

Muestreo condicional

Sea $g(x)$ una función integrable sobre el intervalo (a, b). Consideremos nuevamente el problema de estimar la siguiente integral

$$\begin{aligned} \theta &= \int_a^b g(x)\,dx \\ &= \int_a^b \frac{g(x)}{f(x)}\, f(x)\,dx \\ &= E\left(\frac{g(X)}{f(X)}\right), \end{aligned}$$

en donde X es cualquier variable aleatoria con función de densidad $f(x)$, cuyo soporte contiene al intervalo (a, b). Si $X_1, \ldots, X_n$ es una muestra aleatoria de esta distribución, entonces el estimador de Monte Carlo de la media muestral para θ es

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}. \tag{4.6}$$

Este estimador es insesgado para θ, es decir, $E(\hat{\theta}_1) = \theta$, y su varianza es

$$\text{Var}\left(\hat{\theta}_1\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(\frac{g(X_i)}{f(X_i)}\right) = \frac{1}{n}\text{Var}\left(\frac{g(X)}{f(X)}\right).$$

Por otro lado, sabemos que algunas probabilidades y esperanzas son más fáciles de calcular cuando se condicionan sobre la ocurrencia de algún evento. En nuestra situación, supongamos que la siguiente esperanza condicional puede ser calculada con facilidad

$$E\left(\frac{g(X)}{f(X)}\,\middle|\, Y = y\right), \tag{4.7}$$

para alguna variable aleatoria Y que guarda alguna relación con X. Si contamos con una expresión para esta esperanza condicional para cada posible valor y, entonces se puede encontrar una expresión para la variable aleatoria

$$E\left(\frac{g(X)}{f(X)}\,\middle|\, Y\right), \tag{4.8}$$

Observemos que (4.7) es un número que depende del valor y, mientras que (4.8) es una variable aleatoria, a la que hemos llamado esperanza condicional. Esto sugiere que si $Y_1, \ldots, Y_n$ es una muestra aleatoria de la misma distribución que Y, entonces se puede definir el siguiente estimador por muestreo condicional

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} E\left(\frac{g(X)}{f(X)} \middle| Y_i \right). \tag{4.9}$$

Observe que no es necesario tomar el mismo tamaño de muestra n para definir a $\hat{\theta}_1$ y a $\hat{\theta}_2$, pero lo hemos hecho así a fin de comparar sus varianzas. Observemos también que cuando X y Y son independientes, el muestreo sobre Y no es provechoso pues $\hat{\theta}_2$ se reduce a la cantidad desconocida $\theta = E(g(X)/f(X))$. Las propiedades del estimador $\hat{\theta}_2$ y su comparación con $\hat{\theta}_1$ se muestran a continuación.

Proposición 4.3 *Los estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ definidos en (4.6) y (4.9) son ambos insesgados para θ, sin embargo,*

$$\text{Var}\left(\hat{\theta}_2\right) \leqslant \text{Var}\left(\hat{\theta}_1\right).$$

Demostración. El insesgamiento de $\hat{\theta}_1$ es evidente y se ha señalado antes. Demostraremos ahora el insesgamiento de $\hat{\theta}_2$,

$$\begin{aligned} E\left(\hat{\theta}_2\right) &= E\left(\frac{1}{n} \sum_{i=1}^{n} E\left(\frac{g(X)}{f(X)} \middle| Y_i \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E\left(E\left(\frac{g(X)}{f(X)} \middle| Y_i \right) \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E\left(\frac{g(X)}{f(X)} \right) \\ &= E\left(\frac{g(X)}{f(X)} \right) \\ &= \theta. \end{aligned}$$

Respecto de las varianzas, tenemos los siguientes cálculos. Para obtener la desigualdad se utiliza el resultado de la Proposición 4.2.

$$\begin{aligned}
\text{Var}\left(\hat{\theta}_2\right) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{g(X)}{f(X)}\,\middle|\, Y_i\right)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(E\left(\frac{g(X)}{f(X)}\,\middle|\, Y_i\right)\right) \\
&\leqslant \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(E\left(\frac{g(X)}{f(X)}\right)\right) \\
&= \frac{1}{n}\text{Var}\left(E\left(\frac{g(X)}{f(X)}\right)\right) \\
&= \text{Var}\left(\hat{\theta}_1\right).
\end{aligned}$$

■

Por lo tanto, el estimador condicional θ_2 es mejor que θ_1 en el sentido de tener posiblemente varianza más pequeña, sin embargo, es necesario enfatizar que la ventaja es efectiva siempre y cuando nos sea posible calcular las esperanzas condicionales indicadas en el estimador θ_2. Veamos algunos ejemplos.

Ejemplo 4.1 *Sea (X, Y) un vector aleatorio con función de distribución $F(x, y)$. Supongamos que deseamos estimar la siguiente cantidad*

$$\theta = P(X + Y \leqslant u) = E(1_{(X+Y\leqslant u)}),$$

para algún número real u fijo. Si $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una muestra aleatoria de la distribución $F(x, y)$, entonces el estimador para θ por el método directo de Monte Carlo es

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i+Y_i\leqslant u)}.$$

Alternativamente, podemos usar muestreo condicional y escribir

$$\begin{aligned}
\theta &= E(1_{(X+Y\leqslant u)}) \\
&= E(\,E(1_{(X+Y\leqslant u)}\,|\,Y)) \\
&= E(\,E(1_{(X\leqslant u-Y)}\,|\,Y)) \\
&= E(F_X(u-Y)),
\end{aligned}$$

en donde $F_X(x)$ es la función de distribución marginal de X. Por lo tanto, si $Y_1, \ldots, Y_n$ es una muestra aleatoria de la distribución $F_Y(y)$ (la función de distribución marginal de Y), entonces el estimador por muestreo condicional para θ es

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} F_X(u - Y_i),$$

en donde sabemos que la varianza de este estimador es menor o igual a la varianza de $\hat{\theta}_1$. Es interesante observar que los elementos de la suma que aparece en $\hat{\theta}_1$ son 0 ó 1, mientras que los elementos de la suma en $\hat{\theta}_2$ son valores en el intervalo $(0, 1)$. Intuitivamente esto explica el hecho de que los valores de $\hat{\theta}_2$ presentan menor dispersión.

Como un ejemplo aún más particular, podemos tomar el vector (X, Y) con función de densidad

$$f(x, y) = \begin{cases} (x + 2y)/4 & si \;\; 0 < x < 2, \quad 0 < y < 1, \\ 0 & en \; otro \; caso. \end{cases}$$

Observemos que X y Y no son independientes. Supongamos que deseamos encontrar $\theta = P(X + Y \leqslant u)$ con $u = 1$. Podemos encontrar o estimar esta probabilidad de varias maneras:

- *Se puede calcular de manera exacta*

$$\theta = \int\int_{\{(x,y):x+y\leqslant 1\}} f(x, y)\, dx\, dy.$$

  Después de algunos cálculos sencillos puede comprobarse que $\theta = 1/8$.

- *Se puede encontrar un valor aproximado usando el método de Monte Carlo directo: si $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una muestra aleatoria de la función de densidad $f(x, y)$, entonces*

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i + Y_i \leqslant 1)}.$$

  Este método requiere conocer una manera de obtener valores de la distribución con función de densidad conjunta $f(x, y)$.

- *Se puede usar el método de Monte Carlo condicional: si* $Y_1, \ldots, Y_n$ *es una muestra aleatoria de la distribución de* Y*, entonces*

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} F_X(1 - Y_i).$$

Este método requiere encontrar las distribuciones marginales de X *y de* Y*, y conocer un mecanismo para generar valores de* Y*. Evidentemente, los papeles de* X *y* Y *pueden invertirse a conveniencia.*

Por la desigualdad de la Proposición 4.3, debe ocurrir que los valores de los sumandos de $\hat{\theta}_2$ *están más concentrados alrededor de* θ *que los valores de los sumandos de* $\hat{\theta}_1$*.* ∎

Ejemplo 4.2 *Sea* $X_1, X_2 \ldots$ *una sucesión de variables aleatorias independientes con idéntica distribución* $F(x)$*. Consideremos una variable aleatoria* N *con valores* $0, 1, \ldots$ *e independiente de* $X_1, X_2 \ldots$ *Defina la suma aleatoria*

$$S_N = \sum_{i=1}^{N} X_i.$$

Observe que el número de sumandos es aleatorio. A las distribuciones de este tipo de sumas se les conoce como distribuciones compuestas. Por ejemplo, se conoce como modelo Poisson compuesto a aquella suma en donde N *tiene distribución* Poisson (λ) *y los sumandos* X_i *tienen cualquier distribución general.*

En general, el poder encontrar la distribución de S_N *es un problema complicado. Consideremos entonces el problema de estimar, usando el método de Monte Carlo, la siguiente probabilidad, en donde* x *es cualquier número real fijo,*

$$\theta = P(S_N \leqslant x) = E(\, 1_{(S_N \leqslant x)}\,).$$

- *Método Monte Carlo directo. Si* $(N_1, X_1^{(1)}, \ldots, X_{N_1}^{(1)}), \ldots, (N_n, X_1^{(n)}, \ldots, X_{N_1}^{(n)})$ *es una muestra aleatoria de tamaño* n *del vector* $(N, X_1, \ldots, X_N)$*,*

entonces la aplicación del método directo de Monte Carlo sugiere el estimador

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} 1_{(S_N^{(i)} \leqslant x)} \ ,$$

en donde $S_N^{(i)} = X_1^{(i)} + \cdots + X_{N_i}^{(i)}$ *para* $i = 1, \ldots, n$. *Sabemos que este es un estimador insesgado para* θ *y su varianza es*

$$\mathrm{Var}(\hat{\theta}_1) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(1_{(S_N^{(i)} \leqslant x)}\right) = \frac{1}{n}\,\theta\,(1-\theta).$$

- *Método Monte Carlo condicional. Para* $n \geqslant 0$ *entero fijo, la función de distribución de* S_N *con* $N = n$ *es la* n*-ésima convolución de* $F(x)$. *Esta es una nueva función de distribución denotada por* $F^{*n}(x)$. *Representa la función de distribución de la suma de* n *variables aleatorias independientes* $X_1, \ldots, X_n$ *con idéntica distribución* $F(x)$, *es decir,*

$$F^{*n}(x) = P(X_1 + \cdots + X_n \leqslant x), \quad -\infty < x < \infty.$$

Cuando $n \geqslant 2$ *podemos escribir esta probabilidad de la siguiente forma*

$$\begin{aligned}
F^{*n}(x) &= P\left(\sum_{i=1}^{n} X_i \leqslant x\right) \\
&= P\left(X_1 \leqslant x - \sum_{i=2}^{n} X_i\right) \\
&= F\left(x - \sum_{i=2}^{n} X_i\right) \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} F\left(x - \sum_{i=2}^{n} X_i \,\middle|\, X_2 = x_2, \ldots, X_n = x_n\right) \\
&\qquad\qquad f\,(x_2, \ldots, x_n)\; dx_2 \cdots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} F\left(x - \sum_{i=2}^{n} x_i\right) f\,(x_2, \ldots, x_n)\; dx_2 \cdots dx_n \\
&= E\left(\, F\left(x - \sum_{i=2}^{n} X_i\right)\right).
\end{aligned}$$

Como esta igualdad se cumple para cualquier entero $n \geqslant 1$ (cuando $n = 0$ y $n = 1$ la suma es vacía y se define como cero, de modo que la esperanza es $F(x)$), concluimos la validez de la expresión general,

$$F^{*N}(x) = E\left(F\left(x - \sum_{i=2}^{N} X_i\right)\right).$$

De esta manera, la esperanza de la variable $1_{(S_N \leqslant x)}$ se ha simplificado un poco al condicionar sobre $Y = (X_2, \ldots, X_N)$. Por lo tanto, si $(N_1, X_1^{(1)}, \ldots, X_{N_1}^{(1)}), \ldots, (N_n, X_1^{(n)}, \ldots, X_{N_1}^{(n)})$ es una muestra aleatoria de tamaño n del vector $(N, X_1, \ldots, X_N)$, se propone el estimador insesgado

$$\hat{\theta}_2 = \frac{1}{n} \sum_{k=1}^{n} F\left(x - \sum_{i=2}^{N_k} X_i^{(k)}\right),$$

el cual tiene varianza menor que el estimador que se obtiene del método de Monte Carlo directo. En este caso no se usan las primeras variables $X_1^{(k)}$ y no es necesario generarlas.

■

4.3. Variables comunes y antitéticas

Estas son variables que pueden ayudar a reducir la varianza de un estimador en algunas situaciones particulares. Ilustraremos su utilidad con algunos ejemplos.

Variables comunes

Sea (X, Y) un vector aleatorio compuesto por variables aleatorias que pueden guardar cualquier relación de dependencia. En ciertos problemas se busca estimar por simulación una esperanza de la forma

$$\theta = E(X - Y). \tag{4.10}$$

La característica principal de una esperanza de este tipo es que se puede llevar a cabo el cálculo por separado en cada variable y, en este sentido, es posible usar únicamente la distribución marginal de cada variable. Si $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_n$ son muestras aleatorias de las distribuciones marginales de X y Y, respectivamente, entonces se puede proponer el estimador insesgado

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i), \tag{4.11}$$

cuya varianza es

$$\text{Var}(\hat{\theta}) = \frac{1}{n}(\text{Var}(X) + \text{Var}(Y)). \tag{4.12}$$

Veremos que esta cantidad puede posiblemente hacerse más pequeña cuando se generan valores del vector (X, Y) de una manera muy particular.

Supondremos que X y Y tienen funciones de distribución marginales $F(x)$ y $G(y)$, respectivamente. Consideraremos que el vector (X, Y) puede tener cualquier distribución, es decir, se puede establecer cualquier grado de dependencia entre X y Y, siempre y cuando las distribuciones marginales sean las indicadas: $F(x)$ y $G(y)$, respectivamente. De esta manera, para estimar el valor θ, se puede considerar la familia de todas las distribuciones bivariadas para (X, Y) que compartan las mismas distribuciones marginales $F(x)$ y $G(y)$.

Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una muestra aleatoria del vector (X, Y) con una distribución dada. El estimador insesgado por el método de Monte Carlo directo dado en (4.11) ahora tiene varianza

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \frac{1}{n}\text{Var}(X - Y) \\ &= \frac{1}{n}(\text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y)). \end{aligned}$$

Observemos que la distribución conjunta particular de X y Y se utiliza en el cálculo de la covarianza y distintas distribuciones para (X, Y) pueden producir distintos valores de la covarianza. En contraste, los términos $\text{Var}(X)$ y $\text{Var}(Y)$ permanecen sin cambio pues las distribuciones marginales están fijas. Es claro que $\text{Var}(\hat{\theta})$ se minimiza cuando $\text{Cov}(X, Y)$ es máxima.

Demostraremos más adelante que, cuando se generan valores de X y Y usando la representación que aparece en la siguiente definición, la covarianza entre X y Y es máxima.

Definición 4.3 *Sea (X, Y) un vector aleatorio. Suponga que X tiene función de distribución marginal $F(x)$ y Y tiene función de distribución marginal $G(y)$. Cuando X y Y se expresan en términos de una misma variable $U \sim \text{unif}\,(0,1)$ como aparece abajo, se dice que se toma una variable común para representarlas.*

$$X = F^{-1}(U), \tag{4.13}$$
$$Y = G^{-1}(U). \tag{4.14}$$

Observe que en las identidades (4.13) y (4.14) aparecen las funciones inversas generalizadas de $F(x)$ y $G(y)$. Estas funciones inversas fueron definidas en la Sección 2.1. Además, por la Proposición 2.2 que se encuentra en la sección mencionada, las igualdades (4.13) y (4.14) son igualdades de variables aleatorias en el sentido de distribución, es decir, las variables aleatorias de uno y otro lado de cada igualdad tienen la misma distribución.

Así, bajo la representación de la definición anterior (variables comunes),

$$\begin{aligned}
\text{Var}(\hat{\theta}) &= \frac{1}{n}\,\text{Var}(X - Y) && (4.15)\\
&= \frac{1}{n}\,\text{Var}(F^{-1}(U) - G^{-1}(U)) \\
&= \frac{1}{n}\,\big[\,\text{Var}(F^{-1}(U)) + \text{Var}(G^{-1}(U)) - 2\,\text{Cov}(F^{-1}(U), G^{-1}(U))\,\big] \\
&= \frac{1}{n}\,\big[\,\text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(F^{-1}(U), G^{-1}(U))\,\big]. && (4.16)
\end{aligned}$$

Es posible demostrar que $\text{Cov}(F^{-1}(U), G^{-1}(U)) \geqslant 0$ y, por lo tanto, la expresión (4.16) es menor o igual la varianza (4.12) que se obtiene cuando se generan valores separados para X y Y. Demostraremos más adelante que la expresión (4.16) es el valor más pequeño para $\text{Var}(\hat{\theta})$, dentro de todas las distribuciones que se pueden considerar para (X, Y) con las mismas

distribuciones marginales $F(x)$ y $G(y)$, respectivamente.

Debe observarse que las representaciones (4.13) y (4.14) crean una dependencia completa entre X y Y, y que la distribución conjunta de estas variables posee distribuciones marginales $F(x)$ y $G(y)$.

Variables antitéticas

Otro problema similar al de la subsección anterior es estimar por simulación una esperanza de la forma

$$\theta = E(X + Y). \tag{4.17}$$

Nuevamente, la característica principal de este tipo de esperanzas es que se puede llevar a cabo un cálculo por separado en cada variable y es posible usar únicamente la distribución marginal de las variables. Como antes, supondremos que X y Y tienen funciones de distribución marginales $F(x)$ y $G(y)$, respectivamente.

Si $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_n$ son muestras aleatorias de las distribuciones marginales de X y Y, respectivamente, entonces se puede proponer el estimador insesgado

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (X_i + Y_i), \tag{4.18}$$

cuya varianza es

$$\text{Var}(\hat{\theta}) = \frac{1}{n}(\text{Var}(X) + \text{Var}(Y)). \tag{4.19}$$

Como antes, veremos que esta cantidad puede posiblemente hacerse más pequeña cuando se generan valores del vector (X, Y) de una manera muy particular.

Si $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una muestra aleatoria del vector (X, Y), el estimador insesgado por el método de Monte Calo directo para θ es

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (X_i + Y_i),$$

cuya varianza es

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \frac{1}{n}\text{Var}(X+Y) \\ &= \frac{1}{n}(\text{Var}(X)+\text{Var}(Y)+2\,\text{Cov}(X,Y)). \end{aligned}$$

Nuevamente, la distribución conjunta que se suponga para X y Y es utilizada en el cálculo de la covarianza y distintas distribuciones para (X,Y) producen distintos valores de la covarianza. Es claro que $\text{Var}(\hat{\theta})$ se minimiza cuando $\text{Cov}(X,Y)$ es mínima. Demostraremos más adelante que, cuando se generan valores de X y Y usando la representación que aparece en la siguiente definición, la covarianza entre X y Y es mínima.

Definición 4.4 *Sea (X,Y) un vector aleatorio. Suponga que X tiene función de distribución marginal $F(x)$ y Y tiene función de distribución marginal $G(y)$. Cuando X y Y se expresan en términos de una variable $U \sim \text{unif}\,(0,1)$ como aparece abajo, se dice que se toman variables antitéticas para representarlas.*

$$X = F^{-1}(U), \tag{4.20}$$
$$Y = G^{-1}(1-U). \tag{4.21}$$

Así, bajo la representación de la definición anterior (variables antitéticas),

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \frac{1}{n}\,\text{Var}(X+Y) \\ &= \frac{1}{n}\,\text{Var}(F^{-1}(U)+G^{-1}(1-U)) \\ &= \frac{1}{n}\,\big[\,\text{Var}(F^{-1}(U))+\text{Var}(G^{-1}(1-U)) \\ &\qquad +2\;\text{Cov}(F^{-1}(U),G^{-1}(1-U))\,\big] \\ &= \frac{1}{n}\,\big[\,\text{Var}(X)+\text{Var}(Y)+2\,\text{Cov}(F^{-1}(U),G^{-1}(1-U))\,\big]. \end{aligned} \tag{4.22}$$

Es posible demostrar que $\text{Cov}(F^{-1}(U),G^{-1}(1-U)) \leqslant 0$ y, por lo tanto, la expresión (4.22) es menor o igual la varianza (4.19) que se obtiene cuando

se generan valores separados para X y Y. Demostraremos más adelante que la expresión (4.22) es el valor más pequeño para $\mathrm{Var}(\hat{\theta})$, dentro de todas las distribuciones que se pueden considerar para (X, Y) con las mismas distribuciones marginales $F(x)$ y $G(y)$, respectivamente.

Debe observarse nuevamente que las representaciones (4.20) y (4.21) crean una dependencia completa entre X y Y, y que la distribución conjunta de estas variables posee distribuciones marginales $F(x)$ y $G(y)$.

Cotas para la covarianza

El resultado general que aparece abajo permite comprobar que, para cualquier distribución del vector (X, Y) con distribuciones marginales $F(x)$ y $G(y)$, respectivamente,

$$\mathrm{Cov}(F^{-1}(U), G^{-1}(1-U)) \leqslant \mathrm{Cov}(X, Y) \leqslant \mathrm{Cov}(F^{-1}(U), G^{-1}(U)). \quad (4.23)$$

La expresión en el extremo derecho corresponde a tomar variables comunes para representar a X y Y, mientras que la expresión en el extremo izquierdo corresponde a usar variables antitéticas.

Teorema 4.2 *Sean X y Y dos variables aleatorias, no necesariamente independientes, con funciones de distribución $F(x)$ y $G(y)$, respectivamente. Sean h_1 y h_2 dos funciones monótonas, diferenciables, ambas crecientes o ambas decrecientes. Sea $U \sim \mathrm{unif}\,(0, 1)$. Entonces la covarianza*

$$\mathrm{Cov}(h_1(X), h_2(Y)) \quad (4.24)$$

1. *Se maximiza cuando X y Y se expresan mediante una variable común, es decir, $X = F^{-1}(U)$ y $Y = G^{-1}(U)$.*
2. *Se minimiza cuando X y Y se expresan mediante variables antitéticas, es decir, $X = F^{-1}(U)$ y $Y = G^{-1}(1-U)$.*

Demostración. Por ahora mantendremos la notación explícita para las funciones de densidad y de distribución de las variables aleatorias. Observe-

mos primero la validez de la siguiente relación general

$$\begin{aligned} P(X > x, Y > y) &= P(X > x) - P(X > x, Y \leqslant y) \\ &= P(X > x) - P(Y \leqslant y) + P(X \leqslant x, Y \leqslant y) \\ &= [1 - F_X(x)] - F_Y(y) + F_{X,Y}(x, y). \end{aligned}$$

Utilizaremos la siguiente función auxiliar,

$$\begin{aligned} H(x, y) &= P(X > x, Y > y) - P(X > x)\, P(Y > y) \\ &= [1 - F_X(x)] - F_Y(y) + F_{X,Y}(x, y) \\ &\qquad -[1 - F_X(x)] \cdot [1 - F_Y(y)]. \end{aligned}$$

Derivando respecto de ambas variables,

$$\frac{\partial^2}{\partial x \partial y} H(x, y) = f_{X,Y}(x, y) - f_X(x)\, f_Y(y).$$

Ahora podemos calcular la covarianza buscada. En el análisis que aparece a continuación aplicaremos el método de integración por partes dos veces. Inicialmente, por definición,

$$\begin{aligned} &\mathrm{Cov}(h_1(X), h_2(Y)) \\ &= E(h_1(X) \cdot h_2(Y)) - E(h_1(X)) \cdot E(h_2(Y)) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_1(x) \cdot h_2(y)\, f(x, y)\, dx\, dy - E(h_1(X)) \cdot E(h_2(Y)) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_1(x) \cdot h_2(y) \left[\frac{\partial^2}{\partial x \partial y} H(x, y) + f_X(x) \cdot f_Y(y) \right] dx\, dy \\ &\qquad -E(h_1(X)) \cdot E(h_2(Y)) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_1(x) \cdot h_2(y)\, \frac{\partial^2}{\partial x \partial y} H(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} h_2(y) \int_{-\infty}^{\infty} h_1(x) \cdot \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} H(x, y) \right) dx\, dy. \end{aligned}$$

Llamando $u = h_1(x)$ y dv a la doble derivada, tenemos que

$$\begin{aligned}\mathrm{Cov}(h_1(X), h_2(Y)) &= \int_{-\infty}^{\infty} h_2(y) \underbrace{\left[h_1(x)\frac{\partial}{\partial y}H(x,y)\right]_{-\infty}^{\infty}}_{=0} dy \\ &\quad - \int_{-\infty}^{\infty} h_2(y) \int_{-\infty}^{\infty} \frac{\partial}{\partial y}H(x,y)\, h_1'(x)\, dx\, dy \\ &= - \int_{-\infty}^{\infty} h_1'(x) \int_{-\infty}^{\infty} h_2(y)\frac{\partial}{\partial y}H(x,y)\, dy\, dx.\end{aligned}$$

Ahora llamamos $u = h_2(y)$ y dv a la derivada respecto de y,

$$\begin{aligned}\mathrm{Cov}(h_1(X), h_2(Y)) &= - \int_{-\infty}^{\infty} h_1'(x) \underbrace{\left[h_2(y)\, H(x,y)\right]_{-\infty}^{\infty}}_{=0} dx \\ &\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y) h_1'(x) h_2'(y)\, dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y) h_1'(x) h_2'(y)\, dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X > x, Y > y) \cdot h_1'(x) h_2'(y)\, dxdy \qquad (4.25) \\ &\quad - \left[\int_{-\infty}^{\infty} P(X > x)\, h_1'(x)\, dx\right] \\ &\quad \cdot \left[\int_{-\infty}^{\infty} P(Y > y)\, h_2'(y)\, dy\right].\end{aligned}$$

Observemos que, como $h_1(x)$ y $h_2(x)$ comparten el mismo tipo de monotonía, $h_1'(x)h_2'(y) \geqslant 0$, pues ambos factores son positivos o ambos son negativos. A continuación nos concentraremos en el término $P(X > x, Y > y)$ que aparece en (4.25). Encontraremos una cota inferior y una cota superior para este término.

- *Cota superior.* Sean x y y dos valores cualesquiera y, sin pérdida de

generalidad, supongamos que $F(x) \geqslant G(y)$. Entonces

$$\begin{aligned}
P(X > x, Y > y) &= P((X > x) \cap (Y > y)) \\
&= 1 - P((X \leqslant x) \cup (Y \leqslant y)) \\
&= 1 - [F(x) + G(y) - P(X \leqslant x, Y \leqslant y)] \\
&= 1 - F(x) - G(y) + P(X \leqslant x, Y \leqslant y) \\
&\leqslant 1 - F(x) \underbrace{-G(y) + P(Y \leqslant y)}_{=0} \\
&= 1 - F(x) \\
&= 1 - \text{máx}\{F(x), G(y)\} \\
&= P(U > \text{máx}\{F(x), G(y)\} \\
&= P(U > F(x), U > G(y)) \\
&= P(F^{-1}(U) > x, G^{-1}(U) > y).
\end{aligned}$$

Es decir, hemos demostrado que el evento general $(X > x, Y > y)$, en donde las variables X y Y pueden mantener cualquier relación, tiene probabilidad menor o igual a la probabilidad del evento $(F^{-1}(U) > x, G^{-1}(U) > y)$, en donde, individualmente,

$$X \stackrel{d}{=} F^{-1}(U),$$

$$Y \stackrel{d}{=} G^{-1}(U),$$

y en donde las variables $F^{-1}(U)$ y $G^{-1}(U)$ son completamente dependientes, una es función de la otra.

Por lo tanto, retomando la probabilidad que aparece indicada en (4.25), el número $\text{Cov}(h_1(X), h_2(Y))$ se maximiza cuando se toma una variable común U para obtener valores de X y Y. En particular, cuando h_1 y h_2 son ambas la función identidad, se tiene el caso mencionado antes de la estimación de $E(X - Y)$ con varianza mínima.

- *Cota inferior.* Nuevamente, sean x y y dos valores cualesquiera y, sin

pérdida de generalidad, supongamos que $F(x) \leqslant 1 - G(y)$. Entonces

$$\begin{aligned} P(X > x, Y > y) &= P((X > x) \cap (Y > y)) \\ &= 1 - P((X \leqslant x) \cup (Y \leqslant y)) \\ &= 1 - [F(x) + G(y) - P(X \leqslant x, Y \leqslant y)] \\ &\geqslant (1 - G(y)) - F(x) \\ &= P(F(x) < U < 1 - G(y)) \\ &= P(U > F(x), 1 - U > G(y)) \\ &= P(F^{-1}(U) > x, G^{-1}(1 - U) > y). \end{aligned}$$

Por lo tanto, retomando nuevamente la probabilidad que aparece indicada en (4.25), el número $\text{Cov}(h_1(X), h_2(Y))$ se minimiza cuando se toman variables antitéticas U y $1 - U$ para generar valores de X y Y. En particular, cuando h_1 y h_2 son ambas la función identidad, se tiene el caso mencionado antes de la estimación de $E(X + Y)$ con varianza mínima.

■

En conclusión, bajo la notación e hipótesis del teorema anterior, tenemos las siguientes cotas para la covarianza entre $h_1(X)$ y $h_2(Y)$,

$$\underbrace{\text{Cov}(h_1(F^{-1}(U)), h_2(G^{-1}(1 - U)))}_{\text{Variables antitéticas}} \leqslant \text{Cov}(h_1(X), h_2(Y))$$

$$\text{Cov}(h_1(X), h_2(Y)) \leqslant \underbrace{\text{Cov}(h_1(F^{-1}(U)), h_2(G^{-1}(U)))}_{\text{Variables comunes}}$$

Tomando $h_1(x) = x$ y $h_2(x) = x$, las cotas anteriores producen la expresión (4.23) que aparece al inicio de la presente sección.

Se puede realizar un análisis para determinar si es mejor utilizar variables comunes o variables antitéticas en los casos cuando la funciones h_1 y h_2 presentan monotonía en dirección contraria, tanto en el caso $E(h_1(X) - h_2(Y))$ como en $E(h_1(X) + h_2(Y))$. Los resultados se pueden extender al caso de vectores aleatorios $X = (X_1, \ldots, X_n)$ y $Y = (Y_1, \ldots, Y_n)$, y para funciones vectoriales h_1 y h_2 que son monótonas en cada variable.

4.4. Variables de control

El uso de variables de control constituye otro método que puede ayudar a reducir la varianza de un estimador. Supongamos que $\hat{\theta}$ es un estimador insesgado para una cantidad desconocida θ. En nuestro contexto, el estimador $\hat{\theta}$ se construye a partir de una muestra aleatoria generada por algún algoritmo de simulación.

Definición 4.5 *A una variable aleatoria no constante C con esperanza finita y conocida, y que está correlacionada con un estimador $\hat{\theta}$ se le llama variable de control.*

Las variables de control pueden ser usadas para construir otros estimadores insesgados para θ pero de varianza menor como el siguiente.

Definición 4.6 *Sea $\hat{\theta}$ un estimador insesgado para una cantidad desconocida θ. Sea C una variable de control y sea α cualquier número real. Se define el estimador*

$$\hat{\theta}_\alpha := \hat{\theta} - \alpha(C - E(C)). \tag{4.26}$$

El siguiente resultado muestra algunas propiedades de $\hat{\theta}_\alpha$.

Proposición 4.4 *El estimador $\hat{\theta}_\alpha$ definido en (4.26) es insesgado y tiene varianza menor o igual a la varianza de $\hat{\theta}$. El valor más pequeño para $\mathrm{Var}(\hat{\theta}_\alpha)$ se alcanza cuando $\alpha^* = \mathrm{Cov}(\hat{\theta}, C)/\mathrm{Var}\,(C)$ y es*

$$\mathrm{Var}\left(\hat{\theta}_{\alpha*}\right) = \left(1 - \rho^2\left(\hat{\theta}, C\right)\right)\mathrm{Var}(\hat{\theta}) \leqslant \mathrm{Var}(\hat{\theta}),$$

en donde $\rho\left(\hat{\theta}, C\right)$ es el coeficiente de correlación entre $\hat{\theta}$ y C.

Demostración. Tomando esperanza en (4.26), es inmediato comprobar que $E(\hat{\theta}_\alpha) = E\left(\hat{\theta}\right) = \theta$, es decir, el estimador $\hat{\theta}_\alpha$ también es insesgado para θ, para cualquier valor de la constante α. La varianza del nuevo estimador es

$$\begin{aligned}
\operatorname{Var}(\hat{\theta}_\alpha) &= \operatorname{Var}\Big(\hat{\theta} - \alpha C + \alpha E(C)\Big) \\
&= \operatorname{Var}\Big(\hat{\theta} - \alpha C\Big) \\
&= \operatorname{Var}(\hat{\theta}) + \alpha^2 \operatorname{Var}(C) - 2\alpha \operatorname{Cov}(\hat{\theta}, C).
\end{aligned}$$

Esta es una función cuadrática en α y tiene un valor mínimo en el punto en donde su derivada se anula, es decir, en aquel valor α tal que

$$2\alpha \operatorname{Var}(C) - 2 \operatorname{Cov}(\hat{\theta}, C) = 0.$$

Es decir, cuando

$$\alpha^* = \frac{\operatorname{Cov}(\hat{\theta}, C)}{\operatorname{Var}(C)}.$$

El valor mínimo de la varianza es entonces

$$\begin{aligned}
\operatorname{Var}(\hat{\theta}_{\alpha*}) &= \operatorname{Var}(\hat{\theta}) + \frac{\operatorname{Cov}^2(\hat{\theta}, C)}{\operatorname{Var}^2(C)} \operatorname{Var}(C) - 2 \frac{\operatorname{Cov}(\hat{\theta}, C)}{\operatorname{Var}(C)} \operatorname{Cov}(\hat{\theta}, C) \\
&= \operatorname{Var}(\hat{\theta}) - \frac{\operatorname{Cov}^2(\hat{\theta}, C)}{\operatorname{Var}(C)} \\
&= \operatorname{Var}(\hat{\theta}) - \left(\frac{\operatorname{Cov}(\hat{\theta}, C)}{\sqrt{\operatorname{Var}(\hat{\theta}) \operatorname{Var}(C)}} \right)^2 \operatorname{Var}(\hat{\theta}) \\
&= (1 - \rho^2(\hat{\theta}, C)) \operatorname{Var}(\hat{\theta}).
\end{aligned}$$

Claramente,

$$\operatorname{Var}\left(\hat{\theta}_{\alpha*}\right) \leqslant \operatorname{Var}(\hat{\theta}).$$

∎

De esta manera, mientras mayor sea el valor del coeficiente del correlación $\rho(\hat{\theta}, C)$, menor varianza tendrá el nuevo estimador insesgado $\hat{\theta}_{\alpha*}$. Se pueden distinguir los siguientes dos casos extremos:

- Cuando $\hat{\theta}$ y C son independientes, se cumple que $\text{Cov}(\hat{\theta}, C) = 0$. Por lo tanto, $\rho(\hat{\theta}, C) = 0$ y las varianzas de $\hat{\theta}$ y $\hat{\theta}_{\alpha *}$ coinciden, de modo que no hay ninguna ganancia en la reducción de la varianza.
- Cuando se toma $C = \hat{\theta}$, se tiene que $\rho(\hat{\theta}, C) = 1$, $\text{Cov}(\hat{\theta}, C) = \text{Var}(C)$ y $\alpha^* = 1$. En este caso, el estimador $\hat{\theta}_\alpha$ se reduce a la constante θ y no representa realmente un estimador.

Ejemplo 4.3 *Este es un ejemplo genérico en el que se muestra la manera en la que se puede reducir la varianza en la aproximación de una integral mediante el estimador clásico de Monte Carlo usando una variable de control. Sea $g(x)$ una función integrable sobre el intervalo acotado (a, b). Defina θ como la integral que aparece abajo y la cual deseamos estimar debido a que no conocemos un método analítico para calcularla,*

$$\theta := \int_a^b g(x)\, dx.$$

Si $X_1, \ldots, X_n$ es una muestra aleatoria de la distribución $\text{unif}(a, b)$*, entonces el estimador Monte Carlo clásico para θ es*

$$\hat{\theta} := (b - a)\, \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Puede comprobarse que $\hat{\theta}$ es un estimador insesgado para θ, es decir, $E(\hat{\theta}) = \theta$. El cálculo de la varianza de $\hat{\theta}$ nos regresa al problema de encontrar θ pues, si X representa cualquier elemento de la muestra aleatoria,

$$\begin{aligned}
\text{Var}(\hat{\theta}) &= (b-a)^2\, \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(g(X_i)) \\
&= (b-a)^2\, \frac{1}{n} \text{Var}(g(X)) \\
&= (b-a)^2\, \frac{1}{n} [E(g^2(X)) - E^2(g(X))] \\
&= (b-a)\, \frac{1}{n} \int_a^b g^2(x)\, dx - \frac{1}{n} \left(\int_a^b g(x)\, dx \right)^2.
\end{aligned}$$

Por otro lado, sea $h(x)$ otra función integrable sobre el mismo intervalo (a, b), pero tal que su integral puede calcularse con facilidad. Defina la variable de control

$$C := (b-a)\,\frac{1}{n}\,\sum_{i=1}^{n} h(X_i),$$

cuya esperanza es

$$E(C) = E(h(X)) = \int_a^b h(x)\,dx,$$

y su varianza es

$$\begin{aligned}
\mathrm{Var}\,(C) &= (b-a)^2\,\frac{1}{n}\mathrm{Var}(h(X)) \\
&= (b-a)^2\,\frac{1}{n}[E(h^2(X)) - E^2(h(X))] \\
&= (b-a)\,\frac{1}{n}\int_a^b h^2(x)\,dx - \frac{1}{n}\,\left(\int_a^b h(x)\,dx\right)^2.
\end{aligned}$$

Estamos suponiendo que las integrales de $h^2(x)$ y $h(x)$ son fáciles de encontrar, de modo que podemos calcular $\mathrm{Var}(C)$. *De acuerdo a la Definición 4.6 y por la Proposición 4.4, el estimador insesgado $\hat{\theta}_\alpha := \hat{\theta} - \alpha(C - E(C))$ alcanza su varianza mínima en $\alpha^* = \mathrm{Cov}(\hat{\theta}, C)/\mathrm{Var}(C)$. Podremos encontrar este valor para α si el término* $\mathrm{Cov}(\hat{\theta}, C)$ *puede calcularse, en realidad sólo podremos proponer una aproximación. Tenemos que*

$$\begin{aligned}
\mathrm{Cov}(\hat{\theta}, C) &= \mathrm{Cov}((b-a)\,\frac{1}{n}\,\sum_{i=1}^{n} g(X_i), (b-a)\,\frac{1}{n}\,\sum_{i=1}^{n} h(X_i)) \\
&= (b-a)^2\,\frac{1}{n^2}\,\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(g(X_i), h(X_j)).
\end{aligned}$$

Cuando $i \neq j$, las variables aleatorias $g(X_i)$ y $h(X_j)$ son independientes y, por lo tanto, $\mathrm{Cov}(g(X_i), h(X_j)) = 0$. *Entonces*

$$\begin{aligned}
\mathrm{Cov}(\hat{\theta}, C) &= (b-a)^2\,\frac{1}{n}\,\mathrm{Cov}(g(X), h(X)) \\
&= (b-a)^2\,\frac{1}{n}\,[E(g(X)h(X)) - E(g(X))E(h(X))].
\end{aligned}$$

Esta cantidad sólo podrá encontrarse si la función $h(x)$ es tal que la esperanza $E(g(X)h(X))$ puede calcularse y, por otro lado, encontrar el término $E(g(X))$ es equivalente a conocer el valor de θ. Sobre este último término puede substituirse el valor de $E(g(X))$ por su aproximación Monte Carlo y obtener así un valor de α^* aproximado. ∎

Veremos a continuación un caso particular del ejemplo anterior.

Ejemplo 4.4 *Supongamos que deseamos aproximar la integral*

$$\theta := \int_0^1 e^{-x^2/2}\, dx.$$

Se puede escribir $\theta = E(g(X))$, en donde $g(x) = e^{-x^2/2}$ y X es una variable aleatoria con distribución $\text{unif}(0,1)$. *Si $X_1, \ldots, X_n$ es una muestra aleatoria de esta distribución, el estimador Monte Carlo clásico para θ es*

$$\hat{\theta} := \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

La esperanza de $\hat{\theta}$ es $E(\hat{\theta}) = \theta$ y la varianza es desconocida pues $\text{Var}(\hat{\theta}) = (1/n)\text{Var}(g(X)) = (1/n)[E(g^2(X)) - E^2(g(X))]$.

Sea $h(x) = x$ y defina la variable de control

$$C := \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$

con esperanza $E(C) = E(X) = 1/2$ y varianza $\text{Var}(C) = (1/n)\text{Var}(X) = (1/n)(1/12)$. *El estimador que incorpora a la variable de control es $\hat{\theta}_\alpha =$*

$\hat{\theta} - \alpha(C - E(C))$ *en donde el valor óptimo para* α *es*

$$\begin{aligned}
\alpha^* &= \frac{\operatorname{Cov}(\hat{\theta}, C)}{\operatorname{Var}(C)} \\
&= 12n \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i), \frac{1}{n}\sum_{j=1}^{n} h(X_j)\right) \\
&= 12 \operatorname{Cov}(g(X), h(X)) \\
&= 12\left[E(g(X)h(X)) - E(g(X))E(h(X))\right] \\
&= 12\left[\int_0^1 x e^{-x^2/2}\,dx - \left(\int_0^1 e^{-x^2/2}\,dx\right)\left(\int_0^1 x\,dx\right)\right] \\
&= 12\left[\left(1 - e^{-1/2}\right) - \frac{1}{2}\left(\int_0^1 e^{-x^2/2}\,dx\right)\right] \\
&\approx 12\left[\left(1 - e^{-1/2}\right) - \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} e^{-X_i^2/2}\right)\right],
\end{aligned}$$

en donde $x_1, \ldots, x_n$ *son los valores de la muestra aleatoria* $X_1, \ldots, X_n$. *Tomando* $n = 100$ *se obtuvo el valor aproximado* $\alpha^* \approx -0.4218966$. *Los resultados de la reducción de la varianza se muestran en la Tabla 4.1. Una aproximación numérica (redondeada a 5 dígitos) para* θ *arroja el valor* 0.85562 . *El estimador con la variable de control incorporada* $\hat{\theta}_{\alpha *}$ *es más preciso y presenta menor varianza.*

	Estimación	Varianza
Estimador clásico $\hat{\theta}$	0.85725	0.01455488
Estimador $\hat{\theta}_{\alpha *}$ con variable de control	0.85665	0.0005362

Tabla 4.1: Resultados de estimación y varianza del Ejemplo 4.4.

■

Múltiples variables de control

La definición que aparece en la ecuación (4.26) se puede extender para incorporar varias variables de control de la siguiente manera.

Definición 4.7 *Sea $\hat{\theta}$ un estimador insesgado para una cantidad desconocida θ y sea $C = (C_1, \ldots, C_m)$ un vector de variables de control. Para cualquier valor del vector $\alpha = (\alpha_1, \ldots, \alpha_m)$ se propone el nuevo estimador*

$$\hat{\theta}_\alpha := \hat{\theta} - \sum_{i=1}^{m} \alpha_i \left(C_i - E(C_i)\right). \tag{4.27}$$

Claramente, la ecuación (4.27) se puede escribir en notación matricial como sigue

$$\hat{\theta}_\alpha = \hat{\theta} - (\alpha_1, \ldots, \alpha_m) \begin{pmatrix} C_1 - E(C_1) \\ \vdots \\ C_m - E(C_m) \end{pmatrix}.$$

En el siguiente resultado se muestran las principales propiedades del nuevo estimador $\hat{\theta}_\alpha$ definido en (4.27).

Proposición 4.5 *Sea Σ la matriz de covarianzas $\mathrm{Cov}\,(C_i, C_j)$ y suponga que es invertible. El estimador $\hat{\theta}_\alpha$ definido en (4.27) es insesgado y tiene varianza menor o igual a la varianza de $\hat{\theta}$. El valor más pequeño para $\mathrm{Var}(\hat{\theta}_\alpha)$ se alcanza cuando $\alpha^* = (\Sigma)^{-1} \cdot \mathrm{Cov}(\hat{\theta}, C)$ y es*

$$\begin{aligned} \mathrm{Var}\left(\hat{\theta}_{\alpha^*}\right) &= \left[1 - \frac{\mathrm{Cov}(\hat{\theta}, C) \cdot \Sigma^{-1} \cdot \left(\mathrm{Cov}(\hat{\theta}, C)\right)^t}{\mathrm{Var}(\hat{\theta})}\right] \cdot \mathrm{Var}(\hat{\theta}) \\ &\leqslant \mathrm{Var}(\hat{\theta}), \end{aligned}$$

en donde $\mathrm{Cov}(\hat{\theta}, C) = \left(\mathrm{Cov}\left(\hat{\theta}, C_1\right), \ldots, \mathrm{Cov}\left(\hat{\theta}, C_m\right)\right)$.

Demostración. Aplicando la propiedad de linealidad de la esperanza puede comprobarse con facilidad que el estimador $\hat{\theta}_\alpha$ es insesgado y su varianza ahora es

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_\alpha) &= \mathrm{Var}\left(\hat{\theta} - \sum_{i=1}^{m} \alpha_i \, (C_i - E\,(C_i))\right) \\
&= \mathrm{Var}\left(\hat{\theta} - \sum_{i=1}^{m} \alpha_i \, (C_i)\right) \\
&= \mathrm{Var}(\hat{\theta}) + \mathrm{Var}\left(\sum_{i=1}^{m} \alpha_i \, C_i\right) - 2\,\mathrm{Cov}\left(\hat{\theta}, \sum_{i=1}^{m} \alpha_i \, C_i\right) \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_1\alpha_j \mathrm{Cov}\,(C_i, C_j) \\
&\qquad -2\sum_{i=1}^{m} \alpha_i \,\mathrm{Cov}\left(\hat{\theta}, C_i\right). \qquad (4.28)
\end{aligned}$$

Esta es una función cuadrática del vector $\alpha = (\alpha_1, \ldots, \alpha_m)$ y posee un valor mínimo. Derivando respecto de α_i e igualando a cero se encuentra que

$$2\sum_{j=1}^{m} \alpha_j \,\mathrm{Cov}\,(C_i, C_j) - 2\,\mathrm{Cov}\left(\hat{\theta}, C_i\right) = 0, \quad i = 1, 2, \ldots, m,$$

es decir, en notación matricial,

$$\begin{pmatrix} \mathrm{Cov}\,(1,1) & \mathrm{Cov}\,(1,2) & \cdots & \mathrm{Cov}\,(1,m) \\ \mathrm{Cov}\,(2,1) & \mathrm{Cov}\,(2,2) & \cdots & \mathrm{Cov}\,(2,m) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}\,(m,1) & \mathrm{Cov}\,(m,2) & \cdots & \mathrm{Cov}\,(m,m) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(\hat{\theta}, C_1) \\ \mathrm{Cov}(\hat{\theta}, C_2) \\ \vdots \\ \mathrm{Cov}(\hat{\theta}, C_m) \end{pmatrix},$$

en donde $\mathrm{Cov}(i,j) = \mathrm{Cov}(C_i, C_j)$. Por lo tanto, el punto en donde la varianza se minimiza es

$$\begin{aligned}
\alpha^* &= \begin{pmatrix} \mathrm{Cov}(1,1) & \mathrm{Cov}(1,2) & \cdots & \mathrm{Cov}(1,m) \\ \mathrm{Cov}(2,1) & \mathrm{Cov}(2,2) & \cdots & \mathrm{Cov}(2,m) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(m,1) & \mathrm{Cov}(m,2) & \cdots & \mathrm{Cov}(m,m) \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{Cov}(\hat{\theta}, C_1) \\ \mathrm{Cov}(\hat{\theta}, C_2) \\ \vdots \\ \mathrm{Cov}(\hat{\theta}, C_m) \end{pmatrix} \\
&= (\Sigma)^{-1} \cdot (\mathrm{Cov}(\hat{\theta}, C))^t,
\end{aligned}$$

en donde Σ es la matriz de varianzas y covarianza indicada, la cual es simétrica, es decir, $\Sigma = \Sigma^t$. A partir de (4.28) y de la notación anterior, el valor mínimo de la varianza es

$$\begin{aligned}
\text{Var}(\hat{\theta}_{\alpha*}) &= \text{Var}(\hat{\theta}) + (\alpha^*)^t \cdot \Sigma \cdot \alpha^* - 2(\alpha^*)^t \cdot \text{Cov}(\hat{\theta}, C) \\
&= \text{Var}(\hat{\theta}) + (\text{Cov}(\hat{\theta}, C))^t (\Sigma^{-1})^t \cdot \underbrace{\Sigma \cdot (\Sigma^{-1})}_{=1} \cdot \text{Cov}(\hat{\theta}, C) \\
&\qquad -2(\text{Cov}(\hat{\theta}, C))^t (\Sigma^{-1})^t \cdot \text{Cov}(\hat{\theta}, C) \\
&= \text{Var}(\hat{\theta}) - (\text{Cov}(\hat{\theta}, C))^t \cdot (\Sigma^{-1})^t \cdot \text{Cov}(\hat{\theta}, C) \\
&= \text{Var}(\hat{\theta}) - \text{Cov}(\hat{\theta}, C) \cdot \Sigma^{-1} \cdot (\text{Cov}(\hat{\theta}, C))^t \\
&= \left[1 - \underbrace{\frac{\text{Cov}(\hat{\theta}, C) \cdot \Sigma^{-1} \cdot (\text{Cov}(\hat{\theta}, C))^t}{\text{Var}(\hat{\theta})}} \right] \cdot \text{Var}(\hat{\theta}),
\end{aligned}$$

en donde el término subrayado es el así llamado coeficiente de correlación múltiple entre $\hat{\theta}$ y C. Puede comprobarse que este coeficiente múltiple siempre toma un valor en el intervalo $[0, 1]$, de modo que

$$\text{Var}\left(\hat{\theta}_{\alpha*}\right) \leqslant \text{Var}(\hat{\theta}).$$

■

De esta manera, mientras mayor sea el valor del coeficiente del correlación múltiple entre $\hat{\theta}$ y C, menor varianza tendrá el nuevo estimador insesgado $\hat{\theta}_{\alpha*}$. Se pueden distinguir los siguientes casos particulares:

- Cuando C es unidimensional, tenemos que $\Sigma = \text{Var}(C)$ y el coeficiente de correlación múltiple se reduce al cuadrado del coeficiente de correlación usual, es decir,

$$\begin{aligned}
\frac{\text{Cov}(\hat{\theta}, C) \cdot \Sigma^{-1} \cdot (\text{Cov}(\hat{\theta}, C))^t}{\text{Var}(\hat{\theta})} &= \frac{(\text{Cov}(\hat{\theta}, C))^2 \cdot (\text{Var}(C))^{-1}}{\text{Var}(\hat{\theta})} \\
&= \frac{(\text{Cov}(\hat{\theta}, C))^2}{\text{Var}(\hat{\theta}) \cdot \text{Var}(C)} \\
&= \rho^2(\hat{\theta}, C).
\end{aligned}$$

- Cuando $\hat{\theta}$ y C son independientes, se cumple que $\mathrm{Cov}(\hat{\theta}, C) = (0, \ldots, 0)$. En consecuencia, el coeficiente de correlación múltiple se anula y las varianzas de $\hat{\theta}$ y $\hat{\theta}_{\alpha *}$ coinciden, de modo que no hay ninguna ganancia en la reducción de la varianza.

- Cuando se toma $C = (\hat{\theta}, \ldots, \hat{\theta})$, se tiene que

$$\mathrm{Cov}(\hat{\theta}, C) = (\mathrm{Var}(\hat{\theta}), \ldots, \mathrm{Var}(\hat{\theta})),$$

  y la matriz Σ tiene en todas sus entradas el mismo valor $\mathrm{Cov}(\hat{\theta}, \hat{\theta}) = \mathrm{Var}(\hat{\theta})$ y no es invertible (excepto en el caso unidimensional). De cualquier manera, el estimador $\hat{\theta}_\alpha$ definido en (4.27) se reduce a la constante θ y no representa realmente un estimador.

4.5. Muestreo estratificado

Consideremos nuevamente el problema de estimar por simulación una integral de la siguiente forma

$$\theta = \int_D g(x)\, f_X(x)\, dx = E(g(X)), \tag{4.29}$$

en donde X es una variable aleatoria con valores en la región $D \subseteq \mathbb{R}$ y con función de densidad $f(x)$, y $g(x)$ es una función integrable con dominio el rango de valores de X. Si $x_1, \ldots, x_n$ son valores de la distribución $f(x)$, entonces el estimador Monte Carlo para θ es

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

Sea $D_1, \ldots, D_m$ una partición de la región D. Véase la Figura 4.1.

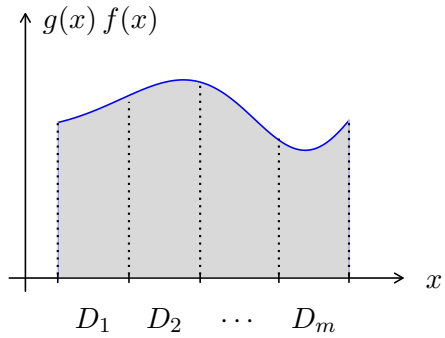


Figura 4.1: Partición.

El objetivo no es seleccionar una función de densidad $f_X(x)$ pues, en este caso, la función ya está especificada y es parte de los datos del problema. Lo que se busca ahora es tomar más muestras en aquellos subconjuntos o estratos D_i que son más importantes en el sentido que explicaremos más adelante. Definamos las siguientes cantidades: para $i = 1, \ldots, m$,

$$\theta_i := \int_{D_i} g(x)\, f_X(x)\, dx,$$

$$p_i := \int_{D_i} f_X(x)\, dx,$$

$$g_i(x) := \begin{cases} g(x) & \text{si } x \in D_i, \\ 0 & \text{en otro caso.} \end{cases}$$

Estos son los elementos de nuestro interés restringidos a cada región D_i de la partición. Es evidente que $\theta = \theta_1 + \cdots + \theta_m$ y puede comprobarse que

$$\theta_i = p_i\, E(g_i(X_i)),$$

en donde X_i es una variable aleatoria con función de densidad

$$f_{X_i}(x) = \begin{cases} f(x)/p_i & \text{si } x \in D_i, \\ 0 & \text{en otro caso.} \end{cases}$$

Sea n_i el número de muestras que se toman del subconjunto D_i. Supongamos que el total de muestras es n, es decir, $n = n_1 + \cdots + n_m$. Tomando ahora una muestra $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ de la distribución $f_X(x)/p_i$ sobre D_i, se pueden definir los estimadores Monte Carlo:

$$\hat{\theta}_i := \frac{p_i}{n_i} \sum_{k=1}^{n_i} g(X_k^{(i)}), \tag{4.30}$$

$$\hat{\theta} := \sum_{i=1}^{m} \hat{\theta}_i. \tag{4.31}$$

Denotando por σ_i^2 a $\mathrm{Var}(g(X_1^{(i)}))$, a partir de las expresiones para $\hat{\theta}$ y $\hat{\theta}_i$ se puede comprobar que

$$\mathrm{Var}(\hat{\theta}) = \sum_{i=1}^{m} \frac{p_i^2\, \sigma_i^2}{n_i}. \tag{4.32}$$

Esta cantidad cambia según los valores que se asignen a los tamaños de muestra locales $n_1, \ldots, n_m$. Ahora podemos establecer el objetivo del muestreo estratificado: se busca calcular los tamaños de muestra locales n_i tales que la varianza (4.32) sea mínima, sujeto a la condición $n_1 + \cdots + n_m = n$. El resultado es el siguiente.

Proposición 4.6 (Muestreo estratificado) *Bajo el esquema de estratificación para estimar la cantidad θ definida en* (4.29)*, el número de muestras n_i que deben tomarse de la región D_i para que la varianza (4.32) del estimador (4.31) sea mínima es*

$$n_i = n\, p_i\, \sigma_i \,/ \sum_{j=1}^{m} p_j\, \sigma_j, \quad i = 1, 2, \ldots, m.$$

En tal caso, el valor mínimo de la varianza es $(1/n)(\sum_{j=1}^{m} p_j\, \sigma_j)^2$.

Demostración. Se considera artificialmente que las cantidades n_i son números reales y se aplica el método de multiplicadores de Lagrange para minimizar (4.32). Así, se define la función

$$J = \sum_{i=1}^{m} \frac{p_i^2\,\sigma_i^2}{n_i} + \lambda \left(\sum_{i=1}^{m} n_i - n \right),$$

en donde las variables son $n_1, \ldots, n_m$. Derivando respecto de n_i, igualando a cero y resolviendo para n_i se encuentra que

$$n_i = \frac{p_i\,\sigma_i}{\sqrt{\lambda}}. \tag{4.33}$$

Substituyendo en la restricción del problema se encuentra que el valor del multiplicador es

$$\lambda = \frac{1}{n^2} \left(\sum_{i=1}^{m} p_i\,\sigma_i \right)^2.$$

Finalmente, substituyendo este valor de λ en (4.33) se llega a la expresión para n_i que aparece en el enunciado. Al evaluar la varianza (4.32) en estos tamaños de muestra locales óptimos n_i se llega a la expresión para la varianza mínima. ∎

Debe observarse que el método de muestreo estratificado no se puede aplicar directamente sin antes determinar ciertos parámetros. Primero, deben encontrarse los valores de σ_i^2, $i = 1, \ldots, m$, y segundo, es necesario contar con un tamaño m de partición preestablecido y con la partición misma $D_1, \ldots, D_m$. Por otro lado, la fórmula encontrada para n_i no produce necesariamente un número entero, de modo que un redondeo al entero más cercano debe efectuarse, lo que puede modificar ligeramente el tamaño de muestra total n. También es interesante observar que n_i es más grande conforme la probabilidad p_i es más grande (esto es el muestreo por importancia a nivel de regiones de la partición), o bien cuando la varianza σ_i^2 es más grande.

4.6. Ejercicios

Muestreo por importancia

94. Elabore un programa de cómputo para aproximar la integral θ que aparece abajo usando muestreo por importancia. Utilice como función

de densidad las funciones $f(x)$ indicadas en los incisos.

$$\theta = \int_0^{\pi/2} \operatorname{sen}^2 x \, dx.$$

a) $f(x) = 2/\pi \quad \text{si} \quad 0 < x < \pi/2.$

b) $f(x) = 8x/\pi^2 \quad \text{si} \quad 0 < x < \pi/2.$

95. Sea $X \sim \text{N}(0, 1)$. Estime $\theta = P(Z > 2)$ mediante muestreo por importancia utilizando la función de densidad exponencial trasladada $f(x)$ que aparece abajo. Tome un tamaño de muestra suficientemente grande para obtener precisión de dos dígitos. Compare con el valor que se obtiene de una tabla de probabilidades de la distribución normal.

$$f(x) = \begin{cases} e^{-(x-2)} & \text{si} \quad x > 2, \\ 0 & \text{en otro caso.} \end{cases}$$

Muestreo condicional

96. Sea (X, Y) un vector aleatorio con distribución unif$(-1, 1) \times (-1, 1)$. Use muestreo condicional para encontrar una aproximación a las probabilidades que aparecen abajo. Calcule el valor exacto de estas probabilidades y compruebe si las aproximaciones obtenidas son razonables.

a) $P(X + Y > 0)$.

b) $P(X - Y > 0)$.

c) $P(2X + Y < 0)$.

d) $P(4X - 2Y > 0)$.

e) $P(X^2 + Y^2 < 1)$.

f) $P(X^2/3 + Y^2/2 < 1)$.

97. Sea (X, N) un vector aleatorio discreto con función de probabilidad $f(x, n)$ como aparece abajo y suponga que X_1 y X_2 son dos copias independientes de X.

$$f(x, n) = \frac{3}{10}\left(\frac{1}{2}\right)^{nx}, \quad n = 1, 2; \; x = 0, 1, \ldots$$

a) Usando muestreo condicional sobre N, encuentre una aproximación a la esperanza de la variable aleatoria S indicada abajo.

$$S = \sum_{i=1}^{N} X_i.$$

b) Encuentre el valor exacto de $E(S)$ y compruebe si la aproximación obtenida es razonable.

98. Defina la variable aleatoria $S := X^N$, en donde X y N son independientes y son tales que $X \sim \text{unif}(0,1)$ y $N \sim \text{unif}\{1,\ldots,10\}$.

a) Usando muestreo condicional sobre N, encuentre una aproximación para $E(S)$.

b) Encuentre el valor exacto de $E(S)$ y compruebe si la aproximación obtenida es razonable.

Variables comunes y antitéticas

99. Sean X y Y dos variables aleatorias con función de distribución $F(x)$ y $G(y)$, respectivamente. Sea $U \sim \text{unif}\,(0,1)$. Demuestre que:

a) $\text{Cov}(F^{-1}(U), G^{-1}(U)) \geqslant 0$.

b) $\text{Cov}(F^{-1}(U), G^{-1}(1-U)) \leqslant 0$.

100. *Distribución conjunta.* Sean X y Y dos variables aleatorias con funciones de distribución invertibles $F(x)$ y $G(y)$, respectivamente. Suponga que X y Y pueden ser representadas mediante una variable común $U \sim \text{unif}\,(0,1)$, es decir, $X \stackrel{d}{=} F^{-1}(U)$ y $Y \stackrel{d}{=} G^{-1}(U)$.

a) Demuestre que la distribución conjunta de X y Y es

$$F(x,y) = \text{mín}\,\{F(x), G(y)\}.$$

b) Demuestre que $F(x,y)$ es, efectivamente, una función de distribución bivariada, es decir, compruebe que se cumplen las propiedades de la Definición A.7 que aparece en la página 282.

c) A partir de $F(x,y)$, demuestre que las distribuciones marginales son, efectivamente, $F(x)$ y $G(y)$.

101. *Distribución conjunta.* Sean X y Y dos variables aleatorias con funciones de distribución invertibles $F(x)$ y $G(y)$, respectivamente. Suponga que X y Y pueden ser representadas mediante variables antitéticas, es decir, $X \stackrel{d}{=} F^{-1}(U)$ y $Y \stackrel{d}{=} G^{-1}(1-U)$, en donde $U \sim \text{unif}\,(0,1)$.

a) Demuestre que la distribución conjunta de X y Y es

$$F(x,y) = \begin{cases} F(x) + G(y) - 1 & \text{si } (x,y) \text{ es tal que} \\ & F(x) \geqslant 1 - G(y), \\ 0 & \text{en otro caso.} \end{cases}$$

b) Demuestre que $F(x,y)$ es, efectivamente, una función de distribución bivariada, es decir, compruebe que se cumplen las propiedades de la Definición A.7 que aparece en la página 282.

c) A partir de $F(x,y)$, demuestre que las distribuciones marginales son, efectivamente, $F(x)$ y $G(y)$.

Variables de control

102. En ocasiones el estimador $\hat{\theta}_\alpha$ que aparece en la Definición 4.6 se define con signo cambiado como aparece en la expresión abajo indicada. Bajo este cambio, compruebe que el valor mínimo para $\text{Var}(\hat{\theta}_\alpha)$ se alcanza en $\alpha^* = -\text{Cov}(\hat{\theta}, C)/\text{Var}\,(C)$.

$$\hat{\theta}_\alpha := \hat{\theta} + \alpha(C - E(C)). \tag{4.34}$$

103. *Tranformación lineal de una variable de control.* Sea $\hat{\theta}$ un estimador insesgado para θ y sea C una variable de control. Sea α^* la constante que aparece en la Proposición 4.4 que hace que el estimador $\hat{\theta}_\alpha$ definido en (4.26) tenga varianza mínima. Sea $a \neq 0$ una constante.

a) Sea α_a^* el valor óptimo para α cuando se toma $C + a$ en lugar de C. Compruebe que $\alpha_a^* = \alpha^*$ y $\text{Var}(\hat{\theta}_{\alpha_a^*}) = \text{Var}(\hat{\theta}_\alpha^*)$.

b) Sea α_a^* el valor óptimo para α cuando se toma aC en lugar de C. Compruebe que $\alpha_a^* = \alpha^*/a$ y $\text{Var}(\hat{\theta}_{\alpha_a^*}) = \text{Var}(\hat{\theta}_\alpha^*)$.

104. Suponga que se desea usar el método de Monte Carlo para aproximar la integral

$$\theta := \int_0^1 g(x)\,dx, \quad \text{con} \;\; g(x) = \frac{1}{1+x}.$$

Si $X_1, \ldots, X_n$ es una muestra aleatoria de la unif$(0, 1)$, el estimador Monte Carlo clásico para θ es

$$\hat{\theta} := \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Defina la variable de control

$$C := \frac{1}{n} \sum_{i=1}^{n} h(X_i), \quad \text{con} \quad h(x) = 1 + x.$$

a) Genere $n = 500$ valores $x_1, \ldots, x_n$ de la distribución unif$(0, 1)$. Calcule la estimación Monte Carlo clásico $\hat{\theta}$ y $\text{Var}(\hat{\theta})$.

b) Encuentre o aproxime el valor óptimo α^*. Calcule el estimador $\hat{\theta}_{\alpha*}$ y encuentre $\text{Var}(\hat{\theta}_{\alpha*})$.

c) Compruebe que $\text{Var}(\hat{\theta}_{\alpha*}) \leqslant \text{Var}(\hat{\theta})$.

d) Encuentre el valor exacto de θ.

Muestreo estratificado

105. Esta es una situación simple en la que se busca ilustrar el muestreo estratificado. Sea X una variable aleatoria con función de densidad

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

Es inmediato comprobar que $E(X) = 2/3$. Para obtener una aproximación a este valor por simulación usando muestreo estratificado considere la siguiente partición del intervalo unitario $(0, 1)$:

$$D_1 = (0, 1/3), \quad D_2 = (1/3, 2/3), \quad D_3 = (2/3, 1).$$

a) Suponiendo $n = 90$, encuentre los tamaños de muestra (valores enteros aproximados) n_1, n_2 y n_3 que deben tomarse de cada uno de los elementos de la partición siguiendo el esquema del muestreo estratificado, en donde $n_1 + n_2 + n_3 = n$.

b) Genere n_1 valores al azar $x_1, \ldots, x_{n_1}$ en $(0, 1/3)$. Genere n_2 valores al azar $x_{n_1+1}, \ldots, x_{n_1+n_2}$ en $(1/3, 2/3)$. Genere n_3 valores al azar $x_{n_1+n_2+1}, \ldots, x_n$ en $(2/3, 1)$.

c) Con los datos obtenidos en el inciso anterior, encuentre el estimador de Monte Carlo $\bar{x}$ y calcule la varianza muestral

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

d) Tome $n_1 = n_2 = n_3 = 30$, genere 90 valores $u_1, \ldots, u_{90}$ con esta distribución de valores en cada región de la partición, calcule la media muestral $\bar{u}$ y la varianza muestral

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2.$$

e) Verifique si se cumple la relación $s_x^2 \leqslant s_u^2$.

Miscelánea

106. Suponga que desea mejorar la precisión de la estimación del valor de π usando simulación. Aplicando el método adecuado, utilice la función $f(x) = 4/(1+x^2)$ para reducir la varianza de la estimación.

107. Suponga que se desea estimar

$$\theta = \int_0^1 e^{x^2} dx.$$

Sean U, U_1 y U_2 variables aleatorias independientes cada una con distribución unif $(0,1)$. Defina los estimadores $\hat{\theta_1}$ y $\hat{\theta_2}$ como aparecece abajo. Compruebe que ambos estimadores de θ son insesgados y calcule la varianza de cada uno para determinar el que es preferible.

$$\hat{\theta_1} = \frac{e^{U^2}(1 + e^{1-2U})}{2} \quad \text{y} \quad \hat{\theta_2} = \frac{e^{U_1^2} + e^{U_2^2}}{2}.$$

108. Considere la integral

$$\theta = \int_0^1 4x^3 dx.$$

a) Estime θ usando el método de integración Monte Carlo clásico.

b) Utilizando alguna variable antitética, realice una estimación de θ con menor varianza.

c) Utilizando estratificación, realice otra estimación de θ.

d) Utilizando una variable de control, realice de nuevo una estimación de θ.

109. Para estimar
$$\theta = \int_b^\infty x^{\alpha-1} e^{-x} dx,$$
en donde $\alpha \leqslant 1$ y $b > 0$, se utiliza una función de densidad de muestreo por importancia $f(x) = e^{-(x-b)} 1_{(x>b)}$. Suponga que $U \sim \text{unif}\,(0,1)$ y defina
$$\hat{\theta} = X^{\alpha-1}\, e^{-b},$$
en donde $X := b - \ln U$. Demuestre que $\hat{\theta}$ es un estimador insesgado para θ y que $\text{Var}(\hat{\theta}) < \theta(b^{\alpha-1}e^{-b} - \theta)$.

110. Realice la aproximación a la siguiente integral utilizando los distintos métodos que se han expuesto en este capítulo. En cada caso, calcule la varianza de la estimación.
$$\int_0^1 \frac{e^x - 1}{e - 1} dx.$$

111. *Aplicación a finanzas.* Suponga que se tiene un producto financiero llamado "dólar ligero" que el día de hoy tiene precio P y en 6 meses cambia su valor según el siguiente mecanismo:

- Si en 6 meses el precio de un dólar (US) está por arriba de 18 pesos (MX), el dueño del dólar ligero recibe sólo 18 si decide venderlo.
- Si en 6 meses el precio de un dólar (US) es menor o igual a 18 pesos (MX), el dueño del dólar ligero recibe 20 pesos si decide venderlo.

Suponga que $D = \sqrt{2X^2 + 1}$, en donde $X \sim \text{N}\,(3,1)$, y que el precio inicial del dólar ligero se calcula según la siguiente fórmula:
$$P = (18 \cdot P(D > 18) + 20 \cdot P(D \leqslant 18)) \exp\{-0.05\}.$$
Aproxime el valor de P usando simulación.

Capítulo 5

Monte Carlo vía cadenas de Markov

Los métodos de Monte Carlo vía cadenas de Markov, referidos como métodos MCMC (*Markov Chains Monte Carlo*), son procedimientos para obtener muestras aleatorias aproximadas de distribuciones arbitrarias y hacen uso de la teoría de las cadenas de Markov.

El algoritmo original fue propuesto por Nicholas Metropolis y otros autores [39] en el año de 1953, cuando éstos se encontraban estudiando algunos problemas computacionales de la física estadística. Existen varias modificaciones y extensiones al método original planteado por Metropolis *et al.* Para referirse a estos procedimientos se usa también el nombre "muestreo por cadenas de Markov" (*Markov chain sampling*). El nombre proviene del hecho de que el método requiere la construcción de una cadena de Markov cuya distribución límite es la distribución de la que se desea obtener la muestra aleatoria. Los dos algoritmos MCMC más utilizados y que estudiaremos son:

- Algoritmo Metropolis-Hastings.
- Muestreo de Gibbs.

El muestreo de Gibbs es particularmente útil en la estadística bayesiana. Iniciaremos explicando el algoritmo Metropolis-Hastings para distribuciones discretas y después pasaremos al caso continuo.

5.1. Conceptos de cadenas de Markov

En esta sección recordaremos algunos conceptos y resultados sobre cadenas de Markov a tiempo discreto con espacios de estados discreto. Una exposición más detallada de este tema, así como la demostración de los resultados que mencionaremos, se puede encontrar en el excelente texto de Karlin y Taylor [30].

Definición 5.1 *Una cadena de Markov es un proceso estocástico a tiempo discreto* $\{X_t : t = 0, 1, \ldots\}$ *con espacio de estados también discreto y que satisface la propiedad de Markov, esto es, para cualquier tiempo* $t \geqslant 1$ *y para cualesquiera estados* $x_0, \ldots, x_t, x_{t+1}$,

$$P(X_{t+1} = x_{t+1} \,|\, X_0 = x_0, \ldots, X_t = x_t) = P(X_{t+1} = x_{t+1} \,|\, X_t = x_t). \tag{5.1}$$

A la identidad (5.1) se le llama propiedad de Markov. Esta condición establece que la probabilidad del evento futuro $(X_{t+1} = x_{t+1})$ depende del pasado o historia $(X_0 = x_0, \ldots, X_t = x_t)$ sólo a través del presente $(X_t = x_t)$ o del pasado más próximo. Sin pérdida de generalidad tomaremos como espacio de estados el conjunto discreto $S = \{0, 1, 2, \ldots\}$. En particular, una cadena de Markov es finita cuando el espacio de estados es finito, por ejemplo, el conjunto $S = \{0, 1, \ldots, k\}$ es un espacio de estados finito.

Sea $t \geqslant 0$ y sean i y j dos estados cualesquiera de una cadena de Markov. La probabilidad condicional $P(X_{t+1} = j \,|\, X_t = i)$ es la probabilidad de transición del estado i al tiempo t, al estado j al tiempo siguiente $t+1$. Se le denota también por $p_{ij}(t, t+1)$ y se le llama probabilidad de transición en un paso. Cuando las probabilidades $p_{ij}(t, t+1)$ no dependen de t, se dice que la cadena es estacionaria u homogénea en el tiempo. Regularmente se estudian las cadenas de Markov bajo esta hipótesis y las probabilidades de transición en un paso se escriben como p_{ij}. Haciendo variar los subíndices i y j en el espacio de estados $S = \{0, 1, 2, \ldots\}$ se obtiene la matriz de probabilidades

de transición en un paso

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & \cdots \\ p_{10} & p_{11} & \cdots \\ \vdots & \vdots & \end{pmatrix}, \tag{5.2}$$

en donde los renglones y columnas se etiquetan con los valores de los estados $0, 1, \ldots$ de tal forma que p_{ij} es la entrada (i, j) de esta matriz, la cual se escribe $\mathbf{P} = (p_{ij})$ y cumple las siguientes dos propiedades:

- $p_{ij} \geqslant 0$.
- $\sum_{j=0}^{\infty} p_{ij} = 1$.

Recíprocamente, a toda matriz cuadrada que satisfaga las dos propiedades anteriores se le llama matriz estocástica.

Una distribución inicial es cualquier distribución para la variable aleatoria inicial X_0 de una cadena de Markov. Esta es una distribución sobre el espacio de estados $S = \{0, 1, 2, \ldots\}$. En ocasiones se establece que la cadena inicia en el estado $x_0 \in S$, es decir, $X_0 = x_0$. Esto significa que la distribución inicial está completamente concentrada en el punto x_0. Una distribución inicial para X_0 y una matriz estocástica $\mathbf{P}$ como aparece en (5.2) determinan por completo a la cadena de Markov. Esta es la razón por la que, a veces, a la matriz (5.2) se le llama cadena de Markov.

Dinámica puntual

Una cadena de Markov $\mathbf{P} = (p_{ij})$ determina una dinámica estocástica en el espacio de estados $S = \{0, 1, 2, \ldots\}$ de la siguiente manera: Si i es un estado inicial, a partir de él se genera el siguiente estado de acuerdo a la distribución dada por el vector de probabilidad $(p_{i\,0}, p_{i\,1}, \ldots)$. Este es el $i-$ésimo renglón de la matriz $\mathbf{P}$. Supongamos que el valor obtenido es j. A partir de la distribución dada por el $j-$ésimo renglón de $\mathbf{P}$, esto es $(p_{j\,0}, p_{j\,1}, \ldots)$, se genera el siguiente valor. Supongamos que el valor es k. Se procede de manera sucesiva y se obtiene así una secuencia de puntos en el espacio de estados

S. A una sucesión de estos puntos en el espacio de estados se le llama una trayectoria o realización de la cadena de Markov.

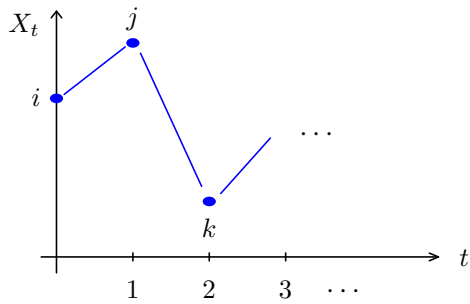


Figura 5.1: Una trayectoria de una cadena de Markov.

Probabilidades de transición en n pasos

Sean $n \geqslant 1$ y $m \geqslant 0$ números enteros, a las cantidades denotadas por $P(X_{n+m} = j \,|\, X_m = i)$, se les llama probabilidades de transición del estado i al estado j en n pasos. Dado que hemos supuesto la condición de homogeneidad en el tiempo, esta probabilidad no depende del tiempo m y coincide con $P(X_n = j \,|\, X_0 = i)$. Haciendo variar los valores de i y de j en el espacio de estados se obtiene la matriz de probabilidades de transición en n pasos,

$$\mathbf{P}(n) = \begin{pmatrix} p_{00}(n) & p_{01}(n) & \cdots \\ p_{10}(n) & p_{11}(n) & \cdots \\ \vdots & \vdots & \end{pmatrix}. \tag{5.3}$$

Cuando $n = 1$ se obtiene la matriz $\mathbf{P}$ como en (5.2), y para $n = 0$ se define $\mathbf{P}(0)$ como la matriz identidad, es decir,

$$\mathbf{P}(0) = (p_{ij}(0)) = \delta_{ij} = \begin{cases} 0 & \text{si } i \neq j, \\ 1 & \text{si } i = j. \end{cases}$$

A pesar de que obtener las probabilidades de transición en n pasos puede parecer un problema muy complicado, los cálculos se clarifican debido al

resultado que aparece abajo. Su demostración está basada en condicionar con respecto al valor de la cadena al tiempo r y aplicar la propiedad de Markov.

Proposición 5.1 *(Ecuación de Chapman-Kolmogorov)*
Sean $0 \leqslant r \leqslant n$ dos números enteros. Las probabilidades de transición $p_{ij}(n)$ de una cadena de Markov satisfacen

$$p_{ij}(n) = \sum_{k=0}^{\infty} p_{ik}(r) \cdot p_{kj}(n-r).$$

Aplicando este resultado en $n-1$ ocasiones usando $r = 1$, se encuentra que $p_{ij}(n)$ resulta ser la entrada (i,j) del producto $\mathbf{P} \cdots \mathbf{P} = \mathbf{P}^n$, es decir,

$$p_{ij}(n) = (\underbrace{\mathbf{P} \cdots \mathbf{P}}_{n \text{ veces}})_{ij} = (\mathbf{P}^n)_{ij}.$$

De forma explícita,

$$\begin{pmatrix} p_{00}(n) & p_{01}(n) & \cdots \\ p_{10}(n) & p_{11}(n) & \cdots \\ \vdots & \vdots & \end{pmatrix} = \begin{pmatrix} p_{00} & p_{01} & \cdots \\ p_{10} & p_{11} & \cdots \\ \vdots & \vdots & \end{pmatrix}^n.$$

Cuando se conoce la matriz potencia $\mathbf{P}^n$ y se tiene una distribución inicial denotada por $p_i := P(X_0 = i)$, entonces la distribución de la variable X_n se puede calcular mediante la siguiente expresión,

$$P(X_n = j) = \sum_{i=0}^{\infty} p_i \cdot p_{ij}(n), \quad j = 0, 1, \ldots \tag{5.4}$$

Dinámica sobre el espacio de distribuciones

Una cadena de Markov $\mathbf{P} = (p_{ij})$ establece una dinámica sobre el conjunto de todas las distribuciones de probabilidad definidas sobre el espacio de estados $S = \{0, 1, 2, \ldots\}$ de la siguiente manera:

- Sea $\pi^{(\mathbf{0})} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}, \ldots)$ una distribución de probabilidad inicial. Esta es la distribución de la variable aleatoria inicial X_0.

- La fórmula (5.4) con $n = 1$ establece la forma de encontrar la distribución de la variable aleatoria X_1, esto es,

$$P(X_1 = j) = \sum_{i=0}^{\infty} \pi_i^{(0)} \cdot p_{ij},$$

para valores de $j = 0, 1, \ldots$ Esta es la distribución de la variable aleatoria X_1 y puede ser denotada por $\pi^{(\mathbf{1})} = (\pi_0^{(1)}, \pi_1^{(1)}, \pi_2^{(1)}, \ldots)$, es decir, la igualdad anterior se puede escribir como

$$\pi^{(\mathbf{1})} = \pi^{(\mathbf{0})} \cdot \mathbf{P}.$$

- De manera análoga se puede calcular la distribución de X_2,

$$\pi^{(\mathbf{2})} = \pi^{(\mathbf{1})} \cdot \mathbf{P} = \pi^{(\mathbf{0})} \cdot \mathbf{P}^2.$$

- En general, la distribución de la variable X_n es

$$\pi^{(\mathbf{n})} = \pi^{(\mathbf{n-1})} \cdot \mathbf{P} = \pi^{(\mathbf{0})} \cdot \mathbf{P}^n, \qquad n = 1, 2, \ldots$$

Con el procedimiento anterior queda definida la sucesión de distribuciones de probabilidad $\pi^{(\mathbf{0})}, \pi^{(\mathbf{1})}, \pi^{(\mathbf{2})}, \ldots$ Estaremos interesados en la convergencia de esta sucesión. Mencionaremos condiciones bajo las cuales la convergencia se verifica. Observemos que la convergencia de $\pi^{(\mathbf{n})}$ es la convergencia de las potencias $\mathbf{P}^n$.

Comunicación

Se dice que el estado j es accesible desde el estado i si existe un número entero $n \geqslant 0$ tal que $p_{ij}(n) > 0$. Esta definición establece que el estado j es accesible desde i si, con probabilidad positiva, existe alguna trayectoria de i a j sin importar el número de pasos necesario. Esto se representa por el símbolo $i \to j$. Véase la Figura 5.2. Por otro lado, se dice que los estados i y j son comunicantes si se cumple al mismo tiempo que $i \to j$ y $j \to i$, no necesariamente en el mismo número de pasos cada transición. La comunicación entre dos estados se denota por $i \leftrightarrow j$. Véase la Figura 5.2.

Figura 5.2: Accesibilidad y comunicación.

Definición 5.2 *Una cadena de Markov es irreducible si todos sus estados se comunican entre sí.*

La comunicación entre estados es una relación de equivalencia e induce una partición en el espacio de estados de la cadena de Markov. Los elementos de la partición son las así llamadas clases de comunicación: dos estados pertenecen a la misma clase de comunicación si se comunican entre ellos. En el caso de una cadena irreducible, sólo hay una clase de comunicación dada por el espacio completo de estados, ya que todos los estados se comunican.

Periodo

Este es un número entero no negativo que se calcula para cada estado i de una cadena de Markov. Se denota por $d(i)$ y se define como

$$d(i) := \text{mcd}\,\{n \geqslant 1 : p_{i\,i}(n) > 0\}\,,$$

en donde m.c.d. significa máximo común divisor. Se define $d(i) = 0$ cuando ocurre que $p_{ii}(n) = 0$ para toda $n \geqslant 1$.

Definición 5.3 *Se dice que un estado i es*

1. *Periódico de periodo k, cuando $d(i) = k \geqslant 2$.*
2. *Aperiódico cuando $d(i) = 1$.*

Más aún, una cadena de Markov es aperiódica si todos sus estados lo son. Puede comprobarse que todos los estados dentro de una clase de comunicación tienen el mismo periodo.

Recurrencia y transitoriedad

Un estado i de una cadena de Markov $\{X_t : t = 0, 1, \ldots\}$ es recurrente si la probabilidad de eventualmente regresar al estado i, partiendo de i es uno, es decir, si

$$P(X_t = i \text{ para alguna } t \geqslant 1 \,|\, X_0 = i) = 1.$$

A un estado que no es recurrente se le llama transitorio. En este caso, la probabilidad arriba indicada es estrictamente menor a 1. De esta manera, cada estado de una cadena de Markov se clasifica en una de estas dos categorías: es recurrente o es transitorio.

La recurrencia y la transitoriedad son propiedades que comparten todos los estados en una misma clase de comunicación, es decir, si i es un estado recurrente e $i \leftrightarrow j$, entonces j también es un estado recurrente. Como la transitoriedad es la contraparte de la recurrencia, tenemos también el resultado complementario: si el estado i es un estado transitorio e $i \leftrightarrow j$, entonces j también es un estado transitorio. En particular, cuando una cadena es irreducible y algún estado es recurrente, todos los estados son recurrentes. Para el caso de cadenas con un número finito de estados, existe por lo menos un estado que es recurrente. Una cadena es recurrente o transitoria cuando todos sus estados lo son. Observe que no se necesita que la cadena sea irreducible.

Recurrencia positiva y nula

Para dos estados i y j de una cadena de Markov $\{X_t : t = 0, 1, \ldots\}$ se puede definir la variable aleatoria

$$\tau_{ij} = \min\{t \geqslant 1 : X_t = j \,|\, X_0 = i\}.$$

Este es el tiempo de la primera visita al estado j a partir del estado i, y puede tomar el valor infinito. En el caso cuando el estado j es recurrente, se define el tiempo medio de recurrencia del estado j a partir del estado i como

$$\mu_{ij} = E(\tau_{ij}).$$

Cuando $i = j$ es un estado recurrente sólo se escribe τ_j en lugar de τ_{jj} y su tiempo medio de recurrencia se escribe como μ_j. Este número puede ser

finito o infinito y representa el número de pasos promedio que le toma a la cadena regresar al estado recurrente j.

Definición 5.4 *Un estado recurrente j es*

1. *Recurrente positivo si $\mu_j < \infty$.*
2. *Recurrente nulo si $\mu_j = \infty$.*

En consecuencia, todo estado recurrente se clasifica en una de estas dos categorías: es recurrente positivo o es recurrente nulo. La recurrencia positiva y la recurrencia nula son propiedades que comparten todos los elementos de una misma clase de comunicación, es decir, si i es un estado recurrente positivo e $i \leftrightarrow j$, entonces j también es recurrente positivo. Por complemento, si i es un estado recurrente nulo e $i \leftrightarrow j$, entonces j es recurrente nulo.

Distribuciones estacionarias

Una distribución de probabilidad dada por el vector $\pi = (\pi_0, \pi_1, \ldots)$ es estacionaria o invariante en el tiempo para una cadena de Markov $\mathbf{P} = (p_{ij})$ si

$$\pi_j = \sum_{i=0}^{\infty} \pi_i \cdot p_{ij}, \quad j = 0, 1, \ldots$$

En notación matricial, $\pi = \pi \cdot \mathbf{P}$. De donde se sigue que $\pi = \pi \cdot \mathbf{P}^n$. Esto significa que si la variable aleatoria inicial X_0 tiene esa distribución π, entonces para cualquier tiempo $n \geqslant 1$ la variable X_n también tiene distribución π, pues

$$P(X_n = j) = \sum_{i=0}^{\infty} \pi_i \cdot p_{ij}(n) = \pi_j.$$

Para una cadena de Markov dada, se pueden presentar sólo tres situaciones: puede no existir ninguna distribución estacionaria, o bien puede existir una única distribución estacionaria, o bien puede existir una infinidad de distribuciones estacionarias. Puede demostrarse que si $\pi = (\pi_0, \pi_1, \ldots)$ es una distribución estacionaria para una cadena de Markov, entonces si j es un estado recurrente positivo, se cumple que $\pi_j > 0$, y si j es recurrente

nulo o transitorio, entonces $\pi_j = 0$. Es decir, toda distribución estacionaria tiene como soporte el conjunto de todos los estados recurrentes positivos.

Proposición 5.2 *Toda cadena de Markov irreducible y recurrente positiva tiene una única distribución estacionaria dada por*

$$\pi_j = \frac{1}{\mu_j} > 0, \quad j = 0, 1, \ldots,$$

en donde μ_j es el tiempo medio de recurrencia del estado j.

En particular, toda cadena de Markov finita e irreducible tiene una única distribución estacionaria.

Definición 5.5 *Una cadena de Markov $\mathbf{P} = (p_{ij})$ es reversible respecto de una distribución de probabilidad $\pi = (\pi_0, \pi_1, \ldots)$ si se cumple la siguiente identidad llamada ecuación de balance,*

$$\pi_i \cdot p_{ij} = \pi_j \cdot p_{ji}, \quad \text{para} \quad i, j = 0, 1, \ldots$$

Es inmediato comprobar que tal distribución π es estacionaria pues para cualquier $j = 0, 1, \ldots$ se tiene

$$\sum_{i=0}^{\infty} \pi_i \cdot p_{ij} = \sum_{i=0}^{\infty} \pi_j \cdot p_{ji} = \pi_j \cdot \sum_{i=0}^{\infty} p_{ji} = \pi_j.$$

Distribuciones límite

Una cadena de Markov es un sistema dinámico que evoluciona en el tiempo de manera aleatoria. A largo plazo, este sistema puede presentar algún tipo de estabilización en su comportamiento. Sean $p_{ij}(n)$ las probabilidades de transición en n pasos de una cadena de Markov. Cuando el valor de n es muy grande, por la propiedad de Markov, la probabilidad $p_{ij}(n)$ podría no depender del estado inicial i y sólo depender del estado final j. Puede

demostrarse que, cuando las probabilidades límite

$$\pi_j = \lim_{n\to\infty} p_{ij}(n), \quad j = 0, 1, \ldots \tag{5.5}$$

existen y no dependen del estado inicial i, entonces se cumple que:

- $\sum_{j=0}^{\infty} \pi_j \leqslant 1.$
- $\pi_j = \sum_{i=0}^{\infty} \pi_i \cdot p_{ij}, \quad j = 0, 1, \ldots$

El primer resultado advierte que el vector límite $(\pi_0, \pi_1, \ldots)$ podría no ser una distribución de probabilidad genuina. A pesar de esto, a este vector límite se le llama distribución límite, por supuesto bajo la hipótesis de la existencia de los límites (5.5). Cuando el espacio de estados es finito, se puede comprobar que la desigualdad que aparece en el primer resultado se convierte en igualdad y, en este caso, la distribución límite es verdaderamente una distribución de probabilidad. El segundo resultado establece que el vector límite $(\pi_0, \pi_1, \ldots)$ es una "distribución" estacionaria.

En términos matriciales,

$$\lim_{n\to\infty} \mathbf{P}^n = \lim_{n\to\infty} \begin{pmatrix} p_{00}(n) & p_{01}(n) & \cdots \\ p_{10}(n) & p_{11}(n) & \cdots \\ \vdots & \vdots & \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 & \cdots \\ \pi_0 & \pi_1 & \cdots \\ \vdots & \vdots & \end{pmatrix}.$$

Observe que la matriz límite tiene todos sus renglones idénticos a la distribución límite.

El siguiente resultado establece condiciones que garantizan la existencia de la distribución límite. Observe que dentro de estas condiciones se encuentra la existencia de una única distribución estacionaria. Este resultado particular nos será de utilidad más adelante.

Teorema 5.1 *Sea* $\mathbf{P} = (p_{ij})$ *una cadena de Markov que es irreducible, aperiódica y con distribución estacionaria* π*. Para* $j = 0, 1, \ldots$

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j.$$

Omitiendo la hipótesis sobre la distribución estacionaria, el siguiente resultado establece condiciones suficientes que garantizan la existencia de las probabilidades límite en una cadena de Markov, y establece también que esa distribución límite es la única distribución estacionaria de la cadena.

Teorema 5.2 *Sea* $\mathbf{P} = (p_{ij})$ *una cadena de Markov que es*

1. *Irreducible,*
2. *Recurrente positiva y*
3. *Aperiódica.*

Entonces las probabilidades límite π_j *que aparecen abajo existen y son la única solución al sistema de ecuaciones* $\pi = \pi \cdot \mathbf{P}$ *sujeto las condiciones* $\pi_j \geqslant 0$ *y* $\pi_0 + \pi_1 + \cdots = 1$*,*

$$\pi_j = \lim_{n\to\infty} p_{ij}(n) = \frac{1}{\mu_j}, \quad j = 0, 1, \ldots$$

Con estos elementos podemos ahora explicar el algoritmo Metropolis-Hastings en su versión discreta, el cual sirve a manera de introducción al funcionamiento del algoritmo. Más adelante estudiaremos el caso continuo.

5.2. El algoritmo Metropolis-Hastings: distribuciones discretas

Este algoritmo es uno de los métodos denominados MCMC (*Markov Chain Monte Carlo*) y su objetivo es proveer un mecanismo para generar valores de una distribución dada. Se busca que esta distribución objetivo sea la

distribución límite de una cadena de Markov, de modo que la simulación de valores de la cadena sean valores aproximados de la distribución objetivo. El método fue propuesto por Nicholas Metropolis y otros autores [39] en un artículo publicado en 1953. En este trabajo se considera el caso cuando una cierta función, llamada función propuesta, es simétrica. Años después, en 1970, Wilfred Keith Hastings [19] logró extender el algoritmo para el caso general cuando la función propuesta es arbitraria. Explicaremos primero el caso cuando la distribución objetivo es discreta y después extenderemos las ideas al caso continuo.

El algoritmo propone un mecanismo para obtener valores de una variable aleatoria discreta X que toma los valores $1, \ldots, m$ y con una distribución previamente especificada por las probabilidades estrictamente positivas $\pi_1, \ldots, \pi_m$.

Como elemento inicial supondremos que contamos con una matriz estocástica Q de $m \times m$. Esta matriz representa la función propuesta. Supondremos que conocemos una forma de generar valores de las distribuciones dadas por cada renglón de esta matriz.

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mm} \end{pmatrix}_{m \times m}$$

en donde los renglones y columnas se etiquetan con los valores de los estados $1, 2, \ldots, m$ de tal forma que q_{ij} es la entrada (i, j) de la matriz. Haciendo uso de la matriz Q, se construye una cadena de Markov $\{X_t : t = 0, 1, \ldots\}$ sobre el espacio de estados $\{1, \ldots, m\}$ a través de la siguiente dinámica:

Construcción de la cadena de Markov

Sea $x_0 \in \{1, 2, \ldots, m\}$ un estado inicial.
Para $t = 1, 2, \ldots$

- Se denota por i al estado x_{t-1}.
- Se genera un valor j de la distribución $(q_{i1}, q_{i2}, \ldots, q_{im})$.
- Se calcula la probabilidad $\alpha_{ij} = \text{mín}\left\{\frac{\pi_j}{\pi_i} \cdot \frac{q_{ji}}{q_{ij}}, 1\right\}$.
- Se define

$$x_t = \begin{cases} j & \text{con prob. } \alpha_{ij}, \quad \text{(Aceptación)} \\ i & \text{con prob. } 1 - \alpha_{ij}. \quad \text{(Rechazo)} \end{cases}$$

Es decir, en cada paso, con probabilidad α_{ij} se acepta el valor generado j como el siguiente valor del proceso, y con probabilidad complementaria se rechaza quedando el proceso en el mismo estado que en el tiempo anterior. De esta manera se obtiene una trayectoria $x_0, x_1, x_2, \ldots$ de un proceso $\{X_t : t = 0, 1, 2, \ldots\}$ que, por construcción, es una cadena de Markov. En el siguiente resultado se muestran las probabilidades de transición de la cadena de Markov así construida.

Proposición 5.3 *El proceso $\{X_t : t = 0, 1, \ldots\}$ construido en el recuadro anterior es una cadena de Markov con espacio de estados $\{1, \ldots, m\}$ y matriz de probabilidades de transición*

$$P = (p_{ij}) = \begin{cases} \alpha_{ij} \cdot q_{ij} & si \quad j \neq i, \\ 1 - \sum_{\substack{k=1 \\ k \neq i}}^{m} \alpha_{ik} \cdot q_{ik} & si \quad j = i. \end{cases}$$

Demostración. Por construcción, X_{t+1} depende únicamente del valor de X_t y de la matriz propuesta Q, de modo que la propiedad de Markov se cumple. Por otro lado, las probabilidades de transición son:

a) Para $j \neq i$,

$$\begin{aligned} P(X_{t+1} = j \,|\, X_t = i) &= P(\text{“Generar el valor } j \text{ y aceptarlo”} \,|\, X_t = i) \\ &= q_{ij} \cdot \alpha_{ij}. \end{aligned}$$

b) Para $j = i$,

$$\begin{aligned} P(X_{t+1} = j \,|\, X_t = i) &= P(\text{“Generar el valor } i \text{ y aceptarlo”} \,|\, X_t = i) \\ &\quad + P(\text{“Generar el valor } k \text{ y rechazarlo”} \,|\, X_t = i) \\ &= q_{ii} \cdot \alpha_{ii} + \sum_{k=1}^{m} q_{ik} \cdot (1 - \alpha_{ik}) \\ &= 1 - \sum_{\substack{k=1 \\ k \neq i}}^{m} q_{ik} \cdot \alpha_{ik}. \end{aligned}$$

■

La cadena de Markov recién construida será irreducible y aperiódica dependiendo de la matriz propuesta Q.

Proposición 5.4 *(Ecuación de balance) Para $i, j = 1, 2, \ldots, m$, se cumple la identidad*

$$\pi_i \, p_{ij} = \pi_j \, p_{ji}. \tag{5.6}$$

Demostración. La igualdad es evidente cuando $j = i$. Consideremos el caso $j \neq i$. Tenemos que

$$\begin{aligned} \pi_i \, p_{ij} &= \pi_i \, \alpha_{ij} \, q_{ij} \\ &= \pi_i \, \text{mín} \left\{ \frac{\pi_j \, q_{ji}}{\pi_i \, q_{ij}}, 1 \right\} q_{ij} \\ &= \text{mín} \left\{ \pi_j \, q_{ji}, \pi_i \, q_{ij} \right\} \\ &= \pi_j \, \text{mín} \left\{ 1, \frac{\pi_i \, q_{ij}}{\pi_j \, q_{ji}} \right\} q_{ji} \\ &= \pi_j \, \alpha_{ji} \, q_{ji} \\ &= \pi_j \, p_{ji}. \end{aligned}$$

■

Recordemos que el cumplimiento de la ecuación de balance (5.6) implica que la distribución $(\pi_1, \ldots, \pi_m)$ es estacionaria para la cadena de Markov pues

$$\sum_{i=1}^{m} \pi_i \cdot p_{ij} = \sum_{i=1}^{m} \pi_j \cdot p_{ji} = \pi_j \cdot \sum_{i=1}^{m} p_{ji} = \pi_j.$$

Además es la distribución límite de la cadena cuando ésta es irreducible y aperiódica. A la matriz estocástica $Q = (q_{ij})$, o a la función $(i, j) \mapsto q_{ij}$, se le llama función propuesta, y a la probabilidad α_{ij} se le llama probabilidad de aceptación. Como se ha mencionado antes, originalmente Metropolis *et al.* sugirieron (q_{ij}) simétrica, y Hastings modificó el algoritmo para permitir una matriz Q no necesariamente simétrica.

Ejemplo 5.1 *(Muestreo independiente)*
Tomando $q_{ij} = q_j$ en donde $(q_1, \ldots, q_m)$ es un vector de probabilidad, la matriz Q tiene todos sus renglones iguales. En este caso, los valores sucesivos de la cadena de Markov son independientes pues el valor siguiente de la cadena no depende del valor anterior. La probabilidad de aceptación es

$$\alpha_{ij} = \text{mín}\left\{ \frac{\pi_j}{\pi_i} \cdot \frac{q_i}{q_j}, 1 \right\}.$$

Mientras más parecido sea el vector $(q_1, \ldots, q_m)$ a la distribución objetivo $(\pi_1, \ldots, \pi_m)$, la aproximación será mejor. En el caso extremo $q_j = \pi_j$, la probabilidad de aceptación es 1. ■

Ejemplo 5.2 *(Muestreo uniforme)*
Suponga que se desea simular la distribución $(1/m, \ldots, 1/m)$. Se puede tomar, por ejemplo, $q_{j,j-1} = q_{jj} = q_{j,j+1} = 1/3$ para que la cadena sea irreducible y aperiódica. La probabilidad de aceptación es

$$\alpha_{ij} = \text{mín}\left\{ \frac{\pi_j}{\pi_i} \cdot \frac{q_i}{q_j}, 1 \right\} = \text{mín}\left\{ \frac{q_{ji}}{q_{ij}}, 1 \right\}.$$

■

A continuación estudiaremos el algoritmo Metropolis-Hastings en el caso de distribuciones continuas. Antes de ello, revisaremos algunos conceptos generales de cadenas de Markov con espacio de estados continuo. En este caso, las matemáticas son más complejas.

5.3. Cadenas de Markov con espacio de estados continuo

Consideraremos ahora cadenas de Markov con espacio de estados un conjunto S continuo. Este conjunto puede ser, por ejemplo, un intervalo (a, b), o $\mathbb{R}$, o $\mathbb{R}^d$, o algún subconjunto continuo de éstos. El tiempo sigue siendo discreto con valores $0, 1, 2, \ldots$ Los conceptos que enunciamos antes para cadenas de Markov con espacios de estados discretos deben modificarse para espacios continuos pues, por ejemplo, ahora tenemos que la probabilidad de transición $P(X_{t+1} = y \,|\, X_t = x)$ es cero para cualesquiera estados x y y. Las probabilidades de transición que debemos considerar son del tipo

$$P(X_{t+1} \in A \,|\, X_t = x),$$

en donde $A \subseteq S$ debe ser un conjunto medible. Al espacio de estados S se le debe asociar una $\sigma-$álgebra con el fin de hacerlo un espacio medible y poder considerar eventos o conjuntos medibles. La teoría matemática de este tipo de cadenas de Markov es más elaborada y sólo escribiremos algunas ideas generales. Una exposición más completa se puede encontrar en el texto de S.P. Meyn y R.L. Tweedie [42] o en D. Revuz [46].

El objetivo es proporcionar una introducción a la teoría de cadenas de Markov para exponer el algoritmo Metropolis-Hastings en el caso cuando la función objetivo es una función de densidad con soporte un conjunto continuo S. Este fue el contexto original en el que fue escrito el trabajo de Metropolis et al.

Definición 5.6 *Un proceso estocástico* $\{X_t : t = 0, 1, \ldots\}$ *con espacio de estados el conjunto continuo* S *es una cadena de Markov si se cumple la propiedad*

$$P(X_{t+1} \in A \,|\, X_t = x_t, \ldots, X_0 = x_0) = P(X_{t+1} \in A \,|\, X_t = x_t), \qquad (5.7)$$

para cualquier conjunto medible $A \subseteq S$ *y cualesquiera estados* $x_0, \ldots, x_t$ *en* S*, con* $t \geqslant 0$ *entero.*

Supondremos nuevamente que las probabilidades $P(X_{t+1} \in A \,|\, X_t = x)$ no dependen del valor t y se trata, por lo tanto, de probabilidades homogéneas en el tiempo. Mas aún, supondremos que estas probabilidades se pueden escribir en términos de una función de densidad $p(x, y)$ de la siguiente manera:

$$P(X_{t+1} \in A \,|\, X_t = x) = \int_A p(x, y)\, dy.$$

A la función $p(x, y) : S \times S \mapsto [0, \infty)$ se le llama función de densidad de transición o kernel de transición en un paso. Esta es la función de densidad que determina la posibilidad de pasar al estado y, a partir del estado x en un paso, y es la versión continua de la probabilidad de transición en un paso p_{ij} que teníamos antes para cadenas de Markov con espacio de estados discreto. A la función $p(x, y)$ se le escribe también como $p(y \,|\, x)$ o $p_x(y)$ para enfatizar que x representa el estado inicial y y el estado final de la transición.

Siendo una función de densidad en la variable y, la función $y \mapsto p(x, y)$ cumple la igualdad

$$\int_S p(x, y)\, dy = 1,$$

en donde el estado inicial x puede considerarse como un parámetro de la función de densidad de transición $y \mapsto p(x, y)$. Por ejemplo, la siguiente es una función de densidad de transición

$$p(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y - x)^2/(2\sigma^2)\right\}.$$

A partir del estado inicial x, esta función determina el siguiente estado y a través de una función de densidad normal de media x y varianza σ^2.

Es claro que la función de densidad $p(x, y)$ asigna mayor probabilidad a valores cercanos al punto inicial x y menor probabilidad a valores lejanos a x. Véase la Figura 5.3.

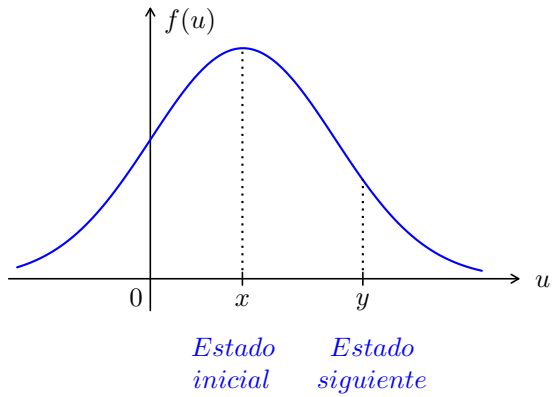


Figura 5.3: Ejemplo de función de transición $p(x, y)$.

Un estado inicial $x_0 \in S$, o una distribución inicial sobre el espacio de estados S, y una función de densidad de transición $p(x, y)$ determinan de manera única (en algún sentido) a la cadena de Markov. Es por ello que a la misma función $p(x, y)$ se le llama cadena de Markov. Un caso particular ocurrirá cuando $p(x, y)$ no depende del estado inicial x. En esta situación, la función de densidad de transición se escribirá simplemente como $p(y)$. Este es el caso análogo cuando una matriz $\mathbf{P} = (p_{ij})$ tiene todos sus renglones idénticos para el modelo con espacio de estados discreto.

Funciones de transición en n pasos

Se pueden considerar también las probabilidades de transición en n pasos de la siguiente manera: para $A \subseteq S$ medible y $n \geqslant 1$, $P(X_n \in A \,|\, X_0 = x)$ es la probabilidad de que, partiendo del estado x, la cadena tome un valor dentro del conjunto A después de n pasos. A esta probabilidad se le denota también como

$$P(X_n \in A \,|\, X_0 = x) = \int_A p^{(n)}(x, y)\, dy,$$

en donde a $p^{(n)}(x,y) \geqslant 0$ se le llama la función de densidad de transición en n pasos. Esta función de densidad se puede expresar en términos de la función de transición en un paso con ayuda de la ecuación de Chapman-Kolmogorov.

Proposición 5.5 *(Ecuación de Chapman-Kolmogorov)*
Para los números enteros $0 \leqslant k \leqslant n$ *se cumple que*

$$p^{(n)}(x,y) = \int_S p^{(k)}(x,z) \cdot p^{(n-k)}(z,y)\, dz.$$

Como el valor de k puede ser cualquier entero entre 0 y n, podemos escoger $k = 1$ y aplicar la ecuación de Chapman-Kolmogorov $n-1$ veces hasta tener sólo probabilidades de transición en un paso. Por ejemplo,

$$p^{(2)}(x,y) = \int_S p(x,z) \cdot p(z,y)\, dz,$$
$$p^{(3)}(x,y) = \int_S \int_S p(x,z) \cdot p(z,w) \cdot p(w,y)\, dz\, dw.$$

Para $n = 0$, se define $p^{(0)}(x,y)$ como la función delta de Dirac $\delta_x(y)$. Esta es una función generalizada que cumple la propiedad

$$\int_S \delta_x(z) \cdot g(z)\, dz = g(x),$$

de modo que, por ejemplo,

$$\begin{aligned} p^{(1)}(x,y) &= \int_S p^{(0)}(x,z) \cdot p^{(1)}(z,y)\, dz \\ &= \int_S \delta_x(z) \cdot p^{(1)}(z,y)\, dz \\ &= p(x,y). \end{aligned}$$

Irreducibilidad

Otra propiedad importante que ya se mencionó en el caso de espacios de estados discreto pero que es necesario adaptar cuando el espacio de

estados es continuo, es la irreducibilidad. La idea sigue siendo que, con probabilidad positiva, la cadena pueda visitar cualquier región medible A del espacio de estados a partir de cualquier estado inicial x, siempre y cuando la región A tenga probabilidad positiva de acuerdo a cierta función de densidad $\pi(x)$. A continuación se presenta con formalidad esta propiedad.

Definición 5.7 *Sea $\pi(x)$ una función de densidad sobre el espacio de estados S de una cadena de Markov $p(x,y)$. La cadena $p(x,y)$ es irreducible o $\pi-$irreducible (i.e. irreducible respecto de π) si para cada $x \in S$ y para cada conjunto medible $A \subseteq S$ tal que $\pi(A) > 0$, existe un número entero $n \geqslant 0$ tal que*

$$p^{(n)}(x, A) > 0.$$

Periodicidad

Definición 5.8 *Sea $p(x,y)$ una cadena de Markov $\pi-$irreducible y sea $d \geqslant 1$ un entero. Se dice que la cadena tiene un ciclo de longitud d si existen conjuntos medibles, no vacíos y disjuntos $A_1, A_2, \ldots, A_d$ tales que:*

1. *Para todo $x \in A_k$, $\quad p(x, A_{k+1}) = 1, \quad k = 1, 2, \ldots, d-1$.*
2. *Para todo $x \in A_d$, $\quad p(x, A_1) = 1$.*
3. $\pi\left((\bigcup_{i=1}^{d} A_i)^c\right) = 0$.

Un ciclo consta de una partición del espacio de estados tal que con probabilidad 1 se cumple la dinámica que se muestra en la Figura 5.4.

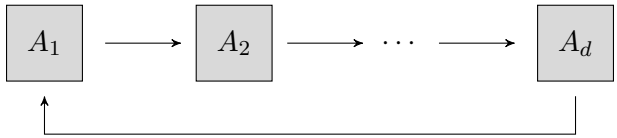


Figura 5.4: Un ciclo.

Definición 5.9 *Sea $p(x, y)$ una cadena de Markov $\pi-$irreducible. El periodo es la longitud $d \geqslant 1$ más grande de un ciclo de la cadena. Cuando $d = 1$ se dice que la cadena es aperiódica.*

Recurrencia

A continuación definiremos dos tipos de recurrencia para cadenas de Markov con espacio de estados continuo.

Definición 5.10 *Sea $\pi(x)$ una función de densidad sobre el espacio de estados S de una cadena de Markov $p(x, y)$. Sea $A \subseteq S$ cualquier conjunto medible tal que $\pi(A) > 0$. La cadena es recurrente si para todo $x_0 \in S$,*

$$P((X_n \in A) \text{ i.o.} \mid X_0 = x_0) > 0.$$

Las letras i.o. provienen del inglés "infinitely often". La expresión $(X_n \in A)$ i.o. representa el evento que ocurre cuando un número infinito de los eventos $(X_1 \in A), (X_2 \in A), (X_3 \in A), \ldots$ ocurre. Esto no significa que todos ellos ocurren, pero se debe verificar que una infinidad de ellos se observan. Esto significa que la cadena visita la región A una infinidad de veces con probabilidad positiva. Se puede escribir

$$((X_n \in A) \text{ i.o.}) = \limsup_{n\to\infty} (X_n \in A) = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} (X_k \in A).$$

Un tipo de recurrencia más fuerte requiere que la probabilidad de este

evento sea 1. La definición es la siguiente.

Definición 5.11 *Sea $\pi(x)$ una función de densidad sobre el espacio de estados S de una cadena de Markov $p(x, y)$. Sea $A \subseteq S$ cualquier conjunto medible tal que $\pi(A) > 0$. Se dice que la cadena es recurrente de Harris si para todo $x_0 \in S$,*

$$P((X_n \in A) \text{ i.o.} \mid X_0 = x_0) = 1.$$

Distribuciones estacionarias

Si $\pi_0(x)$ es una distribución de probabilidad sobre el espacio S de una cadena de Markov en un tiempo dado, entonces la distribución en la siguiente unidad de tiempo es

$$\pi_1(y) = \int_S \pi_0(x) \cdot p(x, y)\, dx, \quad y \in S.$$

Cuando una distribución permanece sin cambio bajo esta transformación, se dice que es estacionaria o invariante en el tiempo.

Definición 5.12 *Una distribución estacionaria para una cadena de Markov $p(x, y)$ con espacio de estados continuo S es una distribución de probabilidad $\pi(x)$ definida sobre S tal que*

$$\pi(y) = \int_S \pi(x)\, p(x, y)\, dx, \quad y \in S.$$

El siguiente resultado establece condiciones que garantizan la unicidad de una distribución estacionaria.

Teorema 5.3 *Sea $\pi(x)$ una distribución estacionaria para una cadena de Markov $p(x, y)$. Si la cadena es $\pi-$irreducible, entonces $\pi(x)$ es la única distribución estacionaria de la cadena.*

Distribuciones límite

Definición 5.13 *Sea $p(x, y)$ una cadena de Markov con espacio de estados continuo S. Cuando el límite de la función de transición en n pasos*

$$\pi(y) = \lim_{n\to\infty} p^{(n)}(x, y), \quad y \in S,$$

existe y no depende de x, a $\{\pi(y) : y \in S\}$ se le llama distribución límite.

En realidad no está garantizado que la distribución límite definida arriba sea una distribución de probabilidad verdadera. Los límites pueden existir pero ser todos cero, por ejemplo. El siguiente resultado establece condiciones para la existencia de las probabilidades límite.

Teorema 5.4 *Sea $p(x, y)$ una cadena de Markov con espacio de estados continuo S tal que:*

a) Posee una distribución estacionaria $\{\pi(x) : x \in S\}$.

b) Es irreducible respecto de $\pi(x)$.

c) Es aperiódica.

Entonces

1. *Para casi cualquier $x \in S$, según la medida $\pi(x)$, se cumple que*

$$\lim_{n\to\infty} ||p^{(n)}(x, A) - \pi(A)||_{TV} = 0,$$

en donde $||\cdot||_{TV}$ es la norma de variación total entre dos medidas de probabilidad, es decir,

$$||\pi_1(A) - \pi_2(A)||_{TV} = \sup_A |\pi_1(A) - \pi_2(A)|.$$

2. *Si $p(x, y)$ es una cadena recurrente de Harris, entonces la convergencia se verifica para cualquier $x \in S$.*

Ecuación de balance y cadena reversible

Definición 5.14 *Sea $p(x, y)$ una cadena de Markov con espacio de estados continuo S. Se dice que la cadena es reversible respecto de una distribución $\{\pi(x) : x \in S\}$ cuando se cumple que*

$$\pi(x)\, p(x, y) = \pi(y)\, p(y, x), \quad x, y \in S. \tag{5.8}$$

A la identidad (5.8) se le llama ecuación de balance. Como en el caso de cadenas de Markov con espacio de estados discreto, es fácil comprobar que toda distribución $\pi(x)$ que satisface la ecuación de balance (5.8) es una

distribución estacionaria. En efecto,

$$\begin{aligned}\int_S \pi(x)\, p(x,y)\, dx &= \int_S \pi(y)\, p(y,x)\, dx \\ &= \pi(y) \int_S p(y,x)\, dx \\ &= \pi(y).\end{aligned}$$

Con estos elementos podemos ahora exponer el algoritmo Metropolis-Hastings para generar valores de distribuciones continuas.

5.4. El algoritmo Metropolis-Hastings: distribuciones continuas

Sea $f(x)$ una función de densidad con soporte un conjunto continuo S como (a, b), $\mathbb{R}$ o $\mathbb{R}^d$. Esta será nuestra función objetivo. Como en el caso de distribuciones discretas, el algoritmo Metropolis-Hastings propone crear una cadena de Markov con espacio de estados S y que tenga a $f(x)$ como su distribución límite. De esta manera, producir valores de la cadena de Markov a largo plazo proporcionará una muestra de valores de la distribución límite $f(x)$, aunque los números obtenidos no serán necesariamente generados de manera independiente y la distribución de los mismos será sólo de manera aproximada.

Como elemento base se requiere contar con una función de densidad de transición $q(x, y)$ definida sobre el espacio producto $S \times S$, en donde x representa el estado inicial y y representa el estado final. Esta es la función propuesta, equivalente a la matriz Q en el caso de espacios de estados discretos. Supondremos que se conoce un mecanismo para generar valores de la función de densidad $y \mapsto q(x, y)$, para cada x fijo.

La construcción de la cadena de Markov $\{X_t : t = 0, 1, \ldots\}$ con espacio de estados S es similar al caso de espacio de estados discretos y es como sigue:

Construcción de la cadena de Markov

Sea $x_0 \in S$ un estado inicial de la cadena.
Para $t = 1, 2, \ldots$

- Se genera un valor y de la distribución $y \mapsto q(x_{t-1}, y)$.
- Se calcula la probabilidad
$$\alpha(x_{t-1}, y) = \text{mín}\left\{ \frac{f(y)}{f(x_{t-1})} \cdot \frac{q(y, x_{t-1})}{q(x_{t-1}, y)}, 1 \right\}.$$
- Se define
$$x_t = \begin{cases} y & \text{con prob. } \alpha(x_{t-1}, y), \quad \text{(Aceptación)} \\ x_{t-1} & \text{con prob. } 1 - \alpha(x_{t-1}, y). \quad \text{(Rechazo)} \end{cases}$$

Es decir, en cada paso, con probabilidad $\alpha(x_{t-1}, y)$ se acepta el valor generado y como el siguiente valor del proceso, y con probabilidad complementaria se rechaza quedando el proceso en el mismo estado que en el tiempo anterior. Una forma equivalente de escribir la regla de decisión es la siguiente:

1. Si $u \leqslant \alpha(x_{t-1}, y)$, en donde u es un valor de la distribución unif $(0, 1)$, se define x_t como y.

2. En cambio, si $u > \alpha(x_{t-1}, y)$, se define x_t como x_{t-1}.

De esta manera se obtiene una trayectoria $x_0, x_1, x_2, \ldots$ de un proceso $\{X_t : t = 0, 1, 2, \ldots\}$ que, por construcción, resulta ser una cadena de Markov. Demostraremos ahora que el proceso estocástico continuo construido es, efectivamente, una cadena de Markov y mostraremos las probabilidades de transición de este proceso.

Proposición 5.6 *El proceso* $\{X_t : t = 0, 1, \ldots\}$ *construido en el recuadro anterior es una cadena de Markov con espacio de estados* S *y función de densidad de transición*

$$p(x, y) = q(x, y)\,\alpha(x, y) + r(x)\,\delta_x(y), \tag{5.9}$$

en donde $\delta_x(y)$ *es la función delta de Dirac y*

$$r(x) = 1 - \int_S q(x, s)\,\alpha(x, s)\,ds.$$

Demostración. Por construcción, el valor de la variable x_t depende únicamente de tres elementos: el valor de x_{t-1}, la función propuesta $q(x, y)$ y la función de densidad objetivo $f(x)$. Por lo tanto, la propiedad de Markov se cumple. Por otro lado, para cualquier región $A \subseteq S$ medible,

$$\begin{aligned} P(X_{t+1} \in A \,|\, X_t = x_t) &= \int_A q(x_t, s)\,\alpha(x_t, s)\,ds \\ &\quad + \int_S q(x_t, s)(1 - \alpha(x_t, s))\,ds\,\delta_A(x_t) \\ &= \int_A q(x_t, s)\,\alpha(x_t, s)\,ds \\ &\quad + \left(1 - \int_S q(x_t, s)\,\alpha(x_t, s)ds\right)\,\delta_A(x_t) \\ &= \int_A [q(x_t, s)\alpha(x_t, s) + r(x_t)\delta_{x_t}(s)]\,ds. \end{aligned}$$

■

El siguiente cálculo comprueba que la función $y \mapsto p(x, y)$ especificada en (5.9) es, efectivamente, una función de densidad.

$$\begin{aligned} \int_S p(x, y)\,dy &= \int_S [q(x, y)\,\alpha(x, y) + r(x)\,\delta_x(y)]\,dy \\ &= \int_S q(x, y)\,\alpha(x, y)\,dy + r(x) \\ &= 1. \end{aligned}$$

A continuación demostraremos que la distribución objetivo $f(x)$ es una distribución estacionaria para la cadena de Markov recién construida. Comprobaremos esto a través de la ecuación de balance detallado. El análisis es análogo al caso de espacios de estados discretos.

Proposición 5.7 *(Ecuación de balance)*
La función objetivo $f(x)$ y la cadena de Markov $p(x,y)$ recién construida satisfacen

$$f(x)\,p(x,y) = f(y)\,p(y,x), \quad x,y \in S.$$

Demostración. La igualdad es evidente cuando $x = y$. Supongamos que $x \neq y$. Entonces

$$\begin{aligned} f(x)\cdot p(x,y) &= f(x)\cdot\alpha(x,y)\cdot q(x,y) \\ &= f(x)\cdot \min\left\{\frac{f(y)}{f(x)}\cdot\frac{q(y,x)}{q(x,y)}, 1\right\}\cdot q(x,y) \\ &= \min\{f(y)\cdot q(y,x), f(x)\cdot q(x,y)\} \\ &= f(y)\cdot \min\left\{1, \frac{f(x)}{f(y)}\cdot\frac{q(x,y)}{q(y,x)}\right\}\cdot q(y,x) \\ &= f(y)\cdot\alpha(y,x)\cdot q(y,x) \\ &= f(y)\cdot p(y,x). \end{aligned}$$

∎

Es inmediato comprobar que toda distribución que satisface la ecuación de balance detallado es una distribución estacionaria. Verificaremos esto a continuación.

Proposición 5.8 *La distribución objetivo $f(x)$ es una distribución estacionaria para la cadena de Markov $p(x,y)$ del algoritmo Metropolis-Hastings.*

Demostración. Para cualquier estado $y \in S$, por la ecuación de balance

detallado,

$$\begin{aligned}\int_S f(x)\cdot p(x,y)\,dx &= \int_S f(y)\cdot p(y,x)\,dx \\ &= f(y)\cdot\int_S p(y,x)\,dx \\ &= f(y).\end{aligned}$$

■

Por lo tanto, la distribución objetivo $f(x)$ es siempre una distribución estacionaria para la cadena de Markov $p(x,y)$ del algoritmo Metropolis-Hastings, aunque no está garantizado que ésta sea su distribución límite, será así cuando la cadena $p(x,y)$ sea irreducible y aperiódica. Veremos ahora algunos casos particulares del algoritmo Metropolis-Hastings.

Ejemplo 5.3 *(Muestreo independiente)*
Supongamos que la función de transición propuesta $q(x,y)$ no depende de x. Esta función de densidad se escribe simplemente como $q(y)$, $y \in S$. De esta manera, la generación del estado propuesto y es independiente del estado anterior x. La probabilidad de aceptación de un estado propuesto y es

$$\alpha(x,y) = \text{mín}\left\{\frac{f(y)}{f(x)}\cdot\frac{q(y,x)}{q(x,y)}, 1\right\} = \text{mín}\left\{\frac{f(y)}{f(x)}\cdot\frac{q(x)}{q(y)}, 1\right\}.$$

Se observa que esta probabilidad de aceptación depende del estado inicial x a través de los términos $f(x)$ y $q(x)$. En el caso extremo y falto de utilidad cuando $q(x) = f(x)$, la probabilidad de aceptación es siempre uno.

Como un ejemplo particular consideremos el caso cuando el espacio de estados S es un intervalo acotado (a,b). La función de densidad objetivo es cualquier función de densidad $f(x)$ con soporte (a,b), y se toma a la función de densidad propuesta $q(y)$ como la distribución unif (a,b)*. Las probabilidades de aceptación se reducen a*

$$\alpha(x,y) = \text{mín}\{f(y)/f(x), 1\}.$$

De esta manera, a partir de valores de la distribución unif (a,b) *se pueden obtener muestras aproximadas de la distribución objetivo $f(x)$. Por ejemplo,*

a partir de valores de la distribución unif $(0,1)$ *se pueden obtener valores de la distribución* beta (a,b). ∎

Ejemplo 5.4 *(Distribución gama)*
Supongamos que deseamos obtener valores de la distribución gama (α,λ). *La función de densidad objetivo es*

$$f(x) = \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}\lambda e^{-\lambda x}, \quad x > 0,$$

en donde los parámetros $\alpha > 0$ *y* $\lambda > 0$ *son conocidos. El espacio de estados* S *en el contexto del algoritmo Metropolis-Hastings es el soporte de esta distribución y en este ejemplo es el intervalo* $(0,\infty)$. *Como función propuesta se puede tomar la distribución exponencial*

$$q(y) = \lambda e^{-\lambda y}, \quad y > 0,$$

cuyos valores se pueden obtener por el método de la función inversa mediante la fórmula $y = -(1/\lambda)\ln u$, *en donde* u *es un valor de la función de densidad* unif $(0,1)$. *Las probabilidades de aceptación son*

$$\alpha(x,y) = \text{mín}\left\{\frac{f(y)}{f(x)}\cdot\frac{q(x)}{q(y)}, 1\right\},$$

en donde

$$\frac{f(y)}{f(x)}\cdot\frac{q(x)}{q(y)} = \frac{\left[(\lambda y)^{\alpha-1}/\Gamma(\alpha)\right]\cdot\lambda e^{-\lambda y}}{\left[(\lambda x)^{\alpha-1}/\Gamma(\alpha)\right]\cdot\lambda e^{-\lambda x}}\frac{\lambda e^{-\lambda x}}{\lambda e^{-\lambda y}} = \left(\frac{y}{x}\right)^{\alpha-1}.$$

Una muestra de la función objetivo $f(x)$ *puede obtenerse de la manera que indica el algoritmo Metropolis-Hastings, es decir, se toma un valor inicial* $x_0 \in (0,\infty)$ *y para* $i = 1, 2\ldots$ *se define*

$$x_i = \begin{cases} y & \textit{con prob.} \quad \alpha(x_{i-1}, y), \\ x_{i-1} & \textit{con prob.} \quad 1-\alpha(x_{i-1}, y). \end{cases}$$

∎

Ejemplo 5.5 *(Método MCMC para aproximar una integral)*
Sea $g(x)$ una función integrable en el intervalo (a, b). Mostraremos una manera de aproximar la integral que aparece abajo usando el método MCMC. Nos interesa estimar

$$\begin{aligned} \theta &= \int_a^b g(x)dx \\ &= \int_a^b \frac{g(x)}{f(x)} f(x)dx \\ &= E\left(\frac{g(X)}{f(X)}\right), \end{aligned} \tag{5.10}$$

en donde X es una variable aleatoria con función de densidad $f(x)$, cuyo soporte contiene al intervalo (a, b). Este es el método de Monte Carlo de la media muestral. Deseamos generar valores de una distribución con función de densidad $f(x)$ y, por lo tanto, esta es nuestra función objetivo. Sea $q(y)$ otra función de densidad con soporte (a, b) y de la cual se conoce una manera sencilla de simular sus valores. Esta es la función propuesta. El algoritmo MCMC es el siguiente.

Algoritmo MCMC para aproximar la integral dada por (5.10)

Se toma un valor inicial x_0 *en el intervalo* (a, b).
Sea N *el tamaño de muestra requerido.*
Para $t = 1, 2, \ldots, N$,

- *Sea* y *un valor de una distribución con función de densidad* $q(y)$.
- *Se calcula la probabilidad*

$$\alpha(x_{i-1}, y) = \text{mín}\left\{\frac{f(y)}{f(x_{i-1})} \cdot \frac{q(x_{i-1})}{q(y)}, 1\right\}.$$

- *Se define*

$$x_i = \begin{cases} y & \text{con prob. } \alpha(x_{i-1}, y), \quad \textit{(Aceptación)} \\ x_{i-1} & \text{con prob. } 1 - \alpha(x_{i-1}, y). \quad \textit{(Rechazo)} \end{cases}$$

Se define la estimación como el promedio de las evaluaciones en los últimos elementos de la muestra, por ejemplo,

$$\hat{\theta} = \frac{1}{N - n + 1} \cdot \sum_{i=n}^{N} \frac{g(x_i)}{f(x_i)}, \quad 1 \leqslant n \leqslant N.$$

Como valor inicial se puede tomar, por ejemplo, $x_0 = (a + b)/2$. *En el procedimiento arriba descrito se descartan las primera* $n - 1$ *observaciones generadas ya que se consideran un periodo de "calentamiento", donde la cadena que se genera posiblemente aún no comienza a mostrar su comportamiento límite.*

El mismo procedimiento se puede utilizar cuando la integral a estimar es, por ejemplo,

$$\theta = E(\varphi(X)) = \int \varphi(x)\, g(x)\, dx,$$

en donde $\varphi(x)$ *es una transformación y* X *es una variable aleatoria con función de densidad* $g(x)$, *posiblemente complicada y de la cual no conocemos*

una forma de generar valores. La estimación es

$$\hat{\theta} = \frac{1}{N-n+1} \sum_{i=n}^{N} \frac{\varphi(x_i)\, g(x_i)}{f(x_i)},$$

en donde $x_1, \ldots, x_N$ son valores de una distribución con función de densidad adecuada $f(x)$. Se pueden también considerar integrales de mayor dimensión como

$$\theta = E(\varphi(X,Y)) = \int\int \varphi(x,y)\, g(x,y)\, dx\, dy,$$

en donde $\varphi(x,y)$ es una transformación y (X,Y) es un vector aleatorio con función de densidad conjunta $g(x,y)$. La estimación es

$$\hat{\theta} = \frac{1}{N-n+1} \sum_{i=n}^{N} \frac{\varphi(x_i,y_i)\, g(x_i,y_i)}{f(x_i,y_i)},$$

en donde $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$ son valores de una distribución con función de densidad adecuada $f(x,y)$.

En cualquier caso se debe contar con una función propuesta adecuada, $q(y)$ o $q(x,y)$, de la cual se conozca una manera (idealmente sencilla) de generar sus valores.

▪

5.5. El muestreo de Gibbs

Este es un caso especial del método Metropolis-Hastings. Fue propuesto en 1984 por dos autores con el mismo apellido: Geman S. y Geman D. [15]. El nombre de Gibbs se debe a que el trabajo de Geman y Geman estaba relacionado con un tipo de distribuciones de probabilidad llamadas distribuciones de Gibbs.

El muestreo de Gibbs proporciona un método para producir valores de un vector aleatorio $(X_1, \ldots, X_n)$ con una función de densidad o de probabilidad conocida $f(x_1, \ldots, x_n)$. El procedimiento requiere que se conozca una forma

de producir valores de cada una de las funciones de densidad condicionales univariadas, es decir, de

$$f(x_i \,|\, x_1, \ldots, x_{i-1}, x_i, \ldots, x_n), \quad i = 1, 2, \ldots, n.$$

Estudiaremos primero el caso de vectores bidimensionales ($n = 2$) y después extenderemos los conceptos y resultados al caso de vectores de cualquier dimensión. Veremos que el aspecto central del método consiste en la construcción de una cadena de Markov con distribución estacionaria la función de densidad $f(x_1, \ldots, x_n)$.

Muestreo de Gibbs para vectores aleatorios bidimensionales

Sea (X, Y) un vector aleatorio con función de densidad conocida $f(x, y)$ y del cual se desean obtener valores. Sean $f(x \,|\, y)$ y $f(y \,|\, x)$ las correspondientes funciones de densidad condicionales. Supondremos que conocemos algún mecanismo para obtener valores de estas funciones de densidad condicionales. El algoritmo de Gibbs es el siguiente.

Muestreo de Gibbs (Vectores de dimensión 2)

Sea x_0 un valor inicial de X.
Sea n el tamaño de muestra requerido.
Para $t = 1, 2, \ldots, n$,

- Sea y_t un valor de $f_{Y \,|\, X}(y \,|\, x_{t-1})$.
- Sea x_t un valor de $f_{X \,|\, Y}(x \,|\, y_t)$.

Este algoritmo produce la sucesión de parejas (x_1, y_1),(x_2, y_2),. . .,(x_n, y_n) a partir del valor inicial x_0. Esta es una realización de la sucesión de vectores aleatorios

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n),$$

en donde la función de densidad de Y_t es la función de densidad de Y condicionada al evento $(X = x_{t-1})$, y la función de densidad de X_t es la función de densidad de X condicionada al evento $(Y = y_t)$. Se trata de un proceso estocástico a tiempo discreto $t = 1, 2, \ldots$ y con espacio de estados

el producto cartesiano del rango de valores de X por el rango de valores de Y.

Por construcción, las dos variables dentro del vector (X_t, Y_t) no son independientes, y los vectores mismos tampoco son independientes. La dependencia y construcción secuencial de los valores de las variables se muestra en la Figura 5.5.

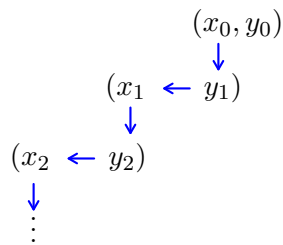


Figura 5.5: Secuencia de valores en el muestreo de Gibbs.

Es necesario enfatizar que la distribución de las variables X_t y Y_t son las distribuciones condicionales $f_{X_t \mid Y_t}(x_t \mid y_t) = f_{X \mid Y}(x_t \mid y_t)$ y $f_{Y_t}(y_t) = f_{Y \mid X}(y_t \mid x_{t-1})$, respectivamente. Por lo tanto, para $t \geqslant 0$, la función de densidad conjunta del vector (X_t, Y_t) es

$$\begin{aligned} f_{X_t, Y_t}(x_t, y_t) &= f_{X_t \mid Y_t}(x_t \mid y_t)\, f_{Y_t}(y_t) \\ &= f_{X \mid Y}(x_t \mid y_t)\, f_{Y \mid X}(y_t \mid x_{t-1}). \end{aligned}$$

Integrando respecto de x_t primero y después respecto de y_t, puede comprobarse que esta función es, efectivamente, una función de densidad bivariada. Observe que esta función de densidad depende del valor x_{t-1}, el cual puede considerarse como un parámetro de la distribución.

Proposición 5.9 *El proceso estocástico* $\{(X_t, Y_t) : t = 1, 2, \ldots\}$ *generado por el algoritmo de Gibbs es una cadena de Markov con espacio de estados* $\mathbb{R}^2$ *y con función de transición*

$$f((x_t, y_t), (x_{t+1}, y_{t+1})) = f_{X \mid Y}(x_{t+1} \mid y_{t+1})\, f_{Y \mid X}(y_{t+1} \mid x_t).$$

Demostración. Por construcción, la variable X_{t+1} depende de Y_{t+1}, y ésta, a su vez, depende de X_t. De esta manera, el vector (X_{t+1}, Y_{t+1}) se determina a partir del vector (X_t, Y_t) y de las funciones de densidad condicionales $f_{X\,|\,Y}(x\,|\,y)$ y $f_{Y\,|\,X}(y\,|\,x)$. Esto comprueba que la propiedad de Markov se cumple. La función de densidad de transición es: para $t = 1, 2, \ldots$

$$\begin{aligned} f((x_t, y_t), (x_{t+1}, y_{t+1})) &= \frac{f_{X_{t+1},Y_{t+1},X_t,Y_t}(x_{t+1}, y_{t+1}, x_t, y_t)}{f_{X_t,Y_t}(x_t, y_t)} \\ &= f_{X_{t+1}\,|\,Y_{t+1}}(x_{t+1}\,|\,y_{t+1}) \cdot f_{Y_{t+1}\,|\,X_t}(y_{t+1}\,|\,x_t) \\ &\qquad \cdot \frac{f_{X_t,Y_t}(x_t, y_t)}{f_{X_t,Y_t}(x_t, y_t)} \\ &= f_{X\,|\,Y}(x_{t+1}\,|\,y_{t+1}) \cdot f_{Y\,|\,X}(y_{t+1}\,|\,x_t). \end{aligned}$$

∎

Observemos que la función de transición es homogénea en el tiempo pues ésta no depende del valor de t. Los números x_t, x_{t+1}, y_t y y_{t+1} que aparecen en los argumentos utilizan el subíndice t para asociarlos a las variables aleatorias correspondientes pero no existe dependencia del tiempo.

Por otro lado, es necesario advertir que la notación se hace más ligera (a riesgo de presentarse alguna confusión) cuando se utiliza la misma letra $f(\cdot)$ para denotar distintas funciones. Si (x, y) y (u, v) son dos valores del vector aleatorio (X, Y), en la Tabla 5.1 se presenta un resumen de significados.

A continuación comprobaremos que la función de densidad $f(x, y)$ es una distribución estacionaria para la cadena de Markov construida en el algoritmo de Gibbs.

Notación	Significado
$f(x,y)$ ó $f(u,v)$	Función objetivo
$f((x,y),(u,v))$	función de densidad de transición del punto (x,y) al punto (u,v)
$f(x)$, $f(u)$	función de densidad marginal de X
$f(y)$, $f(v)$	función de densidad marginal de Y
$f(x \mid y)$, $f(u \mid v)$	función de densidad condicional de X dado Y
$f(y \mid x)$, $f(v \mid u)$	función de densidad condicional de Y dado X

Tabla 5.1: Funciones de densidad.

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\cdot f((x,y),(u,v))\,dx\,dy \\
&\quad= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\cdot f_{X\mid Y}(u\mid v)\cdot f_{Y\mid X}(v\mid x)\,dx\,dy \\
&\quad= \int_{-\infty}^{\infty}\frac{f(u,v)}{f(v)}f(x,y)\cdot\frac{1}{f(x)}\int_{-\infty}^{\infty} f(x,y)\,dy\,dx \\
&\quad= \int_{-\infty}^{\infty}\frac{f(u,v)}{f(v)}f(u,v)\cdot\frac{1}{f(v)}\int_{-\infty}^{\infty} f(x,v)dx \\
&\quad= f(u,v).
\end{aligned}$$

Por lo tanto, cuando la cadena sea convergente y tenga a $f(x,y)$ como su única distribución estacionaria, los puntos (x_1,y_1), $(x_2,y_2),\ldots$ obtenidos a través del algoritmo de Gibbs estarán distribuidos de manera aproximada de acuerdo a la función de densidad $f(x,y)$. Veremos ahora un ejemplo.

Ejemplo 5.6 *(Distribución normal bivariada)*
Recordemos que un vector aleatorio (X,Y) tiene distribución normal bivariada cuando su función de densidad es

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{\mathbf{A}\}, \quad -\infty < x,y < \infty,$$

en donde

$$\mathbf{A} = -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]$$

y los parámetros son tales que $-1 < \rho < 1$, $\sigma_1 > 0$, $\sigma_2 > 0$, $\mu_1, \mu_2 \in \mathbb{R}$. *Se escribe* $(X, Y) \sim \mathrm{N}(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \rho)$. *Para esta distribución bivariada no es difícil comprobar que se cumplen las siguientes propiedades:*

1. $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$.
2. $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$.
3. $\rho(X, Y) = \rho$. *Este número es el coeficiente de correlación entre* X *y* Y.
4. X *y* Y *son independientes si, y sólo si,* $\rho = 0$.
5. $\mathrm{Var}(X, Y) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.
6. *Para cualquier valor de* y,

$$X \,|\, (Y = y) \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), (1-\rho^2)\sigma_1^2\right). \tag{5.11}$$

7. *Para cualquier valor de* x,

$$Y \,|\, (X = x) \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), (1-\rho^2)\sigma_2^2\right). \tag{5.12}$$

∎

Los últimos dos resultados son los que nos interesan para aplicar el muestreo de Gibbs para obtener valores del vector (X, Y). Supondremos que conocemos una manera de generar valores que siguen estas funciones de densidad normales univariadas. El algoritmo es el siguiente.

Muestreo de Gibbs (Distribución normal bivariada)

Se x_0 un valor inicial.
Para $t = 1, 2, \ldots$
- Sea y_t un valor de $Y \,|\, (X = x_{t-1})$ con función de densidad (5.12).
- Sea x_t un valor de $X \,|\, (Y = y_t)$ con función de densidad (5.11).

Se genera de esta manera la sucesión de vectores $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$ El algoritmo de Gibbs garantiza que estos vectores numéricos se pueden considerar valores con distribución aproximada normal bivariada.

Veremos a continuación el muestreo de Gibbs para vectores de dimensión n. El procedimiento se extiende de manera natural.

Muestreo de Gibbs para vectores aleatorios multidimensionales

Supongamos que se desea obtener valores de un vector aleatorio $(X_1, \ldots, X_n)$ con función de densidad conocida $f(x_1, \ldots, x_n)$. Por simplicidad en la notación, las funciones de densidad condicionales de una variable se escribirán de la forma siguiente: para X_1, por ejemplo,

$$f(x_1 \,|\, x_2, \ldots, x_n),$$

es decir, el subíndice indica la variable aleatoria de referencia. Supondremos que se conoce un método para generar valores de cada una de estas distribuciones univariadas. Para el caso de X_1 indicado arriba, observe que los valores $x_2, \ldots, x_n$ pueden ser considerados como parámetros de la distribución condicional. El algoritmo o muestreo de Gibbs es el siguiente.

Muestreo de Gibbs (Vectores de dimensión n)

Sea $(x_1^0, \ldots, x_n^0)$ un valor inicial del vector $(X_1, \ldots, X_n)$.
Para $t = 1, 2, \ldots$

- Sea x_n^t un valor de $f(x_n \,|\, x_1^{t-1}, x_2^{t-1}, \ldots, x_{n-1}^{t-1})$.
- Sea x_{n-1}^t un valor de $f(x_{n-1} \,|\, x_1^{t-1}, \ldots, x_{n-2}^{t-1}, x_n^t)$.

$\vdots$

- Sea x_1^t un valor de $f(x_1 \,|\, x_2^t, \ldots, x_n^t)$.

Observe que en las funciones de densidad condicionales se van incorporando de manera sistemática los valores $x_1^t, \ldots, x_n^t$ generados en los pasos anteriores. Más específicamente, en la parte condicional se tiene la intersección de los siguientes eventos:

$(x_1^{t-1}, \ldots, x_{k-1}^{t-1})$ Entradas del vector anterior,

$(x_{k+1}^t, \ldots, x_n^t)$ Valores generados en los pasos anteriores.

Así, este algoritmo produce la sucesión de vectores $(x_1^1, \ldots, x_n^1), (x_1^2, \ldots, x_n^2), \ldots$ los cuales son una realización de la sucesión de vectores aleatorios

$$(X_1^1, \ldots, X_n^1), (X_1^2, \ldots, X_n^2), \ldots$$

Por ejemplo, para $n = 3$ la dependencia de los valores generados se muestra en el diagrama de la Figura 5.6, el cual se lee de derecha a izquierda: cada número depende de las 2 cantidades más inmediatas que se encuentran a su derecha.

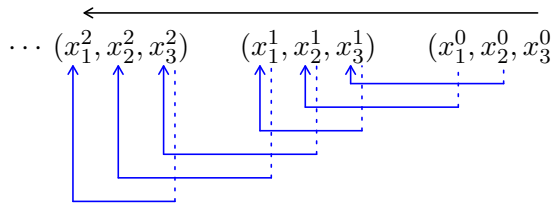


Figura 5.6: Generación de valores en el muestreo de Gibbs.

De manera equivalente se puede definir el muestreo de Gibbs para que la generación (y dependencia) de los números sea como la que muestra en el diagrama de la Figura 5.7. Este diagrama se lee de izquierda a derecha: cada número depende de las 2 cantidades más inmediatas que se encuentran a su izquierda.

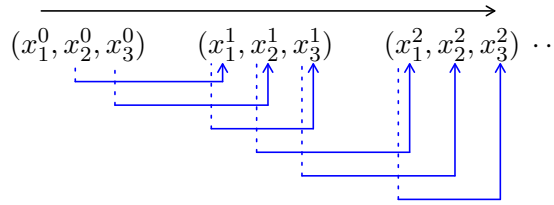


Figura 5.7: Generación de valores en el muestreo de Gibbs.

Como se hizo en el caso $n = 2$, se comprueba que el proceso estocástico $\{(X_1^t, \ldots, X_n^t) : t = 1, 2, \ldots\}$ es una cadena de Markov que tiene a $f(x_1, \ldots, x_n)$ como distribución estacionaria. Cuando la cadena sea convergente y tenga a $f(x_1, \ldots, x_n)$ como única distribución estacionaria, la sucesión de puntos $(x_1^1, \ldots, x_n^1), (x_1^2, \ldots, x_n^2), \ldots$ se distribuirán de manera aproximada de acuerdo a la función de densidad $f(x_1, \ldots, x_n)$.

5.6. El muestreo de Gibbs como caso particular del algoritmo Metropolis-Hastings

Mostraremos la forma en la que el muestreo de Gibbs puede insertarse dentro del esquema del algoritmo Metropolis-Hastings como un caso particular. Sea $f(x_1, \ldots, x_n)$ la función de densidad objetivo y sobre la cual se aplica el muestreo de Gibbs. Será conveniente usar la siguiente notación.

Notación

Sea $x = (x_1, \ldots, x_n)$ un vector numérico y sea y un número. Para cada $i = 1, 2, \ldots, n$,

- El vector x con la $i-$ésima coordenada eliminada es
$$x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$
- El vector x con la $i-$ésima coordenada reemplazada por y es
$$[y, x_{(i)}] = (x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n).$$

No debe confundirse a $x_{(i)}$ con el valor de la i-ésima estadística de orden. En algunos otros trabajos, al vector reducido $x_{(i)}$ se le escribe también como x_{-i}, pero preferimos la primera notación por su simplicidad. En particular, observemos que $[x_i, x_{(i)}]$ es el vector completo $x = (x_1, \ldots, x_n)$ y que $f(x_{(i)})$ es una función de densidad marginal de $f(x_1, \ldots, x_n)$.

En el muestreo de Gibbs se genera una coordenada a la vez de manera secuencial hasta obtener todas las coordenadas del nuevo vector, el cual no pasa por un proceso de aceptación o rechazo, sino que es aceptado de manera implícita de acuerdo a la forma en la que se define el algoritmo de Gibbs. Para la $i-$ésima coordenada se usa la función de densidad condicional

$$f(x_i \,|\, x_{(i)}) = f(x_i \,|\, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),$$

para cierto vector $x_{(i)}$ que depende de algunas coordenadas del vector a un tiempo anterior y de las nuevas coordenadas generadas. Supongamos que y_i es el valor generado usando esta función de densidad. Artificialmente uno puede pensar que se generó el vector $[y_i, x_{(i)}]$ y que la función propuesta que

se ha utilizado es

$$q\left([y_i, x_{(i)}] \,|\, [x_i, x_{(i)}]\right) = f(y_i \,|\, x_{(i)}). \tag{5.13}$$

De esta manera, el $i-$ésimo paso del algoritmo de Gibbs se ha expresado como un paso del algoritmo Metropolis-Hastings. Llamando y al vector $[y_i, x_{(i)}]$, la probabilidad de aceptación de este vector preliminar en el algoritmo de Gibbs es

$$\alpha(x, y) = \text{mín}\left\{\frac{f(y)}{f(x)} \cdot \frac{q(y, x)}{q(x, y)}, 1\right\},$$

en donde, utilizando (5.13), encontramos que

$$\begin{aligned}
\frac{f(y)}{f(x)}\frac{q(y,x)}{q(x,y)} &= \frac{f(y)}{f(x)} \cdot \frac{q(x \,|\, y)}{q(y \,|\, x)} \\
&= \frac{f\left([y_i, x_{(i)}]\right)}{f\left([x_i, x_{(i)}]\right)} \cdot \frac{q\left([x_i, x_{i-1}] \,|\, [y_i, x_{(i)}]\right)}{q\left([y_i, x_{i-1}] \,|\, [x_i, x_{i-1}]\right)} \\
&= \frac{f(y)}{f(x)} \cdot \frac{f(x_i \,|\, x_{(i)})}{f(y_i \,|\, x_{(i)})} \\
&= \frac{f(y)}{f(x)} \cdot \frac{f(x)/f(x_{(i)})}{f(y)/f(x_{(i)})} \\
&= 1.
\end{aligned}$$

Es decir, los vectores intermedios en el algoritmo de Gibbs siempre se aceptan, aunque sabemos que el algoritmo establece que se deben completar todas las coordenadas para producir un nuevo vector, el cual siempre es aceptado.

5.7. El muestreo de Gibbs por rebanadas

Esta es otra aplicación del muestreo de Gibbs. En el idioma inglés se le conoce con el nombre de *slice sampling*. Supongamos que sea desean obtener valores de una variable aleatoria X con función de densidad $f(x)$, la cual puede ser expresada de la siguiente forma

$$f(x) = a \cdot f_1(x) \cdots f_k(x), \quad \text{para} \;\; x \in A \subseteq \mathbb{R}, \tag{5.14}$$

en donde $a > 0$ es una constante y la funciones $f_1(x), \ldots, f_k(x)$ son no negativas, pero no son necesariamente funciones de densidad y $k \geqslant 1$ es un número entero. La función $f(x)$ es nuestra función objetivo. Observemos que su representación (5.14) no es única pues pueden modificarse los factores $f_1(x), \ldots, f_k(x)$ y, sin embargo, producir el mismo producto. Definamos el conjunto

$$S = \{ \ (u_1, \ldots, u_k, x) \quad : \quad 0 \leqslant u_1 \leqslant f_1(x), \qquad (5.15)$$
$$\vdots$$
$$0 \leqslant u_k \leqslant f_k(x), \ x \in A\}.$$

Este conjunto define un volumen en $\mathbb{R}^{k+1}$ y separa el área que determina cada una de las funciones $f_1(x), \ldots, f_k(x)$. Por ejemplo, la rebanada (*slice*) que define la función $f_1(x)$ es el conjunto $\{(x, u_1) : x \in A, \ 0 \leqslant u_1 \leqslant f_1(x)\}$, en donde hemos cambiado el orden de las coordenadas. Este conjunto se muestra en la Figura 5.8.

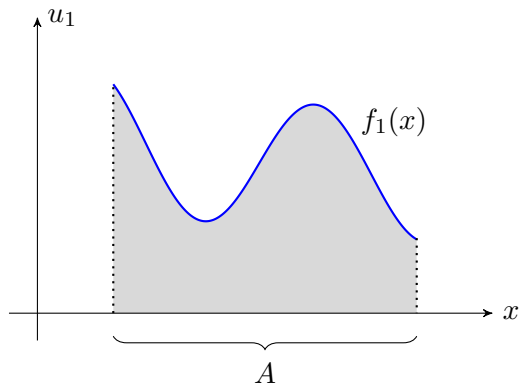


Figura 5.8: Conjunto $\{(x, u_1) : x \in A, \ 0 \leqslant u_1 \leqslant f_1(x)\}$.

Sea $(U_1, \ldots, U_k, X)$ un vector aleatorio con función de densidad uniforme sobre el volumen S, es decir, con función de densidad

$$g(u_1, \ldots, u_k, x) = \begin{cases} 1/|S| & \text{si} \quad (u_1, \ldots, u_k, x) \in S, \\ 0 & \text{en otro caso,} \end{cases}$$

en donde $|S|$ representa el volumen de S. Se comprobará más adelante que $|S| = 1/a$. Esta es una distribución multivariada de $k+1$ variables aleatorias y se puede aplicar el muestreo de Gibbs para generar sus valores. Las distribuciones condicionales son también uniforme y están dadas por:

- $g(u_1 \,|\, u_2, \ldots, u_k, x) = \dfrac{1/|S|}{\int_0^{f_1(x)} 1/|S|\, du_1} = \dfrac{1}{f_1(x)}, \qquad 0 \leqslant u_1 \leqslant f_1(x).$
- $g(u_2 \,|\, u_2, u_3, \ldots, u_k, x) = \dfrac{1/|S|}{\int_0^{f_2(x)} 1/|S|\, du_2} = \dfrac{1}{f_2(x)}, \qquad 0 \leqslant u_1 \leqslant f_2(x).$

$$\vdots$$

- $g(u_k \,|\, u_1, \ldots, u_{k-1}, x) = \dfrac{1/|S|}{\int_0^{f_k(x)} 1/|S|\, du_k} = \dfrac{1}{f_k(x)}, \qquad 0 \leqslant u_1 \leqslant f_k(x).$

Por último, la función de densidad condicional de x es

- $g(x \,|\, u_1, \ldots, u_k) = \dfrac{1/|S|}{\int_D 1/|S|\, du_1} = \dfrac{1}{|D|}, \qquad x \in D,$

en donde $D = \{x \in A : f_1(x) \geqslant u_1, \ldots, f_k(x) \geqslant u_k\}$ y $|D|$ es su longitud. Como todas estas distribuciones son uniformes, es sencillo la generación de sus valores y el muestreo de Gibbs puede aplicarse. Observemos también que la distribución marginal de X es la función objetivo pues, para $x \in A$,

$$\int_0^{f_1(x)} \cdots \int_0^{f_k(x)} g(u_1, \ldots, u_k, x)\, du_1 \cdots du_k = \frac{1}{|S|} \cdot f_1(x) \cdots f_k(x).$$

Este hecho implica que se puede usar el muestreo de Gibbs para generar valores de la función objetivo $f(x)$. Por otro lado, el cálculo anterior muestra que $a = 1/|S|$. En resumen, el muestreo de Gibbs se puede llevar a cabo de la siguiente manera.

Muestreo de Gibbs por rebanadas para una función de densidad de la forma $f(x) = a \cdot f_1(x) \cdots f_k(x)$ con $x \in A \subseteq \mathbb{R}$

Sea x_0 un valor inicial de $f(x)$.
Sea n el tamaño de la muestra requerida.
Para $t = 1, 2, \ldots, n$,

- Sea u_1^t un valor de la distribución unif $(0, f_1(x_{t-1}))$.
- Sea u_2^t un valor de la distribución unif $(0, f_2(x_{t-1}))$.

$\vdots$

- Sea u_k^t un valor de la distribución unif $(0, f_k(x_{t-1}))$.
- Defina $D_t = \{x \in A : f_1(x) \geqslant u_1^t, \ldots, f_k(x) \geqslant u_k^t\}$.
- Sea x_t un valor de la distribución unif (D_t).

Este algoritmo produce la sucesión de números $x_1, x_2, \ldots$ que se distribuye de acuerdo a la función de densidad $f(x)$ dada por (5.14). Observe que no es necesario conocer el valor de la constante a en la especificación de $f(x)$.

Ejemplo 5.7 (Distribución gama truncada)
Sea X una variable aleatoria con función de densidad gama (α, λ)*, en donde $\alpha > 1$, $\lambda = 1$ y está condicionada a tomar valores en el intervalo (c, ∞) para alguna constante $c > 0$. Este función de densidad tiene la forma*

$$f(x) = a \cdot x^{\alpha-1} \cdot e^{-x}, \quad \textit{para} \quad x \in (c, \infty).$$

para alguna constante $a > 0$. Véase la Figura 5.9. Para generar valores de esta distribución se pueden obtener valores de la distribución completa y aceptar sólo aquellos valores que excedan al valor c, sin embargo, si c es muy grande, ocurrirá un alto porcentaje de rechazos y el método se volvería muy ineficiente. Como un método alternativo se puede usar el muestreo de Gibbs por rebanadas. Explicaremos esto a continuación.

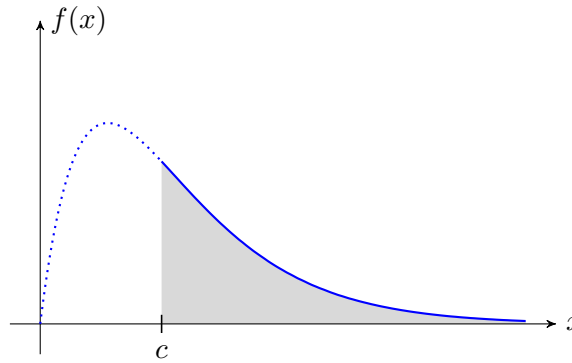


Figura 5.9: Densidad gama truncada.

Sean $f_1(x) = x^{\alpha-1}$ *y* $f_2(x) = e^{-x}$. *Sea* (U_1, U_2, X) *un vector aleatorio con distribución uniforme sobre el conjunto*

$$S = \left\{(u_1, u_2, x) : 0 \leqslant u_1 \leqslant x^{\alpha-1},\ 0 \leqslant u_2 \leqslant e^{-x},\ x > c\right\}.$$

Las funciones de densidad condicionales de U_1 *y* U_2 *son*

$$\begin{aligned} g(u_1 \,|\, u_2, x) &= \frac{1/|S|}{\int_0^{x^{\alpha-1}} 1/|S|\, du_1} = \frac{1}{x^{\alpha-1}}, \quad \textit{para}\ \ 0 \leqslant u_1 \leqslant x^{\alpha-1}, \\ g(u_2 \,|\, u_1, x) &= \frac{1/|S|}{\int_0^{e^{-x}} 1/|S|\, du_1} = \frac{1}{e^{-x}}, \quad \textit{para}\ \ 0 \leqslant u_2 \leqslant e^{-x}. \end{aligned}$$

Por último, para encontrar la función de densidad condicional de X *dados dos valores* u_1 *y* u_2, *se debe descubrir primero su soporte. Este conjunto se obtiene de la definición de* S. *Tenemos que*

$$\begin{aligned} D &= \left\{x : 0 \leqslant u_1 \leqslant x^{\alpha-1},\ 0 \leqslant u_2 \leqslant e^{-x},\ x > c\right\} \\ &= \left\{x : x \geqslant u_1^{1/(\alpha-1)},\ x \leqslant -\ln u_2,\ x > c\right\} \\ &= \left\{x : \max\left(c, u_1^{1/(\alpha-1)}\right) \leqslant x \leqslant -\ln u_2\right\}. \end{aligned}$$

Por lo tanto,

$$g(x \,|\, u_1, u_2) = \frac{1/|S|}{\int_D 1/|S|\, dx} = \frac{1}{|D|}, \quad \textit{para}\ \ x \in D,$$

en donde $|D|$ es la longitud del intervalo D. Así, el algoritmo es el siguiente.

Muestreo de Gibbs por rebanadas
para una función de densidad gama (α, λ)
con $\alpha > 1$, $\lambda = 1$, **y truncada al intervalo** (c, ∞).

Sea $x_0 > c$ un valor inicial.
Sea n el tamaño de la muestra requerida.
Para $t = 1, 2, \ldots, n$,

- *Sea u_1^t un valor de la distribución* unif $(0, (x_{t-1})^{\alpha-1})$.
- *Sea u_2^t un valor de la distribución* unif $(0, e^{-x_{t-1}})$.
- *Sea x_t un valor de la distribución* unif $(\max(c, (u_1^t)^{1/(\alpha-1)}), -\ln u_2^t)$.

Este algoritmo produce la sucesión $x_1, \ldots, x_n$ que sigue la densidad indicada. En la sección de ejercicios se pide verificar numéricamente esta afirmación.

■

5.8. El muestreo de Gibbs por completación

Suponga que se desea obtener una muestra de valores de una variable aleatoria X con un cierta función de densidad $f(x)$. Suponga además que es posible encontrar una función de densidad conjunta $f(x, y)$ tal que

$$\int_{-\infty}^{\infty} f(x, y)\, dy = f(x).$$

Entonces el muestreo de Gibbs resultará de utilidad cuando se conozcan mecanismos para obtener muestras de las funciones de densidades condicionales $x \mapsto f(x \mid y)$ y $y \mapsto f(y \mid x)$. A tales funciones conjuntas se les llama completaciones de $f(x)$.

Definición 5.15 *Sea $f(x)$ una función de densidad univariada. A cualquier función de densidad conjunta $f(x, y)$ que tenga como función de densidad marginal a $f(x)$ se le llama una completación de $f(x)$.*

Por simplicidad hemos mencionado completaciones bidimensionales pero éstas pueden ser de dimensiones mayores $f(x, y_1, \ldots, y_k)$, con $k \geqslant 1$. El muestreo por rebanadas estudiado en la sección anterior es un ejemplo de esta situación. Regresando al caso bidimensional, el muestreo de Gibbs producirá una sucesión de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de la función de densidad conjunta $f(x, y)$, de donde se pueden extraer los valores de cada variable por separado, en particular, los valores de X que se buscan. Veamos algunos ejemplos.

Ejemplo 5.8 (Distribución gama)
Sea $\alpha > 0$ una constante y suponga que (X, Y) es un vector aleatorio con función de densidad

$$f(x,y) = \begin{cases} \dfrac{1}{\Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-y} & si \;\; 0 < x < y, \\ 0 & en\ otro\ caso. \end{cases}$$

Se puede comprobar que las funciones de densidad marginales son

$$\begin{aligned} X &\sim \text{gama}\,(\alpha, 1), \\ Y &\sim \text{gama}\,(1+\alpha, 1). \end{aligned}$$

Es decir, $f(x, y)$ es una completación de cualquiera de estas dos funciones de densidad univariadas. Tampoco es difícil verificar que las funciones de densidad condicionales son

$$x \mapsto f(x \,|\, y) = \frac{f(x,y)}{f(y)} = \alpha\, y^{-\alpha}\, x^{\alpha-1}, \quad para \;\; x \in (0, y), \tag{5.16}$$

$$y \mapsto f(y \,|\, x) = \frac{f(x,y)}{f(x)} = e^{x}\, e^{-y}, \quad para \;\; y \in (x, \infty). \tag{5.17}$$

Para obtener un valor de la función de densidad $f(x \,|\, y)$ se puede usar el método de la función inversa y para obtener un valor de $f(y \,|\, x)$ se debe reconocer que se trata de una distribución exponencial truncada y se puede aplicar el muestreo de Gibbs por rebanadas como se mostró en la sección anterior. De esta manera, el algoritmo es el siguiente.

Muestreo de Gibbs por completación
(Ejemplo 5.8 Distribución gama)

Sea $x_0 > 0$ un valor inicial.
Sea n el tamaño de la muestra requerida.
Para $t = 1, 2, \ldots, n$,

- *Sea y_t un valor de $f_{Y|X}(y \mid x_{t-1})$ dada por (5.16).*
- *Sea x_t un valor de $f_{X|Y}(x \mid y_t)$ dada por (5.17).*

Este algoritmo produce la sucesión $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, en donde las primeras coordenadas son valores de $X \sim \text{gama}(\alpha, 1)$, y las segundas coordenadas son valores de $Y \sim \text{gama}(a + \alpha, 1)$. Finalmente, recordemos que si $X \sim \text{gama}(\alpha, 1)$, entonces $X/\lambda \sim \text{gama}(\alpha, \lambda)$. Esto completa otro mecanismo para generar valores de la distribución gama con parámetros arbitrarios.

■

5.9. Ejercicios

Conceptos de cadenas de Markov

112. Encuentre todas las distribuciones estacionarias de las siguientes cadenas de Markov:

a) $\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$.

b) $\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

c) $\mathbf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$.

d) $\mathbf{P} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix}$.

113. Calcule la distribución límite, cuando exista, de las siguientes cadenas de Markov:

a) $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

b) $\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

c) $\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

d) $\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$.

El algoritmo Metropolis-Hastings para distribuciones discretas

114. Este ejercicio tiene el objetivo de ilustrar el algoritmo Metropolis Hastings en un caso discreto simple.

a) Escoja una distribución objetivo $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ a su elección, con soporte el conjunto de estados $\{1, 2, 3, 4, 5\}$, distinta de la uniforme.

b) Partiendo del algún valor $x_0 \in \{1, 2, 3, 4, 5\}$ y usando la matriz Q (función propuesta) que aparece abajo, elabore un programa de cómputo que genere 150 valores $x_1, \ldots, x_{150}$ mediante el algoritmo Metropolis-Hastings.

$$\mathbf{Q} = \begin{pmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

c) Encuentre explícitamente la matriz $\mathbf{P} = (p_{ij})$ de la cadena de Markov y justifique que es convergente.

d) Elabore un histograma de probabilidad de los valores $x_1, \ldots, x_{50}$ y sobreponga la gráfica de la función objetivo.

e) Elabore un histograma de probabilidad de los valores $x_{51}, \ldots, x_{100}$ y sobreponga la gráfica de la función objetivo.

f) Elabore un histograma de probabilidad de los valores $x_{101}, \ldots, x_{150}$ y sobreponga la gráfica de la función objetivo.

g) Con base en las 3 gráficas anteriores, determine subjetivamente si la última muestra de tamaño 50 es mejor (en algún sentido) que las anteriores.

Cadenas de Markov con espacio de estados continuo

115. Calcule la distribución estacionaria para la función de densidad de transición normal que aparece abajo.

$$p(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-(y-x)^2/2\right\}, \quad -\infty < x, y < \infty.$$

El algoritmo Metropolis-Hastings para distribuciones continuas

116. Sea $\alpha(x,y)$ la probabilidad de aceptación en el algoritmo Metropolis-Hastings. Demuestre que

$$\int_S \alpha(x,y)dy = 1.$$

El muestreo de Gibbs

117. Sea (X,Y) un vector aleatorio con función de densidad

$$f(x,y) = \frac{1}{\pi} \exp\{-x^2 - 2xy - 2y^2\}, \quad -\infty < x, y < \infty.$$

a) Encuentre las densidades marginales de X y Y.

b) Utilice el muestreo de Gibbs para obtener valores del vector (X,Y).

118. Sea (X,Y) un vector aleatorio con función de densidad

$$f(x,y) = \frac{1}{\pi\, e^5} \exp\{-x^2 - y^2 + 2x - 4y\}, \quad -\infty < x, y < \infty.$$

a) Encuentre las densidades marginales de X y Y.

b) Utilice el muestreo de Gibbs para obtener valores del vector (X,Y).

119. Sea (X,Y) un vector aleatorio con distribución tal que se verifican las siguientes distribuciones condicionales:

$$\begin{aligned} X \,|\, (Y=y) &\sim \mathrm{N}(\rho y, \sigma^2), \\ Y \,|\, (X=x) &\sim \mathrm{N}(\rho x, \sigma^2), \end{aligned}$$

en donde ρ es una constante.

a) Encuentre la distribución conjunta $f_{X,Y}(x,y)$.

b) Utilice el muestreo de Gibbs para obtener valores del vector (X,Y).

c) Dado un conjunto de observaciones $x_1,\ldots,x_n$ de X, estime la distribución a posteriori para Y usando el muestreo de Gibbs.

El muestreo de Gibbs por rebanadas

120. Considerando la factorización sugerida, implemente en una computadora el algoritmo del muestreo de Gibbs por rebanadas para obtener valores de las distribuciones que aparecen abajo. Con los datos obtenidos elabore un histograma y compare visualmente con la gráfica de la función de densidad en el caso unidimensional.

a) $f(x) = 3\cdot x\cdot x, \quad \text{si } \ 0<x<1.$

b) $f(x) = \dfrac{6}{5}\cdot x\cdot(x-1), \quad \text{si } \ 1<x<2.$

c) $f(x) = 2\cdot x\cdot \exp\{-x^2\}, \quad \text{si } \ 0<x<\infty.$

d) $f(x) = \dfrac{1}{\sqrt{\pi}}\cdot(1+\cos(x))\cdot\exp\{-x^2\}, \quad \text{si } \ -\infty<x,y<\infty.$

e) $f(x,y) = \dfrac{1}{\pi}\cdot\exp\{-x^2\}\cdot\exp\{-y^2\}, \quad \text{si } \ -\infty<x,y<\infty.$

f) $f(x,y,z) = \dfrac{1}{\pi^{3/2}}\cdot\exp\{-x^2\}\cdot\exp\{-y^2\}\cdot\exp\{-z^2\},$
si $\ -\infty<x,y,z<\infty.$

El muestreo de Gibbs por completación

121. Asigne valores particulares a los parámetros $\alpha>0$ y $\lambda>0$ e implemente en una computadora el muestreo de Gibbs por completación que aparece en Ejemplo 5.8 para obtener valores de la distribución gama(α,λ). Elabore un histograma y compare visualmente con la gráfica de la función de densidad.

122. *Distribución beta.* Sean $\alpha>1$ y $\beta>0$ dos parámetros y suponga que (X,Y) es un vector aleatorio con función de densidad

$$f(x,y) = \frac{\alpha-1}{B(\alpha,\beta)}\, y^{\alpha-2}(1-x)^{\beta-1}, \quad \text{si } \ 0<y<x<1.$$

a) Compruebe que $f(x, y)$ es una completación de la distribución beta demostrando que $X \sim \text{beta}(\alpha, \beta)$ y $Y \sim \text{beta}(\alpha - 1, \beta + 1)$.

b) Encuentre las funciones de densidad condicionales $f(x \,|\, y)$ y $f(y \,|\, x)$.

c) Asigne valores particulares de $\alpha > 1$ y $\beta > 0$. A partir de un punto inicial (x_0, y_0) tal que $0 < y_0 < x_0 < 1$, implemente el muestreo de Gibbs para obtener valores de X.

d) Elabore un histograma con los datos obtenidos y compare con la gráfica de $f(x)$.

123. Sea (X, Y) un vector aleatorio con función de densidad

$$f(x, y) = \frac{1}{\pi}, \quad \text{si} \;\; 0 < x^2 + y^2 < 1.$$

a) Encuentre las funciones de densidad marginales $f(x)$ y $f(y)$.

b) Encuentre las funciones de densidad condicionales $f(x \,|\, y)$ y $f(y \,|\, x)$.

c) Aplique el muestreo de Gibbs para obtener valores de X y Y.

d) Elabore histogramas de los datos obtenidos y compare con las respectivas funciones de densidad.

124. Suponga que se desean obtener valores de una variable aleatoria X con función de densidad

$$f(x) = xe^{-x}, \quad \text{si} \;\; 0 < x < \infty.$$

a) Defina la función de densidad bivariada $f(x, y)$ como aparece abajo y compruebe que la densidad marginal de X es $f(x)$.

$$f(x, y) = e^{-x}, \quad \text{si} \;\; 0 < y < x < \infty.$$

b) Encuentre las funciones de densidad condicionales $f(x \,|\, y)$ y $f(y \,|\, x)$.

c) Aplique el muestreo de Gibbs para obtener valores de X.

d) Elabore un histograma con los datos obtenidos y compare con la gráfica de $f(x)$.

Capítulo 6

Aplicaciones en la estadística

En este último capítulo revisaremos de manera breve algunos procedimientos de la estadística en donde se puede usar la simulación como parte de ciertos algoritmos para el análisis y solución del problema de la estimación de parámetros. En particular estudiaremos dos algoritmos de estimación en presencia de datos faltantes (*missing data*). Estos son: algoritmo EM y algoritmo *data augmentation*. Finalizaremos nuestro trabajo mostrando dos técnicas de remuestreo llamadas *bootstrap* y *jackknife*.

6.1. Introducción al problema de datos faltantes

Sea $f(x;\theta)$ una distribución que depende de un parámetro desconocido θ. Un problema clásico de la estadística consiste en estimar θ a partir de una colección de valores de una variable aleatoria con esta distribución. Por simplicidad denotaremos a estos valores mediante el vector $x = (x_1, \ldots, x_n)$. Usualmente, los métodos para resolver este problema requieren que todas las observaciones estén completamente especificadas, es decir, que se conozcan sus valores exactos.

Sin embargo, en problemas prácticos los valores de la variable aleatoria de interés no se observan directamente del fenómeno en estudio y por diversas razones, algunos datos no aparecen totalmente especificados en su valor y sólo se conoce cierta información parcial de ellos. A estos datos se les denomina datos faltantes o incompletos. Existen diferentes formas acerca de

la incompletitud de la información en un dato o un conjunto de datos. A continuación explicaremos algunas de ellas.

- En ciertas situaciones el conjunto de datos $x = (x_1, \ldots, x_n)$ no son directamente observables u obtenibles y sólo es observable otra colección de datos $y = (y_1, \ldots, y_m)$, no necesariamente con el mismo número de observaciones que los datos originales, pero que son una transformación de los primeros. Esto se puede deber a que nuestros instrumentos de medición no son precisos, o a que no nos es posible observar directamente la variable de interés. Esta situación es similar a la de observar el mundo a través de un cristal que distorsiona la realidad. Dado que únicamente podemos observar los datos del vector y, a estos datos se les llama datos observables. Por ejemplo, en lugar de los datos numéricos $x = (x_1, \ldots, x_n)$, sólo es posible observar el vector $y = (y_1, y_2) = (x_{(1)}, x_{(n)})$, es decir, el mínimo y el máximo. Como otro ejemplo extremo, es posible que la única información observable sea el número $(x_1 + \cdots + x_n)/n$, es decir, la media muestral de los datos completos. Estas transformaciones de datos puede presentarse para la totalidad de los datos originales o sólo para una parte de ellos.

- La falta de información de un dato x_i puede ser completa en el sentido de ser totalmente faltante. Sabemos que una observación o medición debió efectuarse, pero no sabemos nada de su valor.

- En ocasiones no se conoce el valor exacto de un dato u observación particular x_i, pero se conoce, por ejemplo, que $x_i \in (a, \infty)$ para alguna constante a. En otras palabras, sólo se conoce que $x_i > a$, pero no se conoce su valor exacto. En este caso se dice que el dato x_i presenta censura por la derecha. Por ejemplo, suponga que en un estudio se está observando el tiempo en el que ocurre la primera falla en equipos eléctricos y que, por cuestiones de costo y tiempo, se decide concluir las observaciones al tiempo $a > 0$. Si el equipo eléctrico i no ha tenido su primera falla antes de a, entonces la observación x_i quedará censurada por la derecha, es decir, sólo se registrará que $x_i \in (a, \infty)$.

- También puede ocurrir que la única información disponible para un dato x_i sea que $x_i \in (-\infty, a)$, para alguna constante a. Esta es la censura por la izquierda. Por ejemplo, supongamos que la variable en

estudio es la edad en la que las personas desarrollan un cierto tipo de cáncer. En la mayoría de los casos, el cáncer se detecta sólo hasta que la persona acude al médico. De este modo, la mayoría de las observaciones acerca de la edad x_i en la que inicia el desarrollo del cáncer para la persona i no estarán completamente especificadas, sólo se sabrá que $x_i \in (0, a_i)$, en donde a_i es la edad en la que el individuo i acudió al médico y se diagnosticó que padece cáncer.

- También puede presentarse la censura por intervalo, la cual ocurre para un dato x_i cuando sólo se conoce que $x_i \in (a, b)$, para ciertas constantes a y b. Por ejemplo, discretizar una variable continua puede considerarse como una censura por intervalo. En particular, debido a la precisión limitada de los instrumentos de medición, cualquier registro de una variable continua implica necesariamente una censura por intervalo. Por ejemplo, si para medir una distancia sólo contamos con un instrumento cuya precisión máxima indica milímetros, cualquier observación estará censurada por un intervalo de longitud 1 milímetro.

Debemos mencionar que existen otras formas en las que un dato o un conjunto de datos puede presentar deficiencias en su registro. Además, distintos tipos de incompletitud pueden tenerse para un mismo dato. También puede ocurrir que sólo una parte de los datos puede presentar un tipo de incompletitud y otra parte de los datos puede mostrar otro tipo de deficiencia.

Como puede verse, la variedad de incompletitud de información en un conjunto de datos puede ser muy amplia. El problema general consiste en estimar el parámetro o cantidad desconocida θ, a partir de la información posiblemente incompleta que proveen los datos que se pueden observar. La dificultad del problema y el procedimiento a seguir puede ser diferente en cada caso.

6.2. El algoritmo EM

El algoritmo EM (*Expectation-Maximization*) es un procedmiento iterativo cuyo objetivo es encontrar el estimador máximo verosímil de un parámetro desconocido en un modelo. Es particularmente útil en situaciones en donde

existen datos incompletos en la muestra.

Sea $f(x;\theta)$ una distribución que depende de un parámetro desconocido θ y el cual deseamos estimar. Por conveniencia, escribiremos $f(x\,|\,\theta)$. Supondremos que contamos con dos colecciones de datos: un primer conjunto de n_1 datos completos $x = (x_1, \ldots, x_{n_1})$, y un segundo conjunto de n_2 datos incompletos $z = (z_1, \ldots, z_{n_2})$, los cuales no se conocen completamente pues presentan alguna deficiencia en su especificación. La escritura de los datos z sólo es indicativa pues no conocemos sus valores. Supondremos que $n = n_1 + n_2$. Cada iteración del algoritmo EM consiste en los dos pasos que se muestran en la Tabla 6.1.

Paso E (*Expectation*)	Se estiman los datos incompletos z a partir de los datos observados x y de algún valor preliminar θ_0 del parámetro.
Paso M (*Maximization*)	Se maximiza la función de verosimilitud bajo la hipótesis de que los datos incompletos son conocidos. Estos datos faltantes son las estimaciones del paso anterior.

Tabla 6.1: Pasos en el Algoritmo EM.

Consideraremos dos funciones de verosimilitud para θ:

- $L(\theta\,|\,x) = f(x\,|\,\theta)$. Esta es la función de verosimilitud considerando sólo los datos observados x.
- $L(\theta\,|\,x, z) = f(x, z\,|\,\theta)$. Esta es la función de verosimilitud conjunta de los datos observados x y los datos incompletos z.

Estamos usando la misma expresión $f(\cdot\,|\,\theta)$ para denotar a estas funciones pero en realidad son distintas. A partir de estas dos funciones y dado un valor de θ, la verosimilitud de un valor $(z_1, \ldots, z_{n_2})$ de los datos incompletos z es

$$f(z\,|\,\theta, x) = \frac{f(x, z\,|\,\theta)}{f(x\,|\,\theta)}.$$

Con estos elementos podemos ahora explicar el procedimiento iterativo del algoritmo EM.

Paso E (*Expectation*)
Cálculo de la función Q

Sea $\hat{\theta}_0$ una estimación inicial para θ. Este valor puede ser alguna propuesta, o bien, puede estimarse con base en los datos completamente especificados $x = (x_1, \ldots, x_{n_1})$. El logaritmo de la función de verosimilitud de los datos observados es

$$
\begin{aligned}
\ln L(\theta \,|\, x) &= \ln L(\theta \,|\, x) \cdot \underbrace{\int f(z \,|\, \hat{\theta}_0, x)\, dz}_{=1} \\
&= \int [\ln L(\theta \,|\, x)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&= \int [\ln f(x \,|\, \theta)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&= \int \left[\ln \frac{f(x, z \,|\, \theta)}{f(z \,|\, \theta, x)}\right] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&= \int [\ln f(x, z \,|\, \theta)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&\qquad - \int [\ln f(z \,|\, \theta, x)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&= \int [\ln L(\theta \,|\, x, z)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&\qquad - \int [\ln f(z \,|\, \theta, x)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz \\
&= \underbrace{E\left[\ln L(\theta \,|\, x, z) \,\middle|\, \hat{\theta}_0, x\right]}_{=Q(\theta \,|\, \hat{\theta}_0, x)} \\
&\qquad - \int [\ln f(z \,|\, \theta, x)] \cdot f(z \,|\, \hat{\theta}_0, x)\, dz,
\end{aligned}
$$

en donde la esperanza $E\left[\cdot \,\middle|\, \hat{\theta}_0, x\right]$ se calcula respecto de la función de densidad condicional $f(z \,|\, \hat{\theta}_0, x)$. A esta esperanza se le define como la función Q.

Definición 6.1 *La función Q en el algoritmo EM es la esperanza*

$$\theta \mapsto Q(\theta \,|\, \hat{\theta}_0, x) = E\left[\ln L(\theta \,|\, x, z) \,\middle|\, \hat{\theta}_0, x \right].$$

Paso M (*Maximization*)
Maximización de la función Q

Se trata ahora de encontrar el punto en donde la función $\theta \mapsto Q(\theta \,|\, \hat{\theta}_0, x)$ alcanza su máximo. Esto se escribe de la siguiente forma

$$\hat{\theta}_1 = \arg \max_\theta Q(\theta \,|\, \hat{\theta}_0, x).$$

Es decir, $\hat{\theta}_1$ representa el punto en donde la función $Q(\theta \,|\, \hat{\theta}_0, x)$ es máxima. De este modo, a partir de $\hat{\theta}_0$ se obtiene $\hat{\theta}_1$. Procediendo de manera secuencial, se repiten los pasos E y M, y se encuentra la sucesión de estimaciones $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \ldots$ Bajo ciertas condiciones esta sucesión numérica es convergente.

Una versión Monte Carlo del algoritmo EM

En ocasiones la integral múltiple $Q(\theta \,|\, \hat{\theta}_0, x)$ puede ser difícil de calcular y se ha propuesto usar el método de Monte Carlo para aproximarla. Dada una estimación $\hat{\theta}_0$, sean $z^{(1)}, \ldots, z^{(m)}$ valores generados usando la función de densidad $f(z \,|\, \hat{\theta}_0, x)$. La integral $Q(\theta \,|\, \hat{\theta}_0, x)$ se reemplaza por su aproximación Monte Carlo

$$Q(\theta \,|\, \hat{\theta}_0, x) = \frac{1}{m} \sum_{i=1}^{m} \ln L(\theta \,|\, x, z^{(i)}).$$

Se procede después con el paso M, es decir, se busca maximizar $Q(\theta \,|\, \hat{\theta}_0, x)$ para encontrar $\hat{\theta}_1$, y así se procede de manera sucesiva.

En los ejemplos que aparecen a continuación se busca mostrar los cálculos del algoritmo EM para entender su funcionamiento. Las integrales $Q(\theta \,|\, \hat{\theta}_0, x)$ son fáciles de encontrar y no es necesario emplear la aproximación Monte

Carlo. Consideraremos el modelo $\mathrm{N}(\theta, 1)$, en donde nos interesa estimar θ. Primero estudiaremos el caso cuando tenemos datos completos pero también tenemos datos faltantes. En el segundo ejemplo, supondremos que los datos están censurados por la derecha. Los cálculos son muy similares en ambos casos.

Ejemplo 6.1 (Estimación de la media de una distribución normal con algunos datos faltantes) *Sea $\underline{x} = (x_1, \ldots, x_{n_1})$ un vector de n_1 observaciones independientes de la distribución $\mathrm{N}(\theta, 1)$, en donde θ es desconocido y se desea estimar por máxima verosimilitud. Denotaremos por $f(x \,|\, \theta)$ a la función de densidad $\mathrm{N}(\theta, 1)$.*

Sea $\underline{z} = (z_1, \ldots, z_{n_2})$ un vector de n_2 observaciones faltantes de la misma distribución $\mathrm{N}(\theta, 1)$, generadas de manera independiente, e independientes de $\underline{x} = (x_1, \ldots, x_{n_1})$. Los valores $z_1, \ldots, z_{n_2}$ no se conocen.

De este modo, los datos completos son $\underline{x} = (x_1, \ldots, x_{n_1})$ y $\underline{z} = (z_1, \ldots, z_{n_2})$, los cuales producen la función de verosimilitud

$$L(\theta \,|\, \underline{x}, \underline{z}) = f(\underline{x}, \underline{z} \,|\, \theta) = \prod_{i=1}^{n_1} f(x_i \,|\, \theta) \cdot \prod_{i=1}^{n_2} f(z_i \,|\, \theta).$$

La última expresión se obtiene a partir de las hipótesis de independencia. Reiteramos que se usa la misma expresión $f(\cdot)$ para denotar a las funciones de densidad vectoriales $f(\underline{x} \,|\, \theta)$, $f(\underline{z} \,|\, \theta)$, $f(\underline{x}, \underline{z} \,|\, \theta)$, así como a la funciones de una variable $f(x_i \,|\, \theta)$ y $f(z_i \,|\, \theta)$.

Los datos observados son solamente $\underline{x} = (x_1, \ldots, x_{n_1})$, lo que produce la función de verosimilitud observada

$$L(\theta \,|\, \underline{x}) = f(\underline{x} \,|\, \theta) = \prod_{i=1}^{n_1} f(x_i \,|\, \theta).$$

Por la definición de función de densidad condicional y usando las dos funciones de verosimilitud anteriores, tenemos que

$$f(\underline{z} \,|\, \theta, \underline{x}) = \frac{f(\underline{x}, \underline{z} \,|\, \theta)}{f(\underline{x} \,|\, \theta)} = \prod_{i=1}^{n_2} f(z_i \,|\, \theta).$$

Debido a la hipótesis de independencia entre los datos observados y los datos faltantes, la función $f(\underline{z}\,|\,\theta, \underline{x})$ *se puede escribir más brevemente como* $f(\underline{z}\,|\,\theta)$. *Con esta información podemos ahora aplicar el algoritmo EM para estimar* θ *de manera secuencial.*

Paso E (Cálculo de la función Q). *Sea* $\hat{\theta}_0$ *un valor inicial para* θ. *Por definición de la función* Q *y después usando la expresión para* $L(\theta\,|\,\underline{x}, \underline{z})$, *tenemos que*

$$\begin{aligned}
Q(\theta\,|\,\hat{\theta}_0, \underline{x}) &= E\left[\ \ln L(\theta\,|\,\underline{x}, \underline{z}) \ \middle|\ \hat{\theta}_0, \underline{x}\ \right] \\
&= E\left[\sum_{i=1}^{n_2} \ln f(z_i\,|\,\theta) + \sum_{i=1}^{n_1} \ln f(x_i\,|\,\theta) \middle|\ \hat{\theta}_0, \underline{x}\ \right] \\
&= \sum_{i=1}^{n_2} E\left[\ \ln f(z_i\,|\,\theta) \ \middle|\ \hat{\theta}_0, \underline{x}\ \right] + \sum_{i=1}^{n_1} \ln f(x_i\,|\,\theta) \\
&= \sum_{i=1}^{n_2} E\left[\ \ln f(z_i\,|\,\theta) \ \middle|\ \hat{\theta}_0\ \right] + \sum_{i=1}^{n_1} \ln f(x_i\,|\,\theta) \\
&= n_2\, E\left[\ \ln f(z\,|\,\theta) \ \middle|\ \hat{\theta}_0\ \right] + \sum_{i=1}^{n_1} \ln f(x_i\,|\,\theta),
\end{aligned}$$

en donde, recordemos, la esperanza indicada se calcula respecto de la función de densidad multivariada $f(\underline{z}\,|\,\theta, \underline{x}) = f(\underline{z}\,|\,\theta)$, *con* $\underline{z} = (z_1, \ldots, z_{n_2}) \in \mathbb{R}^{n_2}$. *Como para cada integrando en donde aparece* z_i *la esperanza es la misma, se ha escrito como* $E[\ \ln f(z\,|\,\theta)\,|\,\hat{\theta}_0\ ]$, *la cual se calcula respecto de la función de densidad univariada* $f(\underline{z}\,|\,\hat{\theta}_0)$, $z \in \mathbb{R}$. *Observe que las observaciones incompletas* z_i *desaparecen de la última expresión al tomar la esperanza indicada.*

Paso M (Maximización de la función Q). *Derivando la función* Q *respecto de* θ,

$$\frac{\partial}{\partial\theta}\, Q(\theta\,|\,\hat{\theta}_0, \underline{x}) = n_2 \int_{-\infty}^{\infty} \left[\frac{\frac{\partial}{\partial\theta} f(z\,|\,\theta)}{f(z\,|\,\theta)}\right] \cdot f(z\,|\,\hat{\theta}_0)\, dz + \sum_{i=1}^{n_1} \frac{\frac{\partial}{\partial\theta} f(x_i\,|\,\theta)}{f(x_i\,|\,\theta)},$$

en donde, como $f(x\,|\,\theta)$ *es la función de densidad* $\mathrm{N}(\theta, 1)$, *se cumple que*

$$\frac{\partial}{\partial\theta}\, f(x\,|\,\theta) = f(x\,|\,\theta)\,(x - \theta).$$

Por lo tanto,

$$\frac{\partial}{\partial\theta}\, Q(\theta \,|\, \hat{\theta}_0, \underline{x}) = n_2 \int_{-\infty}^{\infty} (z-\theta)\cdot f(z \,|\, \hat{\theta}_0)\, dz + \sum_{i=1}^{n_1}(x_i - \theta).$$

Igualando a cero,

$$n_2\,(\hat{\theta}_0 - \theta) + n_1\bar{x} - n_1\theta = 0.$$

Resolviendo para θ se encuentra que el punto en donde se maximiza la función Q es

$$\hat{\theta}_1 = \frac{n_1}{n}\,\bar{x} + \frac{n_2}{n}\,\hat{\theta}_0,$$

en donde $n = n_1 + n_2$. En general, se tiene la relación

$$\hat{\theta}_{m+1} = \frac{n_1}{n}\,\bar{x} + \frac{n_2}{n}\,\hat{\theta}_m, \quad \textit{para} \;\; m = 0, 1, \ldots$$

De esta manera se obtiene la sucesión de estimaciones $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \ldots$ Tomando el límite cuando $m \to \infty$ y suponiendo que $\hat{\theta}$ es el valor límite, se obtiene la relación $\hat{\theta} = (n_1/n)\bar{x} + (n_2/n)\hat{\theta}$, cuya solución es $\hat{\theta} = \bar{x}$. En palabras, el algoritmo EM produce una sucesión de estimaciones del parámetro desconocido que converge a la media muestral de los datos observados. ∎

Ejemplo 6.2 (Estimación de la media de una distribución normal con algunos datos censurados por la derecha) *Sea $\underline{x} = (x_1, \ldots, x_{n_1})$ un vector de n_1 observaciones independientes de la distribución $\mathrm{N}(\theta, 1)$, en donde θ es desconocido y se desea estimar por máxima verosimilitud. Denotaremos por $f(x \,|\, \theta)$ a la función de densidad $\mathrm{N}(\theta, 1)$.*

Sea $\underline{z} = (z_1, \ldots, z_{n_2})$ un vector de n_2 observaciones censuradas, en donde sólo se sabe que ocurre el evento $(z_j > a)$ para alguna constante conocida a, $j = 1, 2, \ldots, n_2$. Es decir, sólo se sabe que estas observaciones toman un valor en el intervalo (a, ∞), pero no se conocen sus valores exactos. Todos estos datos están censurados por la derecha por una misma constante a.

De este modo, los datos completos son $\underline{x} = (x_1, \ldots, x_{n_1})$ y $\underline{z} = (z_1, \ldots, z_{n_2})$, aunque estos últimos son desconocidos. La función de verosimilitud completa es

$$L(\theta \,|\, \underline{x}, \underline{z}) = f(\underline{x}, \underline{z} \,|\, \theta) = \prod_{i=1}^{n_1} f(x_i \,|\, \theta) \cdot \prod_{i=1}^{n_2} f(z_i \,|\, \theta).$$

Por otro lado, los datos observados son $\underline{x} = (x_1, \ldots, x_{n_1})$ *y* $(z_j > a)$ *para* $j = 1, 2, \ldots, n_2$, *los cuales producen la función de verosimilitud observada*

$$L(\theta \,|\, \underline{x}) = f(\underline{x} \,|\, \theta) = (1 - F(a \,|\, \theta))^{n_2} \cdot \prod_{i=1}^{n_1} f(x_i \,|\, \theta).$$

Por la definición de función de densidad condicional y usando las dos funciones de verosimilitud anteriores, tenemos que

$$\begin{aligned} f(\underline{z} \,|\, \theta, \underline{x}) &= \frac{f(\underline{x}, \underline{z} \,|\, \theta)}{f(\underline{x} \,|\, \theta)} \\ &= \frac{1}{(1 - F(a \,|\, \theta))^{n_2}} \prod_{i=1}^{n_2} f(z_i \,|\, \theta) \\ &= \prod_{i=1}^{n_2} \frac{f(z_i \,|\, \theta)}{(1 - F(a \,|\, \theta))^{n_2}}. \end{aligned}$$

Es decir, la verosimilitud de cada z_i *es la distribución* $\mathrm{N}(\theta, 1)$ *truncada al intervalo* (a, ∞), *eso es,*

$$f(z_i \,|\, \theta, \underline{x}) = \frac{f(z_i \,|\, \theta)}{1 - F(a \,|\, \theta)}, \quad \textit{para} \;\; z_i > a.$$

A continuación aplicaremos el algoritmo EM para estimar θ *de manera secuencial.*

Paso E (Cálculo de la función Q**).** *Sea* $\hat{\theta}_0$ *un valor inicial para* θ. *Por definición de la función* Q *y después usando la expresión para* $L(\theta \,|\, \underline{x}, \underline{z})$,

tenemos que

$$\begin{aligned}
Q(\theta \,|\, \hat{\theta}_0, \underline{x}) &= E\left[\ln L(\theta \,|\, \underline{x}, \underline{z}) \;\middle|\; \hat{\theta}_0, \underline{x} \right] \\
&= E\left[\sum_{i=1}^{n_2} \ln f(z_i \,|\, \theta) + \sum_{i=1}^{n_1} \ln f(x_i \,|\, \theta) \;\middle|\; \hat{\theta}_0, \underline{x} \right] \\
&= \sum_{i=1}^{n_2} E\left[\ln f(z_i \,|\, \theta) \;\middle|\; \hat{\theta}_0, \underline{x} \right] + \sum_{i=1}^{n_1} \ln f(x_i \,|\, \theta) \\
&= n_2 \, E\left[\ln f(z \,|\, \theta) \;\middle|\; \hat{\theta}_0, \underline{x} \right] + \sum_{i=1}^{n_1} \ln f(x_i \,|\, \theta) \\
&= n_2 \int_{-\infty}^{\infty} \ln f(z \,|\, \theta) \cdot f(z \,|\, \hat{\theta}_0, \underline{x}) \, dz + \sum_{i=1}^{n_1} \ln f(x_i \,|\, \theta) \\
&= n_2 \int_{-\infty}^{\infty} \ln f(z \,|\, \theta) \cdot \frac{f(z \,|\, \hat{\theta}_0)}{1 - F(a \,|\, \hat{\theta}_0)} \, dz + \sum_{i=1}^{n_1} \ln f(x_i \,|\, \theta).
\end{aligned}$$

Paso M (Maximización de la función Q). *Derivando la función Q respecto de θ,*

$$\frac{\partial}{\partial \theta} Q(\theta \,|\, \hat{\theta}_0, \underline{x}) = n_2 \int_{-\infty}^{\infty} \left[\frac{\frac{\partial}{\partial \theta} f(z \,|\, \theta)}{f(z \,|\, \theta)} \right] \cdot \frac{f(z \,|\, \hat{\theta}_0)}{1 - F(a \,|\, \hat{\theta}_0)} \, dz + \sum_{i=1}^{n_1} \frac{\frac{\partial}{\partial \theta} f(x_i \,|\, \theta)}{f(x_i \,|\, \theta)},$$

en donde, como $f(x \,|\, \theta)$ es la función de densidad $\mathrm{N}(\theta, 1)$, se cumple que

$(\partial/\partial\theta) f(x \,|\, \theta) = f(x \,|\, \theta) \cdot (x - \theta)$. *Por lo tanto,*

$$\begin{aligned}
\frac{\partial}{\partial\theta} Q(\theta \,|\, \hat{\theta}_0, \underline{x}) &= n_2 \int_a^\infty (z-\theta) \cdot \frac{f(z \,|\, \hat{\theta}_0)}{1 - F(a \,|\, \hat{\theta}_0)}\, dz + \sum_{i=1}^{n_1} (x_i - \theta) \\
&= n_2 \int_a^\infty (z-\hat{\theta}_0) \cdot \frac{f(z \,|\, \hat{\theta}_0)}{1 - F(a \,|\, \hat{\theta}_0)}\, dz - n_2(\theta - \hat{\theta}_0) + n_1(\bar{x} - \theta) \\
&= \frac{n_2}{1 - F(a \,|\, \hat{\theta}_0)} \int_a^\infty (z-\hat{\theta}_0) \cdot \frac{1}{\sqrt{2\pi}}\, e^{-(z-\hat{\theta}_0)^2/2}\, dz \\
&\qquad\qquad -n_2(\theta - \hat{\theta}_0) + n_1(\bar{x} - \theta) \\
&= \frac{n_2}{1 - F(a \,|\, \hat{\theta}_0)} \int_a^\infty -\frac{d}{dz}\left[\frac{1}{\sqrt{2\pi}}\, e^{-(z-\hat{\theta}_0)^2/2} \right] dz \\
&\qquad\qquad -n_2(\theta - \hat{\theta}_0) + n_1(\bar{x} - \theta) \\
&= \frac{n_2}{1 - F(a \,|\, \hat{\theta}_0)}\, f(a \,|\, \hat{\theta}_0) - n_2(\theta - \hat{\theta}_0) + n_1(\bar{x} - \theta).
\end{aligned}$$

Esta es una función lineal en θ. *Igualando a cero y resolviendo para* θ *se encuentra que el punto en donde la función* Q *se maximiza es*

$$\hat{\theta}_1 = \frac{n_2}{n} \cdot \frac{f(a \,|\, \hat{\theta}_0)}{1 - F(a \,|\, \hat{\theta}_0)} + \frac{n_1}{n}\, \bar{x} + \frac{n_2}{n}\, \hat{\theta}_0,$$

en donde $n = n_1 + n_2$. *Esta expresión es similar a la del ejemplo anterior en donde los datos incompletos consistían en datos faltantes, ahora tenemos, además, el primer sumando del lado derecho. En general, se tiene la relación*

$$\hat{\theta}_{m+1} = \frac{n_2}{n} \cdot \frac{f(a \,|\, \hat{\theta}_m)}{1 - F(a \,|\, \hat{\theta}_m)} + \frac{n_1}{n}\, \bar{x} + \frac{n_2}{n}\, \hat{\theta}_m, \quad \textit{para} \;\; m = 0, 1, \ldots$$

Cuando $n_2 = 0$ *y* $n_1 = n$, *es decir, cuando todos los datos son observados y no hay censura, se obtiene la estimación* $\hat{\theta} = \bar{x}$. *En el caso general, haciendo* $m \to \infty$ *y suponiendo convergencia, se encuentra la ecuación*

$$\hat{\theta} = \frac{n_2}{n_1} \cdot \frac{f(a \,|\, \hat{\theta})}{1 - F(a \,|\, \hat{\theta})} + \bar{x},$$

en donde la incógnita es $\hat{\theta}$. *No parece fácil resolver esta ecuación, aunque numéricamente se puede tener una aproximación mediante la estimación*

secuencial.

Haciendo $a \to -\infty$, la condición $(z_j > a)$ de los datos censurados se transforma en $(z_j > -\infty)$, lo que convierte a esos datos en datos faltantes, y la relación iterativa se reduce a $\hat{\theta}_{m+1} = (n_1/n)\,\bar{x} + (n_2/n)\,\hat{\theta}_m$. Esta relación fue encontrada antes en el ejemplo en donde los datos incompletos eran faltantes. La estimación límite cuando $m \to \infty$ es $\hat{\theta} = \bar{x}$. ▪

El siguiente resultado muestra que la sucesión de estimadores $\hat{\theta}_0, \hat{\theta}_0, \ldots$ es siempre creciente, aunque se necesitan hipótesis adicionales para garantizar su convergencia en probabilidad hacia el estimador máximo verosímil cuando $m \to \infty$.

Teorema 6.1 *Sea $\hat{\theta}_0, \hat{\theta}_1, \ldots$ la sucesión de estimaciones del algoritmo EM. Entonces*

$$L(\hat{\theta}_{m+1} \,|\, \underline{x}) \geqslant L(\hat{\theta}_m \,|\, \underline{x}).$$

Demostración. Por definición, el logaritmo del lado izquierdo de la desigualdad a demostrar es

$$\begin{aligned}
\ln L(\hat{\theta}_{m+1} \,|\, \underline{\mathrm{x}}) &= Q(\hat{\theta}_{m+1} \,|\, \hat{\theta}_m, \underline{\mathrm{x}}) - E\left[\ln f(\underline{\mathrm{z}} \,|\, \hat{\theta}_{m+1}, \underline{\mathrm{x}}) \;\middle|\; \hat{\theta}_m, \underline{\mathrm{x}} \right] \\
&\geqslant Q(\hat{\theta}_m \,|\, \hat{\theta}_m, \underline{\mathrm{x}}) - E\left[\ln f(\underline{\mathrm{z}} \,|\, \hat{\theta}_{m+1}, \underline{\mathrm{x}}) \;\middle|\; \hat{\theta}_m, \underline{\mathrm{x}} \right] \\
&= Q(\hat{\theta}_m \,|\, \hat{\theta}_m, \underline{\mathrm{x}}) - E\left[\ln f(\underline{\mathrm{z}} \,|\, \hat{\theta}_m, \underline{\mathrm{x}}) \;\middle|\; \hat{\theta}_m, \underline{\mathrm{x}} \right] \\
&\qquad\qquad -E\left[\ln \frac{f(\underline{\mathrm{z}} \,|\, \hat{\theta}_m, \underline{\mathrm{x}})}{f(\underline{\mathrm{z}} \,|\, \hat{\theta}_{m+1}, \underline{\mathrm{x}})} \middle| \hat{\theta}_m, \underline{\mathrm{x}} \right] \\
&\geqslant \ln L(\hat{\theta}_m \,|\, \underline{\mathrm{x}}),
\end{aligned}$$

en donde la primera desigualdad se obtiene del hecho de que $\hat{\theta}_{m+1}$ maximiza la función $\theta \mapsto Q(\theta \,|\, \hat{\theta}_m, \underline{\mathrm{x}})$, y la segunda desigualdad es una aplicación de

la desigualdad de Jensen para la función convexa $-\ln x$. En efecto,

$$\begin{aligned}
E\left[\ln\frac{f(\underline{z}\,|\,\hat{\theta}_m,\underline{x})}{f(\underline{z}\,|\,\hat{\theta}_{m+1},\underline{x})}\;\middle|\;\hat{\theta}_m,\underline{x}\right] &= E\left[-\ln\frac{f(\underline{z}\,|\,\hat{\theta}_{m+1},\underline{x})}{f(\underline{z}\,|\,\hat{\theta}_m,\underline{x})}\;\middle|\;\hat{\theta}_m,\underline{x}\right]\\
&\geqslant -\ln E\left[\frac{f(\underline{z}\,|\,\hat{\theta}_{m+1},\underline{x})}{f(\underline{z}\,|\,\hat{\theta}_m,\underline{x})}\;\middle|\;\hat{\theta}_m,\underline{x}\right]\\
&= -\ln\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{f(\underline{z}\,|\,\hat{\theta}_{m+1},\underline{x})}{f(\underline{z}\,|\,\hat{\theta}_m,\underline{x})}\cdot f(\underline{z}\,|\,\hat{\theta}_m,\underline{x})\,d\underline{z}\\
&= -\ln\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\underline{z}\,|\,\hat{\theta}_m,\underline{x})\,d\underline{z}\\
&= -\ln(1)\\
&= 0.
\end{aligned}$$

■

En el Ejercicio 125 aparece un modelo discreto aplicado al estudio de ciertos vínculos en genética y en el que se puede aplicar el algoritmo EM. Se trata de una distribución multinomial en donde aparece un parámetro desconocido θ. El modelo fue estudiado por C. R. Rao en el libro [45] y es uno de los primeros ejemplos en los que se muestra la aplicación del algoritmo EM. El mismo modelo aparece también en otras fuentes como el libro de G. J. McLachlan y T. Krishnan [40].

6.3. El algoritmo *data augmentation*

Consideraremos que el parámetro θ a estimar se puede modelar mediante una variable aleatoria con una distribución inicial dada. A esta distribución inicial se le refiere como distribución a priori, es decir, antes de incorporar la información de las observaciones $x = (x_1, \ldots, x_n)$. Por el teorema de Bayes, se cumplen las relaciones

$$f(\theta\,|\,x) = f(x\,|\,\theta)\cdot\frac{f(\theta)}{f(x)} \propto f(x\,|\,\theta)\,f(\theta),$$

en donde $f(\theta\,|\,x)$ es la distribución llamada a posteriori, es decir, después de incorporar la información de las observaciones, y $f(\theta)$ es la distribución

a priori. Además, por el teorema de probabilidad total,

$$f(x) = \int f(x \mid \theta)\, f(\theta)\, d\theta.$$

A diferencia del algoritmo EM en donde se busca estimar puntualmente el parámetro θ (considerado como un parámetro fijo desconocido), en el algoritmo *data augmentation* se encuentra una aproximación de la distribución a posteriori para θ (considerado como una variable aleatoria). El símbolo $\propto$ en el contexto bayesiano significa «proporcional a», y sirve para concentrarse en los factores de la densidad a posteriori que contienen la información relacionada a θ, omitiendo los factores constantes.

Como antes, el vector $x = (x_1, \ldots, x_{n_1})$ representa un conjunto de datos completamente especificados, mientras que $z = (z_1, \ldots, z_{n_2})$ representa un conjunto de datos faltantes. Al vector (x, z) para valores particulares de los datos faltantes z se le llama datos aumentados (*augmented data*). Por su escritura, se podría pensar que los datos x y z están separados pero esto no es así. Por ejemplo, z_1 y z_2 pueden no observarse y lo que es observable es $x_1 = z_1 + z_2$. Esta situación ocurre en el caso de la distribución multinomial de un ejemplo anterior.

En el algoritmo *data augmentation* se utilizan las siguientes funciones:

- $f(\theta \mid x, z)$. Esta es la distribución a posteriori para θ a partir del vector de datos aumentados (x, z). Se espera que esta función tenga una expresión simple y supondremos que es posible generar valores usando esta función de densidad.

- $f(z \mid x)$. Esta es la distribución predictiva para los datos faltantes. Supondremos que es posible producir valores usando esta función de densidad y a los valores generados $z^{(1)}, \ldots, z^{(m)}$ se les llama imputaciones. Estos valores representan substitutos aproximados de los datos faltantes.

- $f(\theta \mid x)$. Esta es la distribución a posteriori observada, es decir, únicamente a partir de la información observada. Esta es la distribución que nos interesa encontrar.

Por ejemplo, para el caso de la distribución multinomial del Ejercicio 125 en el caso de datos faltantes y en donde hay 5 categorías y probabilidades asociadas como se indica en el diagrama 6.6 se tiene que:

$$
\begin{aligned}
& f(\theta \,|\, x, z) \\
& = f(x, z \,|\, \theta) \cdot f(\theta)/f(x, z) \\
& = \frac{n!}{z_1!\, z_2!\, x_2!\, x_3!\, x_4!} \left(\frac{1}{2}\right)^{z_1} \left(\frac{\theta}{4}\right)^{z_2} \left(\frac{1-\theta}{4}\right)^{x_2} \left(\frac{1-\theta}{4}\right)^{x_3} \left(\frac{\theta}{4}\right)^{x_4} \cdot \frac{f(\theta)}{f(x, z)} \\
& \propto \theta^{z_2+x_4}(1-\theta)^{x_2+x_3} \cdot f(\theta),
\end{aligned}
$$

en donde $x_1 = z_1 + z_2$ y $n = x_1 + x_2 + x_3 + x_4$. Observemos que esta es la distribución beta $(z_2 + x_4 + 1, x_2 + x_3 + 1)$ cuando la distribución a priori $f(\theta)$ es unif $(0, 1)$. Por otro lado, se puede comprobar que $f(z \,|\, x)$ es tal que $z_1 \sim \text{bin}\,(x_1, 2/(\theta + 2))$, o equivalentemente, $z_1 \sim \text{bin}\,(x_1, \theta/(\theta + 2))$ pues $z_1 + z_2 = x_1$.

El algoritmo *data augmentation* está basado en las siguientes dos identidades.

Proposición 6.1

1. *La función de densidad a posteriori del parámetro θ se puede expresar como*
$$f(\theta \,|\, x) = \int_z f(\theta \,|\, x, z)\, f(z \,|\, x)\, dz. \tag{6.1}$$
2. *La función de densidad de z dada x se puede escribir como*
$$f(z \,|\, x) = \int_\theta f(z \,|\, \phi, x)\, f(\phi \,|\, x)\, d\phi. \tag{6.2}$$

Demostración. En ambos casos se puede simplificar la integral del lado derecho y llegar a la expresión del lado izquierdo.

1. Por definición de función de densidad condicional,

$$\begin{aligned}\int_z f(\theta \mid x, z)\, f(z \mid x)\, dz &= \int_z \frac{f(\theta, x, z)}{f(x, z)}\, \frac{f(z, x)}{f(x)}\, dz \\ &= \int_z \frac{f(\theta, x, z)}{f(x)}\, dz \\ &= \frac{f(\theta, x)}{f(x)} \\ &= f(\theta \mid x).\end{aligned}$$

2. Por definición de función de densidad condicional,

$$\begin{aligned}\int_\theta f(z \mid \phi, x)\, f(\phi \mid x)\, d\phi &= \int_\theta \frac{f(z, \phi, x)}{f(\phi, x)}\, \frac{f(\phi, x)}{f(x)}\, d\phi \\ &= \int_\theta \frac{f(z, \phi, x)}{f(x)}\, d\phi \\ &= \frac{f(x, z)}{f(x)} \\ &= f(z \mid x).\end{aligned}$$

■

Con estos resultados se puede establecer la siguiente ecuación integral para $f(\theta \mid x)$.

Proposición 6.2 *Se cumple la relación*

$$f(\theta \mid x) = \int_\theta k(\theta, \phi)\, f(\phi \mid x)\, d\phi, \tag{6.3}$$

en donde $k(\theta, \phi)$ es una función definida para cualesquiera dos valores θ y ϕ del espacio parametral y está dada por

$$k(\theta, \phi) = \int_z f(\theta \mid x, z)\, f(z \mid \phi, x)\, dz. \tag{6.4}$$

Demostración. Substituyendo (6.2) en (6.1) e intercambiando el orden de las integrales se tiene que

$$\begin{aligned} f(\theta \,|\, x) &= \int_z f(\theta \,|\, x, z) \left(\int_\theta f(z \,|\, \phi, x)\, f(\phi \,|\, x)\, d\phi \right) dz \\ &= \int_\theta f(\phi \,|\, x) \cdot \underbrace{\int_z f(\theta \,|\, x, z)\, f(z \,|\, \phi, x)\, dz}_{=k(\theta,\phi)}\, d\theta \\ &= \int_\theta k(\theta, \phi)\, f(\phi \,|\, x)\, d\phi. \end{aligned}$$

∎

Se puede ahora proponer el siguiente método iterativo para resolver (6.3): sea $f_0(\theta \,|\, x)$ una función de densidad inicial, se calcula una siguiente función de densidad a partir de (6.3) de la siguiente forma

$$f_1(\theta \,|\, x) = \int_\theta k(\theta, \phi)\, f_0(\phi \,|\, x)\, d\phi.$$

De esta manera se puede obtener una sucesión de funciones de densidad $f_0(\theta \,|\, x)$, $f_1(\theta \,|\, x)$, $f_2(\theta \,|\, x)$,... la cual se acerca a la función desconocida $f(\theta \,|\, x)$ cuando se pueda garantizar algún tipo de convergencia. Sin embargo, debe observarse que este procedimiento requiere conocer la función $k(\theta, \phi)$ y poder llevar a cabo la integral (6.4).

La función $k(\theta, \phi)$ dada por (6.4) es la esperanza de la función $z \mapsto f(\theta \,|\, x, z)$ respecto de la función de densidad predictiva $f(z \,|\, \phi, x)$. Si $z^{(1)}, \ldots, z^{(m)}$ es una muestra de esta función de densidad, entonces la aproximación Monte Carlo para $k(\theta, \phi)$ es

$$\hat{k}(\theta) \approx \frac{1}{m} \sum_{i=1}^{m} f(\theta \,|\, x, z^{(i)}),$$

en donde la variable (valor parametral) ϕ ya no aparece pero influyó en la generación de la muestra. Substituyendo esta aproximación para $k(\theta, \phi)$

en (6.3) se obtiene la estimación Monte Carlo

$$\begin{aligned} \hat{f}(\theta \,|\, x) &= \hat{k}(\theta) \underbrace{\int f(\phi \,|\, x)\, d\phi}_{=1} \\ &= \hat{k}(\theta) \\ &\approx \frac{1}{m} \sum_{i=1}^{m} f(\theta \,|\, x, z^{(i)}). \end{aligned}$$

6.4. Remuestreo

Sea θ una cantidad desconocida que se puede estimar a través de una estadística $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, en donde $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución $F(x)$. Esta cantidad puede ser, por ejemplo, la media, la varianza, la esperanza de una función de la variable aleatoria subyacente, un cuantil o el valor desconocido de un parámetro. En general, la función $F(x)$ es desconocida y puede ser difícil conocer las características numéricas y las propiedades del estimador $\hat{\theta}$. Tenemos las siguientes alternativas:

- Las características de $\hat{\theta}$ podrían encontrarse si se conoce la distribución de esta variable aleatoria o una distribución aproximada de ella, aunque no está garantizado que los cálculos sean fáciles de llevar a cabo.
- Un método alternativo para conocer alguna característica de $\hat{\theta}$ consiste en su estimación a través de la simulación de valores de esta variable aleatoria. Si se conoce la distribución de $\hat{\theta}$, tal vez sea posible usar algún método de simulación para obtener valores $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ de esta variable aleatoria y estimar la característica buscada.
- Cuando la distribución de $\hat{\theta}$ no se conoce, se tendrían que producir varias realizaciones de la muestra aleatoria y para cada una de ellas calcular un valor de $\hat{\theta}$. Se tendría así la colección de valores $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ y, a partir de ellos, estimar la característica de interés. Sin embargo, este método puede ser muy costoso pues para calcular cada valor de $\hat{\theta}$ se tiene que tomar una muestra de $F(x)$. En algunas ocasiones, tomar varias muestras puede ser imposible pues sólo se cuenta con una única muestra $x_1, \ldots, x_n$ de $F(x)$.

- Existen algunos mecanismos conocidos como métodos de remuestreo, y en las siguientes secciones revisaremos dos de ellos, el método *bootstrap* y el método *jackknife*.

Estimadores Monte Carlo

En cualquier de los casos mencionados antes, supongamos que se pueden generar m valores $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ del estimador $\hat{\theta}$. Las características numéricas de este estimador se pueden aproximar usando Monte Carlo de la siguiente manera:

- La esperanza de cualquier función h de $\hat{\theta}$ puede estimarse de la forma usual,

$$E(h(\hat{\theta})) \approx \frac{1}{m} \sum_{i=1}^{m} h(\hat{\theta}_i).$$

  En particular, la esperanza del estimador se puede estimar como la media muestral,

$$E\left(\hat{\theta}\right) \approx \bar{\theta} := \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i.$$

- La varianza se puede aproximar mediante la varianza muestral, en donde se usa la aproximación $\bar{\theta}$ definida antes,

$$\mathrm{Var}(\hat{\theta}) \approx s^2 := \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta})^2.$$

- El sesgo del estimador es

$$E\left(\hat{\theta}\right) - \theta \approx \bar{\theta} - \hat{\theta} = \left(\frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i \right) - \hat{\theta},$$

  en donde $\hat{\theta}$ representa el estimador usando la muestra original y ha reemplazado el parámetro desconocido θ.

- Usando la muestra original, este mismo estimador $\hat{\theta}$ se puede utilizar para aproximar el error cuadrático medio, el cual se calcula como

$$E(\hat{\theta} - \theta)^2 \approx \frac{1}{m} \sum_{i=1}^{m} (\hat{\theta}_i - \hat{\theta})^2.$$

- También se pueden encontrar intervalos de confianza aproximados de la siguiente manera: cuando el tamaño n de la muestra original es grande y $\hat{\theta}$ es insesgado para θ, de manera aproximada,

$$\frac{\hat{\theta}-\theta}{s} \sim \mathrm{N}(0,1).$$

De aquí se obtiene un intervalo de confianza para θ dado por

$$P\left(\hat{\theta} - z_{1-\alpha/2}\cdot s < \theta < \hat{\theta} + z_{1-\alpha/2}\cdot s\right) = 1-\alpha.$$

6.5. El método *bootstrap*

Este método de remuestreo fue propuesto por Bradley Efron [10] en 1979. Consiste en tomar muestras de la función de distribución empírica $F_n(x)$, construida a partir de una muestra particular $x_1,\ldots,x_n$ de $F(x)$. Se trata de una estimación no paramétrica en donde se reemplaza a $F(x)$ por $F_n(x)$. Recordemos que la función de distribución empírica se define como

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(x_i \leqslant x)}, \quad -\infty < x < \infty.$$

Cuando el tamaño de muestra n es grande, se espera que la función $F_n(x)$ se parezca a la función de distribución, posiblemente desconocida, $F(x)$. Este comportamiento está garantizado por el teorema de Glivenko-Cantelli, cuyo enunciado puede consultarse en el Apéndice A.

Definición 6.2 (El método *bootstrap*) *Sea $F_n(x)$ la función de distribución empírica de una realización $x_1,\ldots,x_n$ de una muestra aleatoria de una distribución $F(x)$, posiblemente desconocida. En el método bootstrap se propone tomar muestras de la distribución discreta $F_n(x)$ como aproximación de muestras de $F(x)$.*

Debe observarse que tomar muestras de la distribución empírica $F_n(x)$ es sencillo pues equivale a generar valores de la distribución uniforme sobre la colección de números: $x_1,\ldots,x_n$, en donde no deben omitirse las

repeticiones. De hecho, cuando un valor x_i aparece repetido varias veces, tendrá mayor probabilidad de ser seleccionado en el muestreo *bootstrap*. Además, las nuevas muestras son con reemplazo, es decir, un mismo valor de la muestra original puede seleccionarse varias veces. Véase el Ejercicio 126. Bajo estas condiciones, el total de muestras de tamaño n que se pueden tomar de la distribución $F_n(x)$ es, a lo sumo, n^n. Este número es realmente grande, por ejemplo, si la muestra original es de 100 elementos, entonces el total de muestras *bootstrap* puede ser $100^{100} = 10^{200}$, esto es, un "1" seguido por doscientos "0".

La b-ésima muestra *bootstrap* se denota por

$$x_1^{(b)}, x_2^{(b)}, \ldots, x_n^{(b)}, \quad b = 1, \ldots, B, \tag{6.5}$$

en donde el uso de las letras b como subíndice y B como el total de muestras provienen de la primera letra del término *bootstrap*. Si θ es una cantidad que se desea conocer y se utiliza a la estadística $\hat{\theta}$ como su estimador, entonces la estimación para cada muestra *bootstrap* dada en (6.5) es

$$\hat{\theta}^{(b)} := \hat{\theta}(x_1^{(b)}, x_2^{(b)}, \ldots, x_n^{(b)}), \quad b = 1, \ldots, B. \tag{6.6}$$

Tenemos así la serie de estimaciones $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(B)}$. Promediando estos valores se encuentra el estimador *bootstrap* para θ,

$$\hat{\theta}_{bootstrap} := \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}.$$

Como se explicó antes, a partir de la colección de estimaciones *bootstraps* $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(B)}$ dada en (6.6), se pueden estimar características de la distribución de $\hat{\theta}$.

Ejemplo 6.3 *Usando una computadora se obtuvo una muestra de* 1000 *números* $x_1 = 0.2654018, \ldots, x_{1000} = 0.7785599$ *de la distribución* unif$(0, 1)$. *Esta nube de puntos se muestra en la Figura 6.1.*

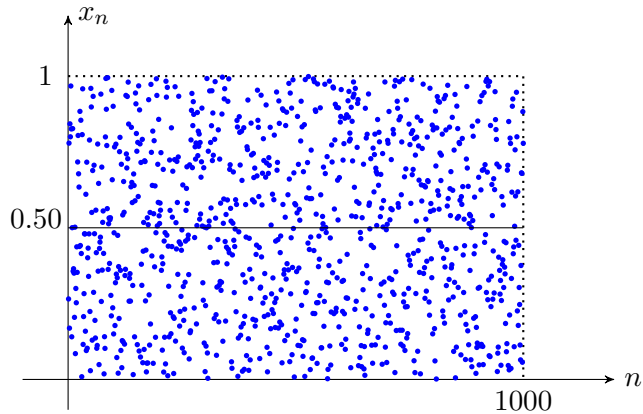


Figura 6.1: Observaciones de la distribución unif$(0, 1)$.

Obtenida la muestra, supondremos que desconocemos el origen de los datos y que deseamos utilizar el método bootstrap para estimar la media θ. La media muestral resultó ser $\hat{\theta} = 0.498636$. A continuación se tomaron 500 *muestras usando el método de remuestreo bootstrap y para cada una de ellas se calculó la media muestral $\hat{\theta}^{(b)}$, $b = 1, \ldots, 500$. Estos valores se muestran gráficamente en la Figura 6.2.*

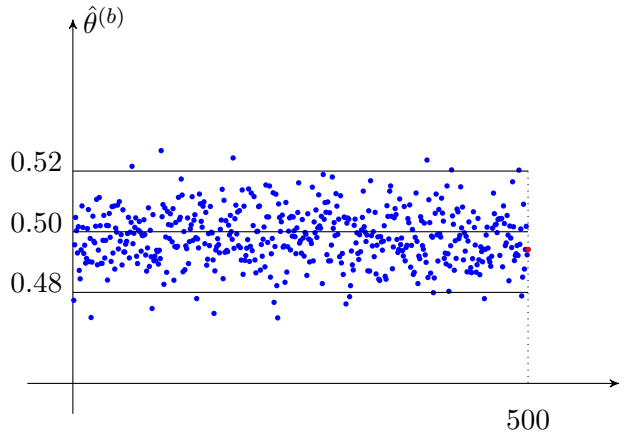


Figura 6.2: Medias de muestras *bootstrap*.

Observamos que las medias muestrales bootstrap $\hat{\theta}^{(b)}$ presentan cierta varia-

bilidad debido a que son calculadas usando las distintas muestras. La colección de valores $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(500)}$ *permiten conocer de manera aproximada la distribución del estimador* $\hat{\theta}$*. Se puede construir un histograma o la función de distribución empírica. En particular, se puede proponer un intervalo de confianza aproximado para* θ *como se indicó antes. La media muestral bootstrap es*

$$\hat{\theta}_{bootstrap} := \frac{1}{500} \sum_{b=1}^{500} \hat{\theta}^{(b)} = 0.4982077\,,$$

y la varianza muestral bootstrap es

$$s^2 := \frac{1}{499} \sum_{b=1}^{500} (\hat{\theta}^{(b)} - \hat{\theta}_{bootstrap})^2 = 0.00007906429\,.$$

■

Cuando se conoce, o se supone conocido, que $F(x)$ pertenece a una familia paramétrica, al método de la Definición 6.5 se le llama *bootstrap paramétrico*. Supondremos que θ es el parámetro y que es desconocido. En esta situación, se puede usar también el siguiente procedimiento para obtener muestras adicionales de $F(x)$. Primero se estima el parámetro θ a traves de los valores de una muestra aleatoria, es decir, se calcula $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$. Después se substituye el valor desconocido θ por la estimación $\hat{\theta}$ en la expresión de $F(x)$. Teniendo ahora completamente especificada la distribución $F(x)$, se pueden generar tantas muestras como se desee suponiendo que se conoce un método de simulación para ello. Cuando no se conoce a $F(x)$, al método de la Definición 6.5 se le llama *bootstrap no paramétrico*.

La palabra *bootstrap* proviene del inglés y se refiere a las cintas, generalmente hechas de cuero por su resistencia, que se colocan en la parte posterior de algunos zapatos y botas, con el fin de jalar de ellas para facilitar la introducción del pie en el calzado. Así, la cinta llamada *bootstrap* es un dispositivo de apoyo. En nuestro contexto estadístico, el término sugiere la idea de una herramienta de autosoporte, es decir, a partir de una muestra de $F(x)$ se pueden obtener muchas más muestras aproximadas de la misma distribución. Por otro lado, en las ciencias de la computación, al proceso de arranque del sistema operativo de una computadora se le llama *boot*, que es una versión corta de *bootstrap process*, entendido este término como el

inicio del arranque, arranque autosuficiente o arranque desde cero.

6.6. El método *jackknife*

Este es un procedimiento que precedió al método *bootstrap*. Se trata, nuevamente, de crear muestras a partir de una muestra dada $x_1, \ldots, x_n$ de una distribución $F(x)$, no necesariamente conocida. Como en el método *bootstrap*, a partir de estas muestras, llamadas ahora *muestras jackknife*, se puede analizar alguna característica asociada a la distribución $F(x)$. El método es más sencillo y se muestra en el siguiente recuadro.

Definición 6.3 (El método *jackknife*) *Sea $x_1, \ldots, x_n$ una realización de una muestra aleatoria de una distribución $F(x)$, posiblemente desconocida. El método jackknife propone crear n muestras a partir de la muestra original de la siguiente forma: para la muestra $j = 1, \ldots, n$, se omite el j-ésimo elemento de la muestra original, es decir, las muestras jackknife son:*

$$\begin{array}{ll} \textit{Muestra } 1\textit{:} & \blacktriangleright, x_2, x_3, \ldots, x_{n-1}, x_n \\ \textit{Muestra } 2\textit{:} & x_1, \blacktriangleright, x_3, \ldots, x_{n-1}, x_n \\ \textit{Muestra } 3\textit{:} & x_1, x_2, \blacktriangleright, \ldots, x_{n-1}, x_n \\ \vdots & \vdots \\ \textit{Muestra } n\textit{:} & x_1, x_2, x_3, \ldots, x_{n-1}, \blacktriangleright \end{array}$$

Claramente, las muestras *jackknife* son de tamaño $n-1$ y se tienen un total de n muestras. Para el análisis estadístico se pueden usar las n muestras *jackknife* y, en tal caso, el remuestreo es completo y no es necesario usar ningún algoritmo de simulación para generar las muestras, aunque se requiere cierto procedimiento para su creación en una computadora. Véase el Ejercicio 131.

La j-ésima muestra *jackknife* se denota por

$$(x_1^{(j)}, x_2^{(j)}, \ldots, x_{n-1}^{(j)}) := (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n), \quad j = 1, \ldots, n,$$

y una estimación de una cantidad desconocida θ basada en esta muestra se escribe de la siguiente manera

$$\hat{\theta}^{(j)} := \hat{\theta}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n), \quad j = 1, \ldots, n.$$

El uso de la letra j como índice en las expresiones anteriores proviene de la primera letra del término *jackknife*. Teniendo la sucesión de estimaciones $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(n)}$ se pueden estimar características de la distribución de $\hat{\theta}$. Debe observarse que si existen elementos repetidos en la muestra original, algunas de las muestras *jackknife* pueden ser idénticas. Por otro lado, dos muestras *jackknife* difieren muy poco, de modo que, en general, no se espera que exista una gran diferencia entre los valores $\hat{\theta}^{(j_1)}$ y $\hat{\theta}^{(j_2)}$ para $j_1 \neq j_2$.

Ejemplo 6.4 *Usando una computadora se obtuvo una muestra de* 100 *números* $x_1, \ldots, x_{100}$ *de una variable aleatoria* X *con una cierta distribución. Esta nube de puntos se muestra en la Figura 6.3.*

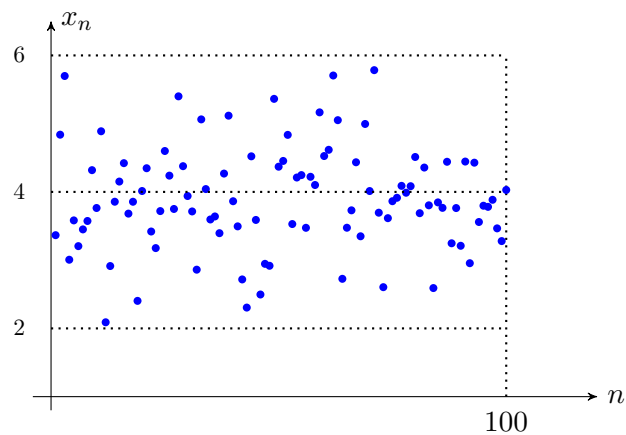


Figura 6.3: Observaciones de una cierta distribución.

Nos interesa estimar $\mathrm{Var}(X)$*, la cual denotaremos por* θ*. La varianza muestral resultó ser* $\hat{\theta} = 0.576919624432$ *. Se generaron las muestras jackknife y*

para cada una de ellas se calculó la varianza muestral $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(100)}$. *Estos valores se muestran gráficamente en la Figura 6.4.*

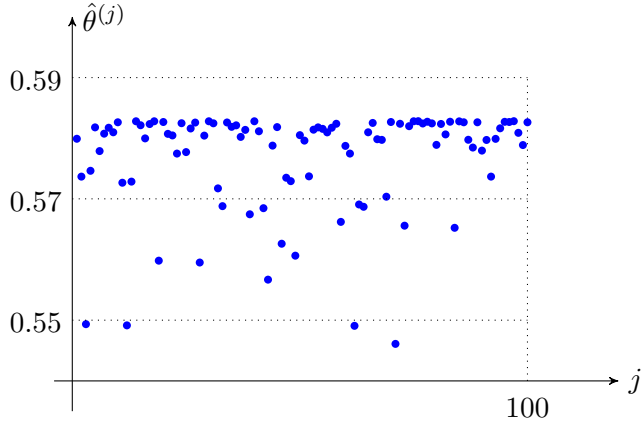


Figura 6.4: Varianzas de muestras *jackknife*.

Se puede construir un histograma o la función de distribución empírica para conocer una aproximación de la distribución del estimador $\hat{\theta}$. *En particular, se puede proponer un intervalo de confianza aproximado para* θ *como se indicó antes. El estimador jackknife para* θ *es*

$$\hat{\theta}_{jackknife} := \frac{1}{500} \sum_{j=1}^{100} \hat{\theta}^{(j)} = 0.576919624432\,,$$

y la varianza muestral jackknife es

$$s^2 := \frac{1}{99} \sum_{j=1}^{100} (\hat{\theta}^{(j)} - \hat{\theta}_{jackknife})^2 = 0.0000724068521268\,.$$

■

Puede usarse también sólo un subconjunto del total de muestras *jackknife*, y en ese caso, tal subconjunto deberá seleccionarse de manera equitativa entre el total de muestras.

Tanto el método *bootstrap* como el método *jackknife* se pueden estudiar con mayor detalle en los textos de A. C. Davison y D. V. Hinkley [8], B. Efron [11] y J. Shao y D. Tu [55].

6.7. Ejercicios

El algoritmo EM

125. *Vínculos en genética (Rao [45], McLachlan-Krishnan [40]).* Considere una distribución multinomial con cuatro categorías C_1, C_2, C_3, C_4, y probabilidades asociadas como se indica en la Figura 6.5, en donde $\theta \in [0,1]$ es un parámetro desconocido que se desea estimar.

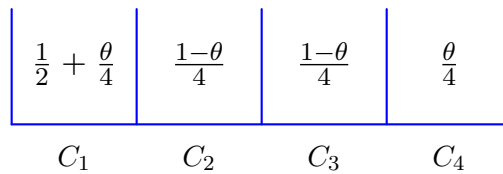


Figura 6.5: Cuatro categorías.

Suponga que en una muestra de tamaño n se obtienen las observaciones $x = (x_1, x_2, x_3, x_4)$ completamente especificadas y en donde x_i indica el número de veces que se observó la categoría C_i, $i = 1, 2, 3, 4$. Se debe cumplir que $x_1 + x_2 + x_3 + x_4 = n$.

a) Compruebe que la función de verosimilitud es

$$L(\theta \,|\, x) = \frac{n!}{x_1!\, x_2!\, x_3!\, x_4!} \left(\frac{1}{2} + \frac{\theta}{4}\right)^{x_1} \left(\frac{1-\theta}{4}\right)^{x_2} \left(\frac{1-\theta}{4}\right)^{x_3} \left(\frac{\theta}{4}\right)^{x_4}.$$

b) Demuestre que

$$\ln L(\theta \,|\, x) \propto c + x_1 \ln(2+\theta) + (x_2 + x_3) \ln(1-\theta) + x_4 \ln \theta,$$

en donde c es una constante.

c) Encuentre el estimador máximo verosímil para θ cuando los datos son $x = (125, 18, 20, 34)$.

Modificaremos ahora el problema creando variables que no son directamente observables. Suponga que la categoría C_1 se divide en dos

categorías $C_1^{(1)}$ y $C_1^{(2)}$ con probabilidades $1/2$ y $\theta/4$, respectivamente, como se muestra en la Figura 6.6 .

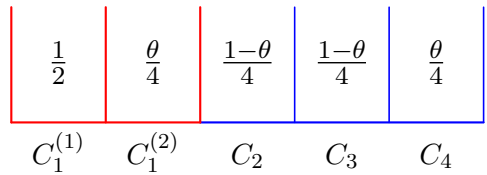


Figura 6.6: Cinco categorías.

Se tienen ahora 5 categorías cuyas frecuencias denotaremos por z_1, z_2, x_2, x_3, x_4, respectivamente. Supondremos que las dos categorías recién construidas no son observables directamente, sólo se observa $x_1 := z_1 + z_2$, sin conocer los valores de z_1 y z_2. El elemento $z = (z_1, z_2)$ representa un vector de datos faltantes, mientras que el vector $x = (x_1, x_2, x_3, x_4)$ es un vector de datos completamente observables. Notemos que los vectores de datos z y x no son objetos separados, el dato x_1 está definido en términos de z_1 y z_2. Como antes, el tamaño de la muestra es $n = x_1 + x_2 + x_3 + x_4$. El problema es nuevamente estimar θ bajo este esquema de datos faltantes.

d) Compruebe que la función de verosimilitud $L(\theta \,|\, x, z)$ es proporcional a una función de densidad beta.

e) Compruebe que la distribución $p(z \,|\, \theta, x)$ es $\text{bin}\,(x_1, 2/(2+\theta))$. Puesto que $z_1 + z_2 = x_1$, la distribución condicional de $z = (z_1, z_2)$ es la distribución univariada indicada. Esta es la distribución asociada al valor z_1. De manera equivalente, la distribución asociada a z_2 es $\text{bin}\,(x_1, \theta/(2+\theta))$.

f) *Paso E.* Dada una estimación inicial $\hat{\theta}_0$ para θ, encuentre la función Q dada por

$$Q(\theta \,|\, \hat{\theta}_0, x) := E(\,\ln L(\theta \,|\, x, z) \,|\, \hat{\theta}_0, x\,).$$

g) *Paso M.* Compruebe que la función Q se maximiza cuando θ toma

el valor

$$\hat{\theta}_1 = \frac{x_1\,\hat{\theta}_0 + 2x_4 + x_4\,\hat{\theta}_0}{n\,\hat{\theta}_0 + 2(x_2 + x_3 + x_4)}.$$

h) Escriba un programa de cómputo que muestre los primeros 20 valores $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{20}$ del algoritmo EM para estimar θ, en donde $\hat{\theta}_0$ es un valor inicial a su elección y los datos son como antes: $(z_1 + z_2, x_2, x_3, x_4) = (125, 18, 20, 34)$, $x_1 = z_1 + z_2$ y $n = 197$.

El método *bootstrap*

126. *Implementación del método de remuestreo bootstrap.*
Considere que se tiene definido en una computadora un arreglo de n números $x_1, \ldots, x_n$. Elabore un programa de cómputo que produzca una lista de B muestras *bootstrap*, en donde B es cualquier entero tal que $1 \leqslant B \leqslant n^n$. Por ejemplo, en el paquete estadístico `R` se puede definir el vector `Datos` como la lista de los siguientes 6 valores:

$$\texttt{Datos} \texttt{<-}\ \texttt{c}(3.5, 8.3, 5.1, 4.2, 1.8, 1.8)$$

Se puede usar el siguiente comando para generar una muestra *bootstrap* del vector `Datos` de tamaño n, con $1 \leqslant n \leqslant 6$. Si se omite la opción `size`, la muestra generada es de tamaño completo, en este caso de tamaño 6.

$$\texttt{sample}(\texttt{Datos}, \texttt{size} = \texttt{n}, \texttt{replace} = \texttt{TRUE})$$

Se repite el comando anterior tantas veces como se necesite para producir el número de muestras *bootstrap* requeridas.

127. Considere que se tiene una muestra de n números distintos $x_1, \ldots, x_n$.

a) ¿Cuántas muestras *bootstrap* se pueden construir exactamente?

b) ¿Cuál es la probabilidad de obtener una de estas muestras *bootstrap*?

128. Considere que se tienen las siguientes $n = 20$ observaciones de una variable aleatoria X con distribución desconocida.

4.2	3.8	2.1	1.7	2.3	3.3	8.6	6.3	5.8	9.5
5.2	4.7	8.9	5.1	6.0	1.6	4.5	2.2	1.7	5.2

Suponga que se desea estimar el coeficiente de variación (CV) definido de la forma siguiente

$$\theta := \mathrm{CV}(X) = \frac{\sqrt{\mathrm{Var}(X)}}{E(X)}.$$

a) Encuentre la estimación puntual $\hat{\theta}$ a partir de la muestra observada, es decir, calcule

$$\hat{\theta} = \frac{\mathrm{Var}(x)}{\bar{x}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\frac{1}{n}\sum_{i=1}^{n} x_i}.$$

b) Genere $B = 1000$ muestras *bootstrap* de la muestra original y para cada una de ellas calcule el coeficiente de variación estimado $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.

c) Elabore un histograma con los valores $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$. Esta gráfica nos proporciona una idea de la dispersión del coeficiente de variación considerando las distintas muestras *bootstrap*.

d) Calcule la media y la varianza del conjunto de valores $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$. En particular, la media es el estimador *bootstrap* para θ, es decir,

$$\hat{\theta}_{bootstrap} := \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{(b)}.$$

e) Bajo la hipótesis de normalidad, encuentre un intervalo de confianza aproximado al 95 % para θ.

129. Usando el algoritmo implementado en una computadora, genere 100 valores $x_1, \ldots, x_{100}$ de la distribución $\mathrm{N}(\mu, 1)$ con $\mu = 0$. Obtenida la muestra, suponga que olvidamos que $\mu = 0$ y que deseamos estimar este parámetro usando el método *bootstrap*.

a) Genere $B = 1000$ muestras *bootstrap* de tamaño 100 de la muestra original. Para cada una de las muestras *bootstrap* generadas $x_1^{(b)}, \ldots, x_{100}^{(b)}$, $b = 1, \ldots, B$, calcule la media muestral

$$\hat{\mu}^{(b)} = \frac{1}{100}\sum_{i=1}^{100} x_i^{(b)}, \quad b = 1, 2, \ldots, B.$$

b) En una gráfica muestre la nube de puntos $(b, \hat{\mu}^{(b)})$, $b = 1, \ldots, B$.

c) Elabore un histograma con los valores $\hat{\mu}^{(b)}$, $b = 1, \ldots, B$.

d) Calcule la media muestral *bootstrap* dada por

$$\hat{\mu}_{\,bootstrap} := \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}^{(b)}.$$

e) Encuentre un intervalo de confianza al 95 % para μ.

130. Usando el algoritmo implementado en una computadora, genere 100 valores $x_1, \ldots, x_{100}$ de la distribución unif$(0, 1)$.

a) Genere 500 muestras *bootstrap*.

b) Calcule la varianza *bootstrap* y el error estándar *bootstrap*.

c) Encuentre un intervalo de confianza al 95 % para la varianza.

El método *jackknife*

131. *Implementación del método de remuestreo jackknife.*
Considere que se tiene definido en una computadora un arreglo de n números $x_1, \ldots, x_n$. Elabore un programa de cómputo que produzca una lista de k muestras *jackknife* al azar, en donde k es cualquier entero tal que $1 \leqslant k \leqslant n$. Por ejemplo, supongamos que en el vector `Datos` tenemos la lista de los $n = 6$ valores que aparecen abajo. Esto se puede declarar en el paquete estadístico `R` de la siguiente manera:

$$\texttt{Datos} \texttt{ <- c}(3.5, 8.3, 5.1, 4.2, 1.8, 1.8)$$

La i-ésima muestra *jackknife* puede obtenerse de la forma que se muestra abajo, en donde `jack` es el arreglo que contiene esta muestra y es declarado al inicio. El valor de i puede ser $1, \ldots, n$, con n el tamaño de la muestra.

```
jack <- numeric(length(Datos)-1)
for (j in 1:length(Datos)){
    if (j < i) jack[j] <- Datos[j]
    else if (j > i) jack[j-1] <- Datos[j]
}
```

Compruebe que este código produce realmente la i-ésima muestra *jackknife*. Se puede completar la programación para que se genere la totalidad de muestras *jackknife* variando el valor de i. Si sólo se requieren k de ellas, se deben elegir de manera uniforme sobre todas las muestras *jackknife*.

132. Considere que se tiene una muestra de n números distintos.

   a) ¿Cuántas muestras *jackknife* se pueden construir exactamente?

   b) ¿Cuál es la probabilidad de obtener una de estas muestras *jackknife*?

133. Genere 500 valores $x_1, \ldots, x_{500}$ de la distribución Poisson(θ) con $\theta = 2$. Obtenida la muestra, olvidemos que $\theta = 2$ y supongamos que deseamos estimar este parámetro.

   a) Para cada una de las 500 muestras *jackknife* $x_1^{(j)}, \ldots, x_{499}^{(j)}$, calcule la media muestral

$$\hat{\theta}^{(j)} := \frac{1}{499} \sum_{i=1}^{499} x_i^{(j)}, \quad j = 1, \ldots, 500.$$

   b) En una gráfica muestre la nube de puntos $(j, \hat{\theta}^{(j)})$, $j = 1, \ldots, 500$.

   c) Elabore un histograma con los valores $\hat{\theta}^{(j)}$, $j = 1, \ldots, 500$.

   d) Calcule la media muestral *jackknife* dada por

$$\hat{\theta}_{jackknife} := \frac{1}{500} \sum_{j=1}^{500} \hat{\theta}^{(j)}.$$

   e) Encuentre un intervalo de confianza al 95 % para θ.

134. Genere 200 valores $x_1, \ldots, x_{200}$ de una variable aleatoria Z con distribución N(0, 1).

   a) Para cada una de las 200 muestras *jackknife* $x_1^{(j)}, \ldots, x_{199}^{(j)}$, $j = 1, \ldots, 200$, calcule $\hat{\theta}^{(j)}$ como la estimación de $\theta := P(Z > 1)$ dada por el cociente del número de valores de la muestra que son mayores a 1 entre el total del valores de la muestra, es decir, 199.

b) En una gráfica muestre la nube de puntos $(j, \hat{\theta}^{(j)})$, $j = 1, \ldots, 200$.

c) Elabore un histograma con los valores $\hat{\theta}^{(j)}$, $j = 1, \ldots, 200$.

d) Calcule la media muestral *jackknife* dada por

$$\hat{\theta}_{jackknife} := \frac{1}{200} \sum_{j=1}^{200} \hat{\theta}^{(j)}.$$

e) Encuentre un intervalo de confianza al 95 % para θ.

135. Considere que se tienen las siguientes $n = 24$ observaciones de una variable aleatoria X con distribución desconocida.

$$\begin{array}{cccccccccccc} -29 & -23 & -24 & -18 & 2 & 26 & 10 & 8 & -17 & -3 & 28 & -29 \\ -7 & -20 & 31 & 11 & -7 & -25 & 6 & 17 & 43 & -22 & -18 & 31 \end{array}$$

Suponga que se desea estimar $\theta = E(X)$.

a) Encuentre la estimación puntual $\hat{\theta} = \bar{x}$ a partir de la muestra observada.

b) Genere las muestras *jackknife* y para cada una de ellas calcule las medias muestrales $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(24)}$.

c) Elabore un histograma con los valores $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(24)}$.

d) Calcule la media y la varianza del conjunto de valores $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(24)}$. En particular, la media es el estimador *jackknife* para θ, es decir,

$$\hat{\theta}_{jackknife} := \frac{1}{24} \sum_{j=1}^{24} \hat{\theta}^{(j)}.$$

e) Bajo la hipótesis de normalidad, encuentre un intervalo de confianza aproximado al 95 % para θ.

Apéndice A: Resultados varios

En esta sección se presentan algunos conceptos y resultados que se han mencionado en el texto y que pueden ser de utilidad para el lector.

Desigualdad de Chebyshev

Este resultado proporciona una cota superior para la probabilidad de que una variable aleatoria tome un valor alejado de su valor medio μ en más de una cierta cantidad ϵ.

Proposición A.3 *Sea X una variable aleatoria con media μ y varianza finita σ^2. Para cualquier $\epsilon > 0$,*

$$P\left(|X-\mu| \geqslant \epsilon\right) \leqslant \frac{\sigma^2}{\epsilon^2}. \tag{7}$$

Demostración.

$$\begin{aligned}
\sigma^2 &= E\left[(X-\mu)^2\right] \\
&= E\left[\ (X-\mu)^2 \cdot 1_{(|X-\mu|\geqslant\epsilon)} + (X-\mu)^2 \cdot 1_{(|X-\mu|<\epsilon)}\right] \\
&\geqslant E\left[\ (X-\mu)^2 \cdot 1_{(|X-\mu|\geqslant\epsilon)}\right] \\
&\geqslant E\left[\ \epsilon^2 \cdot 1_{(|X-\mu|\geqslant\epsilon)}\right] \\
&= \epsilon^2 \cdot P\left(|X-\mu| \geqslant \epsilon\right).
\end{aligned}$$

∎

La desigualdad de Chebyshev es un resultado importante de la teoría de la probabilidad. Puede encontrarse mayor información de ésta y otras desigualdades en la mayoría de los textos de probabilidad, por ejemplo, puede consultarse el libro de A. Gut [18].

Ley de los grandes números

Este resultado es importante y establece que, bajo ciertas condiciones, el promedio de variables aleatorias converge a una constante cuando el número de sumandos crece a infinito.

Teorema A.2 *(Ley débil) Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes idénticamente distribuidas con media finita μ. Cuando $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

Demostración. Supondremos adicionalmente la existencia de la varianza, la cual denotaremor por σ^2. Sea $S_n = (X_1 + \cdots + X_n)/n$. Entonces $E(S_n) = \mu$ y $\text{Var}(S_n) = \sigma^2/n$. La desigualdad de Chebyshev aplicada a la variable S_n asegura que, para cualquier $\epsilon > 0$,

$$P\left(|S_n - \mu| \geqslant \epsilon\right) \leqslant \frac{\sigma^2}{n\,\epsilon^2}.$$

Tomando el límite cuando $n \to \infty$ se obtiene el resultado. ∎

Más explícitamente, la ley débil de los grandes números (convergencia en probabilidad) establece que

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \epsilon\right) = 0.$$

La ley débil establece la convergencia en probabilidad y la ley fuerte garantiza la convergencia casi segura. En consecuencia, la ley fuerte implica la ley débil. Más explícitamente, la ley fuerte de los grandes números (convergencia casi segura) establece que

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

Una demostración de la ley fuerte se puede encontrar en A. Gut [18] ó A. F. Karr [32].

Funciones convexas

Definición A.4 *Una función $\varphi(x)$ es convexa si para cualquier $\lambda \in [0,1]$,*

$$\varphi(\lambda x + (1-\lambda)y) \leqslant \lambda\varphi(x) + (1-\lambda)\varphi(y). \tag{8}$$

Esto significa que la línea recta que une los puntos $(x, \varphi(x))$ y $(y, \varphi(y))$ (lado derecho de la desigualdad) siempre queda por encima de los valores de la función en el intervalo $[a, b]$ (lado izquierdo de la desigualdad). Observe que no se ha especificado el dominio de la función $\varphi(x)$, sin embargo, de manera implícita, estamos suponiendo que si x y y son dos puntos en el dominio, entonces la combinación lineal $\lambda x + (1-\lambda)y$ también pertenece al dominio. La convexidad estricta se cumple cuando la desigualdad (8) se satisface de manera estricta para valores de λ en el intervalo abierto $(0, 1)$, es decir,

$$\varphi(\lambda x + (1-\lambda)y) < \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

Por ejemplo, la función exponencial $\varphi(x) = e^x$ es estrictamente convexa, su dominio es $(-\infty, \infty)$ y sólo toma valores positivos. La siguiente es una condición suficiente para la convexidad.

Proposición A.4 *Sea $\varphi(x)$ una función dos veces diferenciable y tal que $\varphi''(x) \geqslant 0$. Entonces $\varphi(x)$ es convexa.*

Funciones cóncavas

Definición A.5 *Una función $\varphi(x)$ es cóncava cuando $-\varphi(x)$ es convexa, es decir, si para cualquier $\lambda \in [0,1]$,*

$$\varphi(\lambda x + (1-\lambda)y) \geqslant \lambda\varphi(x) + (1-\lambda)\varphi(y). \tag{9}$$

En este caso, la línea recta que une los puntos $(x, \varphi(x))$ y $(y, \varphi(y))$ (lado derecho de la desigualdad) queda por debajo de los valores de la función en el intervalo $[a, b]$ (lado izquierdo de la desigualdad). La concavidad estricta se cumple cuando la desigualdad (9) es estricta. Por ejemplo, la función $\varphi(x) = \ln x$ es estrictamente cóncava. Esta función tiene como dominio de definición el intervalo $(0, \infty)$, y sus valores son tanto positivos como negativos. La siguiente es una condición suficiente para la concavidad.

Proposición A.5 *Sea $\varphi(x)$ una función dos veces diferenciable y tal que $\varphi''(x) \leqslant 0$. Entonces $\varphi(x)$ es cóncava.*

Funciones de distribución

Definición A.6 *Una función $F(x) : \mathbb{R} \to [0,1]$ es de distribución si satisface las siguientes propiedades:*

1. $\lim_{x\to+\infty} F(x) = 1$.
2. $\lim_{x\to-\infty} F(x) = 0$.
3. *Si $x_1 \leqslant x_2$ entonces $F(x_1) \leqslant F(x_2)$.*
4. *$F(x)$ es continua por la derecha, es decir, $F(x+) = F(x)$.*

Definición A.7 *Una función $F(x,y) : \mathbb{R} \times \mathbb{R} \to [0,1]$ es de distribución (bivariada) si satisface las siguientes propiedades:*

1. $\lim_{x,y\to\infty} F(x,y) = 1$.
2. $\lim_{x\to-\infty} F(x,y) = \lim_{y\to-\infty} F(x,y) = 0$.
3. *$F(x,y)$ es no decreciente en cada variable.*
4. *$F(x,y)$ es continua por la derecha en cada variable.*
5. *Si $a_1 < b_1$ y $a_2 < b_2$, entonces*

$$F(b_1,b_2) - F(a_1,b_2) - F(b_1,a_2) + F(a_1,a_2) \geqslant 0.$$

Desigualdad de Jensen

Si X es una variable aleatoria con esperanza finita y $\varphi(x)$ es una función convexa, entonces

$$\varphi(E(X)) \leqslant E(\varphi(X)).$$

A este resultado se le llama desigualdad de Jensen. Cuando $\varphi(x)$ es cóncava, la desigualdad se invierte. Por ejemplo, la función $\varphi(x) = x^2$ es convexa y la desigualdad de Jensen implica que $E^2(X) \leqslant E(X^2)$, suponiendo finitud de estas esperanzas. Esto corrobora que $\text{Var}(X) \geqslant 0$. La desigualdad de

Jensen también es válida para esperanzas condicionales, es decir, para $\varphi(x)$ convexa,

$$\varphi(E(X \,|\, Y)) \leqslant E(\varphi(X) \,|\, Y).$$

Observe, sin embargo, que los dos términos que aparecen en esta última desigualdad son variables aleatorias y la afirmación se cumple en el sentido c.s. La demostración de la desigualdad de Jensen puede encontrarse en el libro de A. Gut [18].

Teorema de Glivenko-Cantelli

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución $F(x)$. La función de distribución empírica es la variable aleatoria

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(X_i \leqslant x)}, \quad -\infty < x < \infty.$$

Para cada x fijo, la ley fuerte de los grandes números implica que la variable aleatoria $F_n(x)$ converge a la constante $F(x)$ en el sentido casi seguro. Es decir, puntualmente en x, $F_n(x) \to F(x)$, c.s. El teorema de Glivenko-Cantelli asegura que la convergencia es uniforme en x pues establece que, cuando $n \to \infty$ y en el sentido casi seguro,

$$\sup_x |F_n(x) - F(x)| \to 0.$$

La demostración de este resultado puede consultarse en textos de probabilidad como A. Gut [18] ó A. F. Karr [32], por ejemplo.

Función gama

La función gama se denota por $\Gamma(z)$ y se define como la integral que aparece abajo para aquellos valores complejos z tal que la integral es absolutamente convergente.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx, \quad z \in \mathbb{C}.$$

En nuestro caso, la función gama se utiliza cuando el argumento z es un número real positivo y, en tal situación, la función está bien definida. Esta

función aparece en la expresión de la función de densidad gama y está también relacionada con la función beta. Puede comprobarse directamente que $\Gamma(1/2) = \sqrt{\pi}$ y que $\Gamma(2) = \Gamma(1) = 1$. Más generalmente, puede demostrarse que:

- $\Gamma(z+1) = z\,\Gamma(z)$.
- $\Gamma(z+1) = z!$ para $z = 0, 1, 2, \ldots$

Observe que cualquiera de estas propiedades establece que la función gama es una generalización del factorial de un número natural. Para mayor información se puede consultar M. Abramowitz e I. A. Stegun [1], P. J. Davis [7] ó H. Hochstad [21].

Función beta

La función beta se denota por $\mathrm{B}(a, b)$ y se define como la integral que aparece abajo para valores $a > 0$ y $b > 0$,

$$\mathrm{B}(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx.$$

Puede comprobarse que se cumple la propiedad $\mathrm{B}(a, b) = B(b, a)$. Esta función aparece en la definición de la distribución beta y está relacionada con la función gama de la siguiente manera,

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

Como la función $\Gamma(z)$ está bien definida para valores $z > 0$, la función $\mathrm{B}(a, b)$ también está bien definida para $a > 0$ y $b > 0$. Se puede encontrar mayor información sobre la función beta en M. Abramowitz e I. A. Stegun [1] ó H. Hochstad [21].

Teoremas de cambio de variable

Los siguientes dos resultados proporcionan fórmulas para la función de densidad de la transformación de una variable aleatoria o de un vector aleatorio.

Proposición A.6 *Sea X una variable aleatoria con función de densidad $f_X(x)$ y con valores en el intervalo $(a,b) \subseteq \mathbb{R}$. Sea $\varphi : (a,b) \to \mathbb{R}$ una función continua, estrictamente creciente o decreciente, y con inversa diferenciable. Entonces la variable aleatoria $Y = \varphi(X)$ tiene función de densidad*

$$f_Y(y) = f_X(\varphi^{-1}(y)) \left| \frac{d}{dy}\varphi^{-1}(y) \right| \quad si \quad y \in \varphi(a,b).$$

Proposición A.7 *Sea (X,Y) un vector aleatorio con función de densidad $f_{X,Y}(x,y)$ y con valores en una región $A \subseteq \mathbb{R}^2$. Sea $\varphi : A \to \mathbb{R}^2$ una función continua y con inversa diferenciable $\varphi^{-1} = (\varphi_1^{-1}, \varphi_2^{-1})$. Entonces el vector aleatorio $(U,V) = \varphi(X,Y)$ tiene función de densidad*

$$f_{U,V}(u,v) = f_{X,Y}(\varphi^{-1}(u,v))\,|J(u,v)| \quad si \quad (u,v) \in \varphi(A),$$

en donde

$$J(u,v) = \begin{vmatrix} \partial_u \varphi_1^{-1} & \partial_v \varphi_1^{-1} \\ \partial_u \varphi_2^{-1} & \partial_v \varphi_2^{-1} \end{vmatrix}.$$

La demostración de estos resultados puede ser consultada en L. Rincón [47].

Función de densidad de la suma, diferencia, producto y cociente de variables aleatorias

Sea (X,Y) un vector aleatorio absolutamente continuo con función de densidad $f_{X,Y}(x,y)$. Entonces

1. $f_{X+Y}(u) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(u-v,v)\,dv.$
2. $f_{X-Y}(u) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(u+v,v)\,dv.$
3. $f_{XY}(u) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(u/v,v)\,|1/v|\,dv.$
4. $f_{X/Y}(u) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(uv,v)\,|v|\,dv.$

Un método para demostrar estas fórmulas hace uso del teorema de cambio de variable. Los detalles pueden encontrarse en L. Rincón [47].

Apéndice B: Distribuciones de probabilidad

En esta sección se muestra información básica acerca de algunas distribuciones de probabilidad univariadas y la forma en la que pueden generarse sus valores. En algunos casos, a partir de los valores de unas variables aleatorias se pueden construir valores de otras variables aleatorias. Los métodos de simulación indicados en esta sección son mayormente los que se han expuesto en este trabajo. Pueden existir otros métodos alternativos de simulación.

Por otro lado, debe advertirse que las distintas referencias bibliográficas pueden emplear una parametrización diferente a la aquí se presentada para las distribuciones. Esto incluye también a los programas de cómputo, en donde la implementación de una distribución de probabilidad puede tener una parametrización diferente. Se puede encontrar mayor información de las distintas distribuciones de probabilidad univariadas en los libros de Johnson, Kemp y Kotz [26] y Jonhson, Kotz y Balakrishnan [27] y [28].

Bernoulli

Una variable aleatoria X que sólo toma los valores 0 y 1 tiene una distribución Bernoulli y se escribe $X \sim \text{Ber}\,(p)$, en donde $p \in (0,1)$ es un parámetro. Su función de probabilidad es

$$f(x) = p^x(1-p)^{1-x}, \qquad x = 0, 1.$$

Su esperanza es $E(X) = p$ y su varianza es $\text{Var}(X) = p\,(1-p)$.

El método de la función inversa puede utilizarse para simular valores de esta distribución. El método se reduce al siguiente resultado: si u es un valor de la distribución unif $(0,1)$, entonces $x = 1_{(0,p]}(u)$ es un valor de la distribución Ber (p), véase el Ejemplo 2.6.

binomial

Una variable aleatoria X que toma los valores en $0, 1, \ldots, n$ tiene una distribución binomial y se escribe $X \sim \text{bin}\,(n,p)$, en donde $n \in \mathbb{N}$ y $p \in (0,1)$ son dos parámetros, cuando su función de probabilidad es

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Su esperanza es $E(X) = np$ y su varianza es $\text{Var}(X) = np\,(1-p)$. Cuando $n = 1$ se obtiene la distribución Ber (p).

Para simular valores de la distribución binomial se pueden usar los siguientes métodos:

- Se puede usar el método de la función inversa siguiendo el algoritmo de la Figura 2.5 que aparece en la Sección 2.2.
- Se puede aplicar el método de von Neumann para variables aleatorias discretas con soporte acotado que aparece en la Proposición 2.4.
- Un valor de la distribución binomial puede obtenerse a partir de n valores de la distribución Bernoulli usando el siguiente resultado: si $X_1, \ldots, X_n$ son variables aleatorias independientes con idéntica distribución Ber (p), entonces $X_1 + \cdots + X_n$ tiene distribución bin (n,p).

unif discreta

Una variable aleatoria X que toma n valores distintos $x_1, \ldots, x_n$ tiene una

distribución uniforme y se escribe $X \sim \text{unif}\,\{x_1, \ldots, x_n\}$ cuando su función de probabilidad es

$$f(x) = \frac{1}{n}, \qquad x = x_1, \ldots, x_n.$$

Su esperanza es $E(X) = \bar{x} = (x_1 + \cdots + x_n)/n$ y su varianza es $\text{Var}(X) = (1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Se puede usar el método de la función inversa para generar valores de la distribución uniforme discreta siguiendo el algoritmo de la Figura 2.5 que aparece en la Sección 2.2.

geométrica

Una variable aleatoria X que toma los valores en $0, 1, 2, \ldots$ tiene una distribución geométrica y se escribe $X \sim \text{geo}\,(p)$, en donde $p \in (0, 1)$ es un parámetro, cuando su función de probabilidad es

$$f(x) = p\,(1-p)^x, \qquad x = 0, 1, 2, \ldots$$

Su esperanza es $E(X) = (1-p)/p$ y su varianza es $\text{Var}(X) = (1-p)/p^2$.

Para simular valores de la distribución geométrica se pueden usar los siguientes métodos:

- Se puede aplicar el método de la función inversa siguiendo el algoritmo de la Figura 2.5 que aparece en la Sección 2.2.
- Se puede aprovechar la facilidad para obtener valores de la distribución exponencial para generar valores de la distribución geométrica como se muestra en el procedimiento del Ejercicio 54. El resultado es el siguiente: Si u es un valor de la distribución unif $(0, 1)$, entonces $\lfloor \ln u / \ln p \rfloor$ es un valor de la distribución geo (p), en donde $\lfloor x \rfloor$ indica la parte entera del número $x \geqslant 0$.

binomial negativa

Una variable aleatoria X que toma los valores $0, 1, 2, \ldots$ tiene una distribución binomial negativa y se escribe $X \sim \text{bin neg}(k, p)$, en donde $k \in \mathbb{N}$ y $p \in (0, 1)$ son dos parámetros, cuando su función de probabilidad es

$$f(x) = \binom{k+x-1}{x} p^k (1-p)^x, \qquad x = 0, 1, 2, \ldots$$

Su esperanza es $E(X) = k(1-p)/p$ y su varianza es $\text{Var}(X) = k(1-p)/p^2$. Cuando $k = 1$ se obtiene la distribución $\text{geo}\,(p)$.

Para simular valores de la distribución binomial negativa se pueden usar los siguientes métodos:

- Se puede aplicar el método de la función inversa siguiendo el algoritmo de la Figura 2.5 que aparece en la Sección 2.2.
- Un valor de la distribución binomial negativa puede obtenerse a partir de k valores de la distribución geométrica usando el siguiente resultado: si $X_1, \ldots, X_k$ son variables aleatorias independientes con idéntica distribución $\text{geo}\,(p)$, entonces $X_1 + \cdots + X_k$ tiene distribución $\text{bin neg}(k, p)$.

Poisson

Una variable aleatoria X que toma los valores $0, 1, 2, \ldots$ tiene una distribución Poisson y se escribe $X \sim \text{Poisson}\,(\lambda)$, en donde $\lambda > 0$ es un parámetro, cuando su función de probabilidad es

$$f(x) = \frac{\lambda^x}{x!}\, e^{-\lambda}, \qquad x = 0, 1, 2, \ldots$$

Su esperanza es $E(X) = \lambda$ y su varianza es $\text{Var}(X) = \lambda$.

Se puede usar el método de la función inversa para generar valores de la distribución Poisson siguiendo el algoritmo de la Figura 2.5 que aparece en la Sección 2.2.

hipergeométrica

Una variable aleatoria X que toma los valores $0, 1, \ldots, n$ tiene una distribución hipergeométrica y se escribe $X \sim \text{hipergeo}\,(N, K, n)$, en donde N, K y n son tres parámetros con valores naturales tales que $n \leqslant K$ y $n \leqslant N - K$, cuando su función de probabilidad es

$$f(x) = \binom{K}{x}\binom{N-K}{n-x}\Big/\binom{N}{n}, \qquad x = 0, 1, \ldots, n.$$

Su esperanza es $E(X) = n(K/N)$ y su varianza es $\text{Var}(X) = n(K/N)((N-K)/N)(N-n)/(N-1)$.

Para simular valores de la distribución hipergeométrica se pueden usar los siguientes métodos:

- Se puede aplicar el método de la función inversa siguiendo el algoritmo de la Figura 2.5 que aparece en la Sección 2.2.
- Se puede aplicar el método de von Neumann para variables aleatorias discretas con soporte acotado que aparece en la Proposición 2.4.

unif continua

Una variable aleatoria X que toma valores en el intervalo acotado (a, b) tiene una distribución uniforme y se escribe $X \sim \text{unif}\,(a, b)$ cuando su función de densidad es

$$f(x) = \frac{1}{b-a} \qquad a < x < b.$$

Su función de distribución es $F(x) = (x-a)/(b-a)$ para $a < x < b$. Su esperanza es $E(X) = (a+b)/2$ y su varianza es $\text{Var}(X) = (b-a)^2/12$.

Se puede usar el método de la función inversa para generar valores de esta distribución. El procedimiento es el siguiente: si u es un valor de la función de densidad $\text{unif}\,(0, 1)$, entonces $a + (b-a)u$ es valor de la función de densidad $\text{unif}\,(a, b)$.

exponencial

Una variable aleatoria X que toma valores en el intervalo $(0, \infty)$ tiene una distribución exponencial y se escribe $X \sim \exp(\lambda)$, en donde $\lambda > 0$ es un parámetro, cuando su función de densidad es

$$f(x) = \lambda\, e^{-\lambda x}, \qquad x > 0.$$

Su función de distribución es $F(x) = 1 - e^{-\lambda x}$, para $x > 0$. Su esperanza es $E(X) = 1/\lambda$ y su varianza es $\text{Var}(X) = 1/\lambda^2$.

Para simular valores de la distribución exponencial se pueden usar los siguientes métodos:

- Se puede aplicar el método de la función inversa. Esta es la forma más simple de simulación. La función de distribución $F(x)$ es invertible y el método asegura que si u es un valor de la distribución $\text{unif}(0, 1)$, entonces $x = (1/\lambda)\ln(1/u)$ es un valor de la distribución $\exp(\lambda)$.
- Se puede aplicar el método del cociente. Los detalles de este procedimiento se muestran en el Ejemplo 2.15.

normal

Una variable aleatoria X que toma valores en el intervalo $(-\infty, \infty)$ tiene una distribución normal y se escribe $X \sim \text{N}(\mu, \sigma^2)$, en donde $\mu \in \mathbb{R}$ y $\sigma^2 > 0$ son dos parámetros, cuando su función de densidad es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad -\infty < x < \infty.$$

A esta función se le denota por $\phi(x)$. A la correspondiente función de distribución se le denota por $\Phi(x)$. Su esperanza es $E(X) = \mu$ y su varianza es $\text{Var}(X) = \sigma^2$. Cuando $\mu = 0$ y $\sigma^2 = 1$ se obtiene la distribución normal estándar, denotada por $\text{N}(0, 1)$, cuya función de densidad es $\phi(x) = (2\pi)^{-1/2} \exp\{-x^2/2\}$, para $-\infty < x < \infty$. El siguiente resultado establece que es suficiente estudiar la distribución normal estándar:

a) $X \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow (X - \mu)/\sigma \sim \mathrm{N}(0, 1)$.

b) $X \sim \mathrm{N}(0, 1) \Rightarrow \mu + \sigma X \sim \mathrm{N}(\mu, \sigma^2)$.

Se pueden usar los siguientes métodos para generar valores de la distribución normal:

- Se puede aplicar el método de aceptación y rechazo como se muestra en el Ejemplo 2.11.
- Se puede aplicar el método de Box y Muller que se expone en la Sección 2.5.
- Se puede aplicar el método de Marsaglia que se expone en la Sección 2.6.
- Se puede aplicar el método del cociente de uniformes como se muestra en el Ejemplo 2.16.

log normal

Una variable aleatoria X que toma valores en el intervalo $(0, \infty)$ tiene una distribución log normal y se escribe $X \sim \log\text{normal}(\mu, \sigma^2)$, en donde $\mu \in \mathbb{R}$ y $\sigma^2 > 0$ son dos parámetros, cuando su función de densidad es

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, \qquad x > 0.$$

Esta es la distribución de una variable aleatoria definida como la exponencial de una variable con distribución $\mathrm{N}(\mu, \sigma^2)$. En símbolos,

$$Y \sim \mathrm{N}(\mu, \sigma^2) \;\Leftrightarrow\; e^Y \sim \log\text{normal}(\mu, \sigma^2).$$

Es decir, el logaritmo natural de la variable $X = e^Y$ tiene distribución normal. De este hecho surge el nombre de esta distribución. Su función de distribución se puede expresar como $F(x) = \Phi((\ln x - \mu)/\sigma)$. Su esperanza es $E(X) = \exp\left\{\mu + \sigma^2/2\right\}$ y su varianza es $\mathrm{Var}(X) = \exp\left\{2\mu + 2\sigma^2\right\} - \exp\left\{2\mu + \sigma^2\right\}$.

Para simular valores de la distribución log normal a partir de valores de la distribución normal se puede usar el resultado arriba inidcado: si $Y \sim \mathrm{N}(\mu, \sigma^2)$, entonces $e^Y \sim \log \mathrm{normal}(\mu, \sigma^2)$.

gama

Una variable aleatoria X que toma valores en el intervalo $(0, \infty)$ tiene una distribución gama y se escribe $X \sim \text{gama}\,(\gamma, \lambda)$, en donde $\gamma > 0$ y $\lambda > 0$ son dos parámetros, cuando su función de densidad es

$$f(x) = \frac{(\lambda x)^{\gamma-1}}{\Gamma(\gamma)}\,\lambda\, e^{-\lambda x}, \qquad x > 0,$$

en donde $\Gamma(\gamma) = \int_0^\infty t^{\gamma-1} e^{-t}\, dt$ es la función gama. Su esperanza es $E(X) = \gamma/\lambda$ y su varianza es $\mathrm{Var}(X) = \gamma/\lambda^2$. Cuando $\gamma = 1$ se obtiene la distribución $\exp(\lambda)$. Cuando γ es un número entero $n \geqslant 1$, la distribución gama se conoce también con el nombre de distribución Erlang y se denota por $\mathrm{Erlang}(n, \lambda)$

Para simular valores de la distribución gama se pueden usar los siguientes métodos:

- Se puede aplicar el método del cociente. Los detalles de este procedimiento se muestran en el Ejercicio 49 en la página 117.
- Cuando el parámetro γ es igual al número entero $n \geqslant 1$ (caso Erlang), un valor de la distribución $\mathrm{Erlang}(n, \lambda)$ puede obtenerse a partir de n valores de la distribución exponencial usando el siguiente resultado: si $X_1, \ldots, X_n$ son variables aleatorias independientes con idéntica distribución $\exp(\lambda)$, entonces $X_1 + \cdots + X_n$ tiene distribución $\mathrm{Erlang}(n, \lambda)$.

log gama

Debe advertirse al lector que existen varias maneras de definir una distribución con este nombre y las fórmulas que se obtienen no son necesariamente

iguales. Seguiremos el mismo procedimiento que se usó para definir a la distribución log normal, el cual es consistente con la definición de la distribución log gama que aparece en el texto de R. V. Hogg y S. A. Klugman [22].

Una variable aleatoria X que toma valores en el intervalo $(1, \infty)$ tiene una distribución log gama y se escribe $X \sim \log\text{gama}\,(\gamma, \lambda)$, en donde $\gamma > 0$ y $\lambda > 0$ son dos parámetros, cuando su función de densidad es

$$f(x) = \frac{(\lambda \ln x)^{\gamma-1}}{\Gamma(\gamma)} \lambda\, x^{-(\lambda+1)}, \qquad x > 1.$$

Esta es la distribución de una variable aleatoria definida como la exponencial de una variable con distribución gama (γ, λ). En símbolos,

$$Y \sim \text{gama}\,(\gamma, \lambda) \;\Leftrightarrow\; e^Y \sim \log\text{gama}\,(\gamma, \lambda).$$

Es decir, el logaritmo natural de la variable $X = e^Y$ tiene distribución gama. De este hecho surge el nombre de esta distribución. Su esperanza es $E(X) = [\lambda/(\lambda - 1)]^\gamma$ para $\lambda > 1$ y su varianza es $\text{Var}(X) = [\lambda/(\lambda - 2)]^\gamma - [\lambda/(\lambda - 1)]^{2\gamma}$ para $\lambda > 2$.

Para simular valores de la distribución log gama a partir de valores de la distribución gama se puede usar el resultado arriba indicado: si $Y \sim \text{gama}\,(\gamma, \lambda)$, entonces $e^Y \sim \log\text{gama}\,(\gamma, \lambda)$.

χ^2

Una variable aleatoria X que toma valores en el intervalo $(0, \infty)$ tiene una distribución ji-cuadrada y se escribe $X \sim \chi^2(n)$, en donde $n \geqslant 1$ es un parámetro con valor en los números naturales y llamado *grados de libertad*, cuando su función de densidad es

$$f(x) = \frac{(1/2)^{n/2}}{\Gamma(n/2)}\, x^{n/2-1}\, e^{-x/2}, \qquad x > 0.$$

Su esperanza es $E(X) = n$ y su varianza es $\text{Var}(X) = 2n$. La distribución $\chi^2(n)$ es la distribución gama (γ, λ) cuando $\gamma = n/2$ y $\lambda = 1/2$.

Para simular valores de la distribución ji-cuadrada se pueden usar los siguientes métodos:

- Se puede usar el siguiente resultado: si $Z_1, \ldots, Z_n$ son variables aleatorias independientes con distribución $\mathrm{N}(0,1)$, entonces $Z_1^2 + \cdots + Z_n^2$ tiene distribución $\chi^2(n)$.
- Se puede usar el método que se conozca para la distribución gama (γ, λ) con los parámetros particulares $\gamma = n/2$ y $\lambda = 1/2$.

Weibull

Una variable aleatoria X que toma valores en el intervalo $(0, \infty)$ tiene una distribución Weibull y se escribe $X \sim \text{Weibull}\,(r, \lambda)$, en donde $r > 0$ y $\lambda > 0$ son dos parámetros, cuando su función de densidad es

$$f(x) = \lambda r(\lambda x)^{r-1} \exp\left\{-(\lambda x)^r\right\}, \qquad x > 0.$$

Su función de distribución es $F(x) = 1 - e^{-(\lambda x)^r}$ para $x > 0$, con inversa $F^{-1}(u) = (1/\lambda) \exp\left\{(1/r)\ln(-\ln(1-u))\right\}$ para $0 < u < 1$. Su esperanza es $E(X) = \Gamma(1+1/r)/\lambda$ y su varianza es $\text{Var}(X) = [\Gamma(1+2/r) - \Gamma^2(1+1/r)]/\lambda^2$. Cuando $r = 1$ la distribución Weibull se reduce a la distribución $\exp(\lambda)$. Cuando $r = 2$ se obtiene la distribución Rayleigh(λ).

Se puede usar el método de la función inversa para generar valores de la distribución Weibull. La función de distribución $F(x)$ es invertible y el método asegura que si u es un valor de la distribución $\text{unif}\,(0,1)$, entonces $x = (1/\lambda) \exp\left\{(1/r)\ln\left(\ln\left(1/u\right)\right)\right\}$ es un valor de la distribución $\text{Weibull}\,(r, \lambda)$. Cuando $r = 1$, la fórmula para x se reduce a la expresión presentada antes para la distribución $\exp(\lambda)$.

Pareto (tipo I)

Sean $a > 0$ y $b > 0$ dos parámetros. Una variable aleatoria X que toma

valores en el intervalo (b, ∞) tiene una distribución Pareto (tipo I) y se escribe $X \sim \text{Pareto}\,(a, b)$, cuando su función de densidad es

$$f(x) = a\, b^a\, x^{-a-1}, \qquad x > b.$$

Su función de distribución es $F(x) = 1 - [b/x]^a$ para $x > b$. Su esperanza es $E(X) = ab/(a-1)$ cuando $a > 1$ y su varianza es $\text{Var}(X) = ab^2/[(a-1)^2(a-2)]$ cuando $a > 2$.

Se puede usar el método de la función inversa para generar valores de la distribución Pareto (tipo I). La función de distribución $F(x)$ es invertible y el método asegura que si u es un valor de la distribución unif $(0, 1)$, entonces $x = bu^{-1/a}$ es un valor de la distribución Pareto (a, b).

t

Una variable aleatoria X que toma valores en el intervalo $(-\infty, \infty)$ tiene una distribución t de Student y se escribe $X \sim \text{t}(n)$, en donde $n > 0$ es un parámetro, cuando su función de densidad es

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \qquad -\infty < x < \infty.$$

Su esperanza es $E(X) = 0$ y su varianza es $\text{Var}(X) = n/(n-2)$ para $n > 2$.

A partir de valores de las distribuciones normal y ji-cuadrada se pueden generar valores de la distribución t usando el siguiente resultado: si Z tiene distribución $\text{N}(0, 1)$ y es independiente de Y, la cual tiene distribución $\chi^2(n)$, entonces $Z/\sqrt{Y/n}$ tiene distribución $\text{t}(n)$.

Cauchy

Una variable aleatoria X que toma valores en el intervalo $(-\infty, \infty)$ tiene una distribución Cauchy y se escribe $X \sim \text{Cauchy}\,(a, b)$, en donde $a \in \mathbb{R}$ y $b > 0$

son dos parámetros, cuando su función de densidad es

$$f(x) = \frac{1}{b\pi}\left[1 + \left(\frac{x-a}{b}\right)^2\right]^{-1}, \qquad -\infty < x < \infty.$$

Su función de distribución es $F(x) = 1/2 + (1/\pi)\arctan((x-a)/b)$ para $-\infty < x < \infty$. Su esperanza y varianza no existen. Cuando $a = 0$ y $b = 1$ se obtiene la distribución Cauchy estándar denotada por Cauchy $(0, 1)$, cuya función de densidad es

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty,$$

y su función de distribución es $F(x) = 1/2 + (1/\pi)\arctan x$, para $-\infty < x < \infty$. El siguiente resultado establece que es suficiente estudiar la distribución Cauchy estándar:

- $X \sim \text{Cauchy}\,(a, b) \Rightarrow (X - a)/b \sim \text{Cauchy}\,(0, 1)$.
- $X \sim \text{Cauchy}\,(0, 1) \Rightarrow a + bX \sim \text{Cauchy}\,(a, b)$.

Se puede usar el método de la función inversa para generar valores de la distribución Cauchy. La función de distribución $F(x)$ es invertible y el método asegura que si u es un valor de la distribución unif $(0, 1)$, entonces $x = a + b\tan((u - 1/2)\pi)$ es un valor de la distribución Cauchy (a, b). También se puede usar el método del cociente de uniformes para generar valores de esta distribución. En el Ejercicio 47 se pide aplicar este método a la distribución Cauchy estándar.

beta

Una variable aleatoria X que toma valores en el intervalo $(0, 1)$ tiene una distribución beta y se escribe $X \sim \text{beta}\,(a, b)$, en donde $a > 0$ y $b > 0$ son dos parámetros, cuando su función de densidad es

$$f(x) = \frac{1}{\text{B}(a, b)}\, x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1,$$

en donde $\mathrm{B}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ es la función beta y $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\,dt$ es la función gama. Su esperanza es $E(X) = a/(a+b)$ y su varianza es $\mathrm{Var}(X) = ab/[(a+b+1)(a+b)^2]$. Cuando $a > 1$ y $b > 1$, la función de densidad tiene un valor máximo en el punto $x^* = (\alpha-1)/(\alpha+\beta-2)$. El valor máximo de la función es $f(x^*) = (x^*)^{\alpha-1}\,(1-x^*)^{\beta-1}/B(\alpha,\beta)$. Cuando $a = b = 1$ la distribución beta se reduce a la distribución unif $(0,1)$. Cuando $a = b = 1/2$ se obtiene la distribución arcoseno.

Cuando $a > 1$ y $b > 1$, la distribución beta puede simularse usando el método de von Neumann que se expone en la Sección 2.3 pues el soporte es acotado y la función de densidad es acotada teniendo un valor máximo, o moda, igual a $M = f(x^*)$, con x^* en punto arriba indicado. El procedimiento se muestra en el Ejemplo 2.7 en la página 76.

F

Una variable aleatoria X que toma valores en el intervalo $(0,\infty)$ tiene una distribución F y se escribe $X \sim \mathrm{F}(n,m)$, en donde n y m son dos parámetros con valores enteros positivos a los que se les llama *grados de libertad*, cuando su función de densidad es

$$f(x) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)}\left(\frac{n}{m}\right)^{n/2}\left(1+\frac{nx}{m}\right)^{-(n+m)/2}, \qquad x > 0.$$

Su esperanza es $E(X) = m/(m-2)$ cuando $m > 2$ y su varianza es $\mathrm{Var}(X) = 2m^2(n+m-2)/[n(m-2)^2(m-4)]$ cuando $m > 4$.

A partir de valores de la distribución χ^2 se pueden generar valores de la distribución F usando el siguiente resultado: si $X \sim \chi^2(n)$ y $Y \sim \chi^2(m)$ son independientes, entonces $(X/n)/(Y/m)$ tiene distribución $\mathrm{F}(n,m)$.

Apéndice C: Códigos en R

En esta sección se muestran algunos códigos en el lenguaje de programación R con el fin de ilustrar la implementación de algunos algoritmos para la generación de muestras aleatorias y para la obtención de estimaciones utilizando simulación. La programación se realizó con fines ilustrativos y se buscó claridad en la sintaxis de las instrucciones, omitiendo el uso de paqueterías en la medida de lo posible. El lector puede reproducirlos, modificarlos y mejorarlos para lograr comprobar los resultados que desee.

Algoritmo para estimar $E(Y)$ en (1.1)

```
set.seed(123) #semilla
n <- 100
y <- numeric(n)
for(i in 1:n){
    w <- sample(1:6,size = 1) # resultado de lanzar un dado
    y[i] <- rbinom(1,w,1/2) # ocasiones en que cae "cruz"
}
EY <- mean(y)
cat('Estimación:',EY,'Valor exacto:',7/4)
```

```
## Estimación: 1.8 Valor exacto: 1.75
```

Algoritmo para estimar el número esperado de lanzamientos de una moneda hasta obtener dos caras consecutivas

```
set.seed(123) #semilla
n <- 100
lanzamientos <- numeric(n)
for(i in 1:n){
    contador <- 0
    aux <- 0
    while(aux<2){
        contador <- contador + 1
        moneda <- sample(0:1,1)
        if(moneda == 0){aux <- 0}else{aux <- aux + 1}
    }
    lanzamientos[i] <- contador
}
Elanzamientos <- mean(lanzamientos)
cat('Estimación:',Elanzamientos,'Valor exacto:',6)
```

```
## Estimación: 6.02 Valor exacto: 6
```

Algoritmo para estimar el valor de la integral $\int_0^1 e^{-x^2} dx$

```
set.seed(123) #semilla
n <- 100
Y <- runif(n,0,1)
E <- mean(exp(-Y^2))
f <- function(x){exp(-x^2)}
I <- as.numeric(integrate(f,0,1)[1])
cat('Estimación:',round(E,4),'Valor exacto:',round(I,4))
```

```
## Estimación: 0.7493 Valor exacto: 0.7468
```

Generador congruencial lineal simple

```
# Ejemplo 1.6
n <- 25
M <- 16; a <- 5; c <- 0
x0 <- 1
x <- numeric(n)
x[1] <- (a*x0+c)%%M
for(i in 2:n){x[i] <- (a*x[i-1]+c)%%M}
x
```

```
##  [1]  5  9 13  1  5  9 13  1  5  9 13  1  5
         9 13  1  5  9 13  1  5  9 13  1  5
```

```
##################
# Ejemplo 1.7
n <- 25
M <- 100; a <- 13; c <- 14
x0 <- 1
x <- numeric(n)
x[1] <- (a*x0+c)%%M
for(i in 2:n){x[i] <- (a*x[i-1]+c)%%M}
x
```

```
##  [1] 27 65 59 81 67 85 19 61  7  5 79 41 47 25
        39 21 87 45 99  1 27 65 59 81 67
```

Generador multiplicativo (1.20)

```
# Ejemplo 1.10 con a=7
n <- 25
M <- 11; a <- 7
x0 <- 1
x <- numeric(n)
```

```
x[1] <- (a*x0)%%M
for(i in 2:n){x[i] <- (a*x[i-1])%%M}
x  #Periodo q=M-1=10
```

```
##  [1]  7  5  2  3 10  4  6  9  8  1  7  5  2
        3 10  4  6  9  8  1  7  5  2  3 10
```

```
######################
# Ejemplo 1.10 con a=19
n <- 25
M <- 11; a <- 19
x0 <- 1
x <- numeric(n)
x[1] <- (a*x0)%%M
for(i in 2:n){x[i] <- (a*x[i-1])%%M}
x #Periodo q=M-1=10
```

```
##  [1]  8  9  6  4 10  3  2  5  7  1  8  9  6
        4 10  3  2  5  7  1  8  9  6  4 10
```

Prueba de Kolmogorov-Smirnov

```
# Ejemplo 1.14
set.seed(425)
algunos_datos <- runif(20,0,1)
ks.test(algunos_datos,'punif',0,1)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  algunos_datos
## D = 0.26275, p-value = 0.1045
## alternative hypothesis: two-sided
```

Prueba Ji-cuadrada

```
# Ejemplo 1.15
set.seed(425)
algunos_datos <- runif(20,0,1)
ad <- algunos_datos
O <- numeric(10)
for(i in 1:10){
    O[i] <- sum(ad>(i-1)/10 & ad<=i/10)
}
O
```

```
##  [1] 0 1 0 3 2 3 3 1 5 2
```

```
TT <- sum((O-2)^2)/2
pvalue <- 1-pchisq(TT,9)
pvalue
```

```
## [1] 0.2757089
```

Prueba de Cox-Stuart

```
# Ejemplo 1.16
set.seed(425)
numeros <- runif(20,0,1)
numeros
```

```
##  [1] 0.6078023 0.9073220 0.5299141 0.8476512 0.7802851
##  [6] 0.9418568 0.3448128 0.4627547 0.6645587 0.3105249
## [11] 0.3324742 0.8224943 0.8121696 0.5116627 0.1037860
## [15] 0.6346132 0.5954693 0.8505040 0.4974272 0.8757268
```

```
n <- length(numeros)
m <- n/2
numeros[1:m]<numeros[(m+1):n]
```

```
##  [1] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
```

```
TT <- sum(numeros[1:m]<numeros[(m+1):n])
p1 <- sum(dbinom(0:TT,m,1/2))
p2 <- sum(dbinom(TT:m,m,1/2))
pvalue <- 2*min(c(p1,p2))
pvalue
```

```
## [1] 0.7539063
```

Método de la función inversa

Ejemplo en donde se realiza una simulación de variables aleatorias cuya función de distribución, función de densidad y la función inversa de la función de distribución son como sigue:

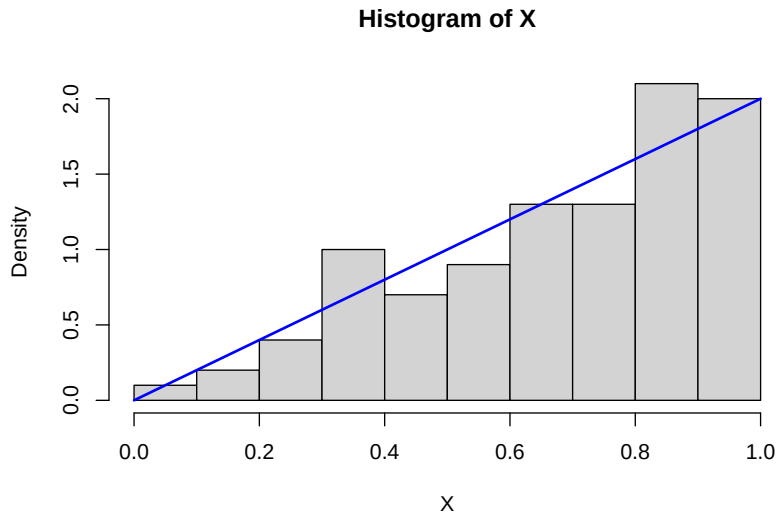
$$F(x) = \begin{cases} 0 & \text{si } x \leqslant 0, \\ x^2 & \text{si } 0 < x < 1, \\ 1 & \text{si } x \geqslant 1. \end{cases}$$

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

$$F^{-1}(u) = \sqrt{u}, \quad 0 < u < 1.$$

```
# Ejemplo 2.3
n <- 100
set.seed(425)
U <- runif(n)
X <- sqrt(U)
```

```
hist(X,freq = F)
curve(2*x,add=T,col='blue',lwd=2)
```



Simulación de variables aleatorias discretas

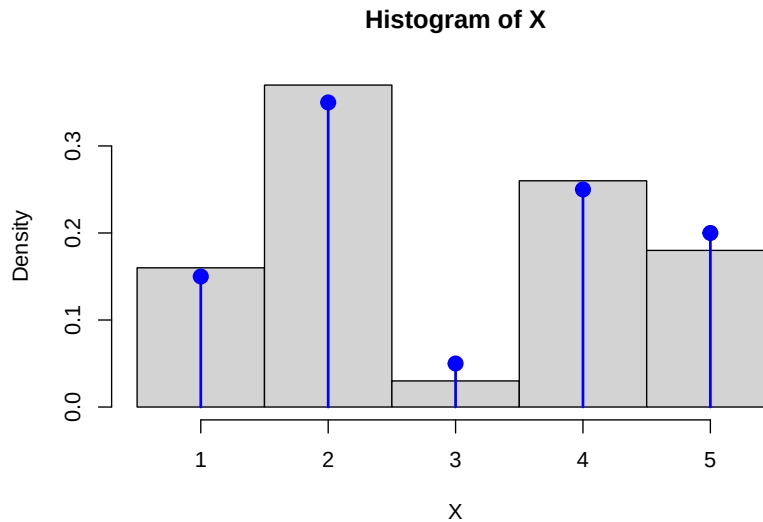
```
# Algoritmo figura 2.5
n <- 100
x <- 1:5
probabilidades <- c(0.15,0.35,0.05,0.25,0.2)
probabilidades
```

```
## [1] 0.15 0.35 0.05 0.25 0.20
```

```
Fx <- cumsum(probabilidades)
Fx
```

```
## [1] 0.15 0.50 0.55 0.80 1.00
```

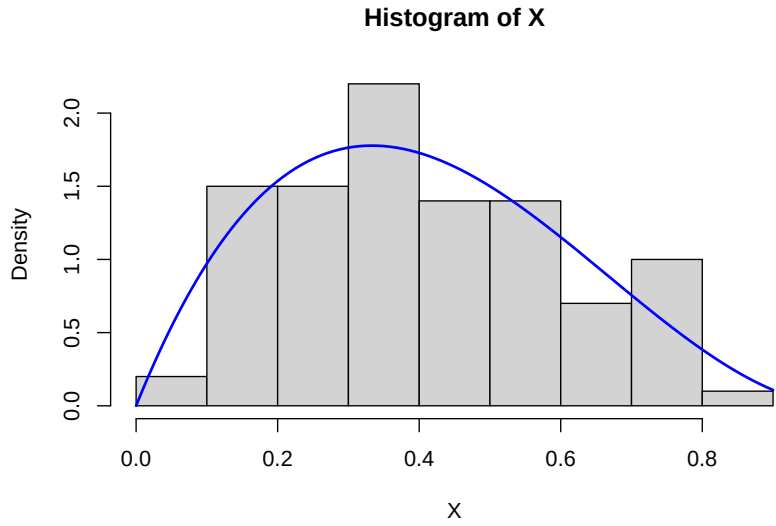
```
###
X <- numeric(n)
set.seed(123)
u <- runif(n)
for(k in 1:n){
  i <- 1
  while(u[k]>=Fx[i]){
    i <- i + 1
  }
  X[k] <- i
}
hist(X,breaks = seq(0.5,5.5,1),freq = F)
segments(1:5,rep(0,5),1:5,probabilidades,
         col='blue',lwd=2)
points(1:5,probabilidades,pch=19,cex=1.5,col='blue')
```



Método de von Newmann

Ejemplo en donde se realiza una simulación de variables aleatorias con distribución beta.

```
# Ejemplo 2.7 (distribución beta)
n <- 100
a <- 2
b <- 3
xstar <- (a-1)/(a+b-2)
M <- dbeta(xstar,a,b)
X <- numeric(n)
rechazos <- 0
set.seed(123)
for(i in 1:n){
  sigue <- TRUE
  while(sigue){
    u1 <- runif(1)
    u2 <- runif(1)
    x <- u1
    u <- M*u2
    if(u<=dbeta(x,a,b)){
      X[i] <- x
      sigue <- FALSE
    }else{
      rechazos <- rechazos + 1
    }
  }
}
###
hist(X,freq = F)
curve(dbeta(x,a,b),add=TRUE,col='blue',lwd=2)
```



Método de von Newmann caso discreto

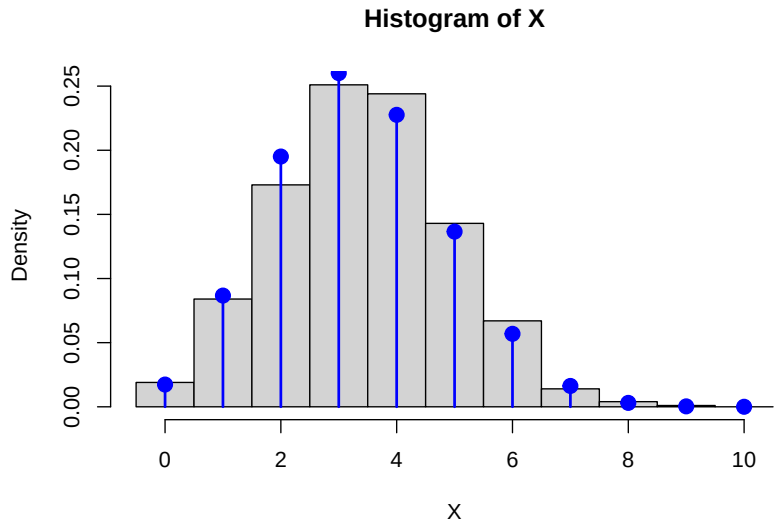
Ejemplo en donde se realiza una simulación de variables aleatorias con distribución binomial.

```
# Ejemplo 2.8 (distribución binomial)
m <- 1000
n <- 10
p <- 1/3
X <- numeric(m)
rechazos <- 0
set.seed(456)
for(i in 1:m){
  sigue <- TRUE
  while(sigue){
    u <- runif(1)
    x <- sample(0:n,1)
    if(u<=dbinom(x,n,p)){
      X[i] <- x
      sigue <- FALSE
    }else{
```

```
      rechazos <- rechazos + 1
    }
  }
}
###
rechazos
```

```
## [1] 9919
```

```
hist(X,breaks = seq(-0.5,n+0.5,1),freq = F)
segments(0:n,rep(0,n+1),0:n,dbinom(0:n,n,p),
         col='blue',lwd=2)
points(0:n,dbinom(0:n,n,p),pch=19,cex=1.5,col='blue')
```



Método de aceptación y rechazo

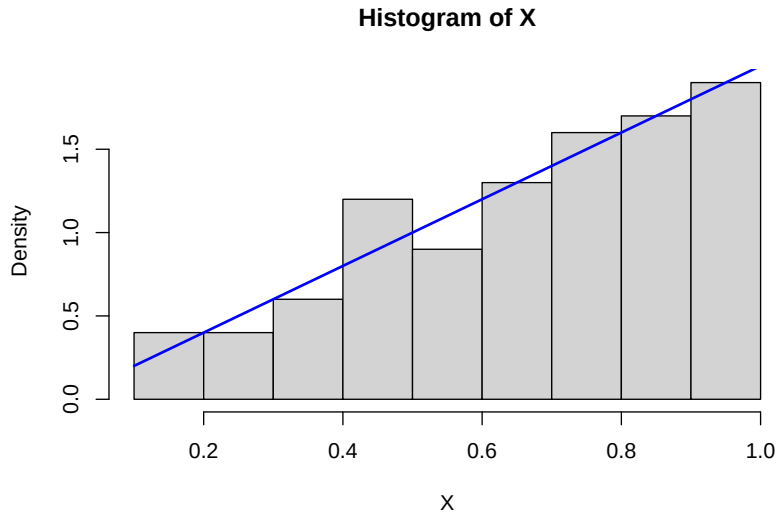
Ejemplo en donde se realiza una simulación de variables aleatorias con función de densidad

$$f(x) = 2x, \quad 0 < x < 1.$$

```
# Ejemplo 2.9 (aceptación y rechazo)
n <- 100
f <- function(x){2*x*(x>0)*(x<1)}
g <- function(x){1*(x>0)*(x<1)}
c <- 2
X <- numeric(n)
rechazos <- 0
set.seed(123)
for(i in 1:n){
  sigue <- TRUE
  while(sigue){
    u <- runif(1)
    x <- runif(1)
    if(u<=f(x)/(c*g(x))){
      X[i] <- x
      sigue <- FALSE
    }else{
      rechazos <- rechazos + 1
    }
  }
}
###
rechazos
```

```
## [1] 98
```

```
hist(X,freq = F)
curve(2*x,add=TRUE,col='blue',lwd=2)
```



Método de aceptación y rechazo discreto

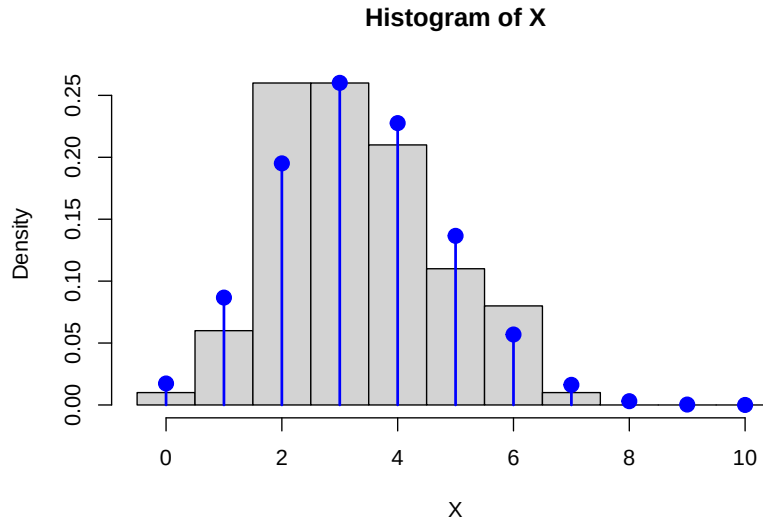
Ejemplo en donde se realiza una simulación de variables aleatorias con distribución binomial.

```
# Ejemplo 2.13 (aceptación y rechazo discreto)
m <- 100
#densidad binomial n=10 y p=1/3
f <- function(x){dbinom(x,10,1/3)}
#densidad propuesta geométrica theta=1-p=2/3
g <- function(x){dgeom(x,2/3)}
c <- (factorial(10)/(factorial(5)*factorial(5)))/(2/3)
X <- numeric(m)
rechazos <- 0
set.seed(456)
for(i in 1:m){
  sigue <- TRUE
  while(sigue){
    u <- runif(1)
    x <- rgeom(1,2/3)
    if(u<=f(x)/(c*g(x)) & x<=10){
```

```
      X[i] <- x
      sigue <- FALSE
    }else{
      rechazos <- rechazos + 1
    }
  }
}
###
rechazos
```

```
## [1] 40479
```

```
hist(X,breaks = seq(-0.5,10.5,1),freq = F)
segments(0:10,rep(0,11),0:10,dbinom(0:10,10,1/3),
         col='blue',lwd=2)
points(0:10,dbinom(0:10,10,1/3),pch=19,cex=1.5,col='blue')
```



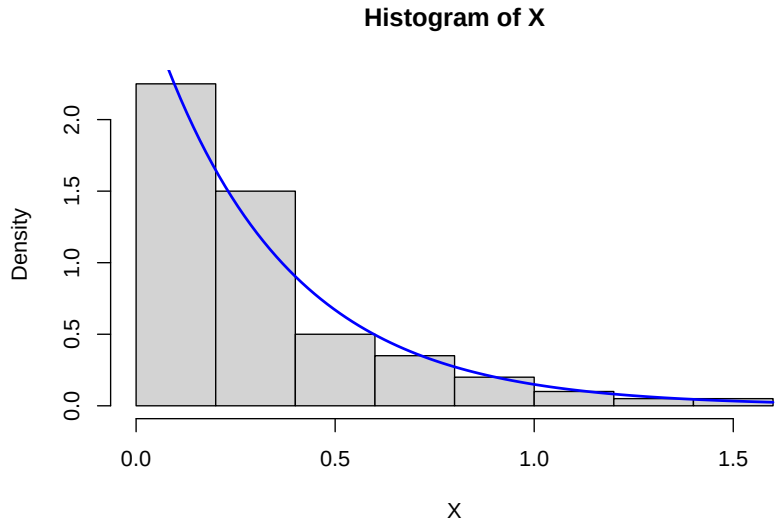
Cociente de uniformes

Ejemplo en donde se realiza una simulación de variables aleatorias con distribución exponencial.

```
# Ejemplo 2.15 (exponencial via cociente de uniformes)
n <- 100
lambda <- 3
X <- numeric(n)
rechazos <- 0
set.seed(123)
for(i in 1:n){
  sigue <- TRUE
  while(sigue){
    u <- runif(1)
    v <- runif(1)
    aux <- -2*u*log(u)
    if(v<aux){
      X[i] <- v/(lambda*u)
      sigue <- FALSE
    }else{
      rechazos <- rechazos + 1
    }
  }
}
###
rechazos
```

```
## [1] 99
```

```
hist(X,freq = F)
curve(dexp(x,lambda),add=TRUE, col='blue',lwd=2)
```



Integración Monte Carlo: método clásico

Ejemplo en donde se realiza una estimación de la integral

$$\int_1^5 \sqrt{x^2+1}\, dx,$$

utilizando la expresión (3.3).

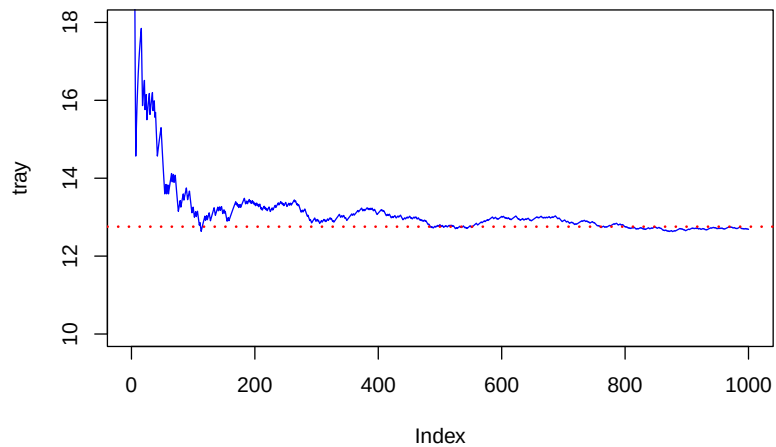
```
# Estimación (3.3)
n <- 1000
g <- function(x){sqrt(x^2+1)}
a <- 1
b <- 5
c <- g(5)
exacto <- round(as.numeric(integrate(g,1,5)[1]),3)
set.seed(123)
x <- runif(n,a,b)
y <- runif(n,0,c)
n_0 <- sum(y<g(x))
aprox <- c*(b-a)*(n_0/n)
cat('Estimación:',aprox,'Valor exacto:',exacto)
```

```
## Estimación: 12.68636 Valor exacto: 12.756
```

```
vm <- var(c*(b-a)*(y<g(x)))
cat('Varianza muestral:',vm)
```

```
## Varianza muestral: 97.90616
```

```
############################
# Evolución del promedio #
############################
tray <- c*(b-a)*cumsum(y<g(x))/1:n
plot(tray,type = 'l',ylim=c(10,18),col='blue')
abline(h=exacto,lty=3,lwd=2,col='red')
```



Integración Monte Carlo: método de la media muestral

Ejemplo en donde se realiza una estimación de la integral

$$\int_1^5 \sqrt{x^2+1}\, dx,$$

utilizando la expresión (3.10).

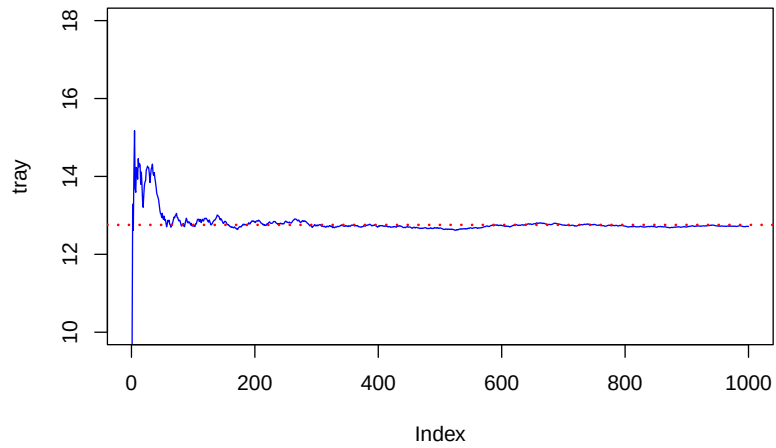
```
# Estimación (3.10)
n <- 1000
g <- function(x){sqrt(x^2+1)}
a <- 1
b <- 5
f <- function(x){dunif(x,a,b)}
exacto <- round(as.numeric(integrate(g,1,5)[1]),3)
set.seed(123)
X <- runif(n,a,b)
aprox <- mean(g(X)/f(X))
cat('Estimación:',aprox,'Valor exacto:',exacto)
```

```
## Estimación: 12.71421 Valor exacto: 12.756
```

```
vm <- var(g(X)/f(X))
cat('Varianza muestral:',vm)
```

```
## Varianza muestral: 18.46032
```

```
############################
# Evolución del promedio #
############################
tray <- cumsum(g(X)/f(X))/1:n
plot(tray,type = 'l',ylim=c(10,18),col='blue')
abline(h=exacto,lty=3,lwd=2,col='red')
```



Variables de control

Ejemplo 4.4 en donde se realiza una estimación de la integral

$$\int_0^1 e^{-x^2/2}\,dx,$$

utilizando la expresión (4.26).

```
# Ejemplo 4.4
n <- 100
g <- function(x){exp(-x^2/2)}
h <- function(x){x}
exacto <- round(as.numeric(integrate(g,0,1)[1]),5)
set.seed(123)
X <- runif(n)
aprox1 <- mean(g(X))
vm1 <- var(g(X))
C <- mean(h(X))
EC <- 1/2
alpha <- 12*(1-exp(-1/2)-aprox1/2)
cat('Valor de alpha=',alpha)
```

```
## Valor de alpha= -0.4218966
```

```
aprox2 <- aprox1-alpha*(C-EC)
vm2 <- var(g(X)-alpha*(h(X)-EC))
################################
cat('Estimación 1:',round(aprox1,5),'Valor exacto:',exacto)
```

```
## Estimación 1: 0.85725 Valor exacto: 0.85562
```

```
cat('Varianza muestral 1:',vm1)
```

```
## Varianza muestral 1: 0.01455488
```

```
################################
cat('Estimación 2:',round(aprox2,5),'Valor exacto:',exacto)
```

```
## Estimación 2: 0.85665 Valor exacto: 0.85562
```

```
cat('Varianza muestral 2:',vm2)
```

```
## Varianza muestral 2: 0.0005361832
```

```
#Varianza mucho menor a la varianza de la primera estimación.
```

Bibliografía

[1] Abramowitz M., Stegun I. A. (Editors) *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover, 1972.

[2] Asmussen S., Glynn P. W. *Stochastic simulation: algorithms and analysis.* Springer,2007.

[3] Brandimarte P. *Handbook in Monte Carlo simulation: applications in financial engineering, risk management, and economics.* Wiley, 2014.

[4] Chan N. H., Wong H. Y. *Simulation techniques in financial risk management.* John Wiley & Sons, 2006.

[5] Conover, W. J. *Practical nonparametric statistics.* John Wiley & Sons, 1999.

[6] Dagpunar J. S. *Simulation and Monte Carlo: with applications in finance and MCMC.* John Wiley & Sons, 2007.

[7] Davis P. J. *Leonhard Euler's integral: A historical profile of the gamma function.* Am. Math. Monthly (66), 849-869.

[8] Davison A. C., Hinkley D. V. (1997). *Bootstrap methods and their application.* Cambridge University Press.

[9] Dunn W. L., Shultis J. K. *Exploring Monte Carlo methods.* Elsevier, 2012.

[10] Efron B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* 7 (1) 1 - 26, January, 1979.

[11] Efron B. *The jackknife, the bootstrap and other resamplig plans.* SIAM, 1982.

[12] Fishman G. S. *Monte Carlo: concepts, algorithms and applications.* Springer, 1996.

[13] Fishman G. S. *Discrete-event simulation: modeling, programming and analysis.* Springer, 2001.

[14] Gamerman D., Lopes H. F. *Markov chain Monte Carlo: stochastic simulation for bayesian inference.* Chapman & Hall/CRC, 2006.

[15] S. Geman, D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1948). vol. PAMI-6, no. 6, pp. 721–741.

[16] Gentle J. E. *Random number generation and Monte Carlo methods.* Springer, 2003. Second edition.

[17] Glasserman P. *Monte Carlo methods in financial engineering.* Springer-Verlag, 2004.

[18] Gut A. *Probability: a graduate course.* Springer, 2005.

[19] Hastings W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* (1970) 57,1, p. 97.

[20] Henderson S. G., Nelson B. L. (Editors) *Handbooks in operations research and management science. Volume 13. Simulation.* North-Holland, 2006.

[21] Hochstadt, H. *The functions of mathematical physics.* Dover, 1986.

[22] Hogg R. V., Klugman S. A. *Loss distributions.* John Wiley & Sons, 1984.

[23] Hull, T. E., Dobell, A. R. Random number generators. *SIAM review* (1962) 4(3), 230-254.

[24] Jäckel P. *Monte Carlo methods in finance.* John Wiley & Sons, 2002.

[25] Johanse A., Evers L. *Monte Carlo methods.* Lecture Notes, Department of Mathematics, University of Bristol. Edited by Nick Whiteley, 2008.

[26] Johnson N. L., Kemp A. W., Kotz S. *Univariate discrete distributions.* 3rd. edition. Wiley-Interscience, 2005.

[27] Johnson N. L., Kotz S., Balakrishnan N. *Continuous univariate distributions.* Vol. 1, 2nd. edition. Wiley-Interscience, 1994.

[28] Johnson N. L., Kotz S., Balakrishnan N. *Continuous univariate distributions.* Vol. 2, 2nd. edition. Wiley-Interscience, 1995.

[29] Jones O., Maillardet R., Robinson A. *Introduction to scientific programming and simulation using* `R`. CRC Press, 2009.

[30] Karlin S., Taylor H. M. *A first course in stochastic processes.* 2nd. ed. Academic Press, 1975.

[31] Kalos M. H., Whitlock P. A. *Monte Carlo methods.* Wiley-VCH, 2004.

[32] Karr A. F. *Probability.* Springer, 1993.

[33] Kinderman A. J., Monahan J. F. Computer generation of random variables using the ratio of uniform deviates. *ACM Trans. Math. Software*, **3** (1977) pp. 257–260

[34] Kron R., Korn E., Kroisandt G. *Monte Carlo methods and models in finance and insurance.* Chapman & Hall/CRC, 2010.

[35] Lehmer D. H. Mathematical methods in large-scale-computing units. *Annals of the Computation Laboratory of Harvard University*, 26:141-146, 1951.

[36] L'Ecuyer P., Blouin F., Couture R. A search for good multiple recursive random number generators. *ACM Transaction on Modeling Computer Simulation* (3), 87-98, 1993.

[37] Matsumoto, M., Nishimura, T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation.* (1998) 8(1), 3-30.

[38] Marsaglia G., Bray T. A. A convenient method for generating normal variables. *SIAM review* **6** (3), 1964, 260–264.

[39] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H. y Teller E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* (1953) 21, 1087-1092.

[40] McLachlan G. J., Krishnan T. *The EM algorithm and extensions.* Wiley-Interscience, 2008. Second edition.

[41] McLeish D. L. *Monte Carlo Simulation and Finance.* John Wiley & Sons, 2005.

[42] Meyn, S. P., Tweedie, R. L. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

[43] Nelson B. L. *Foundations and methods of stochastic simulation: a first course.* Springer, 2013.

[44] Osais Y. E. *Computer simulation: a fundational approach using Python.* Chapman & Hall/CRC, 2018.

[45] Rao C. R. *Linear statistical inference and its applications.* John Wiley & Sons, 1973.

[46] Revuz D. *Markov Chains.* North–Holland, 1984.

[47] Rincón L. *Curso intermedio de probabilidad.* Las Prensas de Ciencias, UNAM, 2007.

[48] Ripley B. D. *Stochastic simulation.* John Wiley & Sons, 1987.

[49] Robert C. P., Casella G. *Monte Carlo statistical methods.* Springer, 2004. Second edition.

[50] Robert C. P., Casella G. *Introducing Monte Carlo methods with* `R`. Springer, 2010.

[51] Ross S. M. *Simulation.* Elsevier, 2006. Fourth edition.

[52] Rubinstein R. Y. *Simulation and the Monte Carlo method.* John Wiley & Sons, 1981.

[53] Rubinstein R. Y., Kroese D. P. *Simulation and the Monte Carlo method.* John Wiley & Sons, 2008. Second edition.

[54] Seydel R. U. *Tools for computational finance.* Springer, 2012. Fifth edition.

[55] Shao J. y Tu D. *The jackknife and boostrap.* Springer, 1995.

[56] Shonkwiler R. W., Mendivil F. *Explorations in Monte Carlo methods.* Springer, 2009.

[57] Sóbol I. M. *Método de Monte Carlo.* Lecturas Populares de Matemáticas. Mir, 1983. Segunda edición.

[58] Suess E. A., Trumbo B. E. *Introduction to probability simulation and Gibbs sampling with* `R`. Springer, 2010.

[59] Wackerly D., Mendenhall, W., Scheaffer, R. L. *Mathematical Statistics with Applications.* Cengage Learning, 2008.

[60] Watanabe M., Yamaguchi K. (Editors) *The EM algorithm and related statistical methods.* Marcel Dekker, 2004.

[61] Williams D. *Probability with martingales.* Cambridge University Press, 1991.

Índice alfabético

Elementos de simulación estocástica
fue editado por la Facultad de Ciencias
de la Universidad Nacional Autónoma de México.
Se terminó de editar el 20 de marzo de 2025.

Publicación electrónica.

El cuidado de la edición estuvo a cargo de
Leticia Pacheco Gasca.