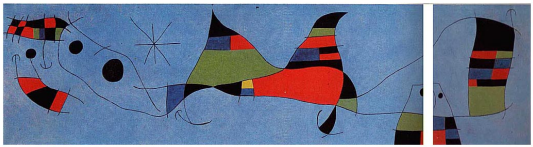
Simulación

Integración Monte Carlo

Jorge de la Vega Góngora

Departamento de Estadística,
Instituto Tecnológico Autónomo de México

Semana 13



Introducción

Introducción

- Dos grandes clases de problemas que surgen en inferencia estadística son:
  1. Optimización (ML, EM)
  2. Integración (estimadores, valores esperados, modelos Bayesianos)
- No siempre es posible resolver analíticamente estos problemas. En ambos casos se requieren soluciones numéricas, porque algunos problemas pueden ser muy complejos o sin solución analítica.
- Para integración, el problema a resolver siempre se puede escribir de la forma:

$$\theta = \int_{\chi} h(x) f(x)\, dx = \mathsf{E}_f[h(X)]$$

  donde f es la densidad de la variable aleatoria X, y h es una función arbitraria.
- Es importante notar que X puede ser un vector de variables aleatorias, y en ese caso la integral es multivariada.
- En este curso nos vamos a concentrar en integración. La parte de optimización se puede extender como proyecto, o revisar en el libro de Casella y Robert.

Métodos para integración

Se pueden tomar los siguientes enfoques para calcular $\theta = E[h(X)]$:

- (a) Encontrar el valor de manera analítica. Obviamente el método preferido si es posible.
- (b) Si la integral no se puede resolver analíticamente, se puede intentar integrales numéricas para tener un valor aproximado de la integral. Este método es bueno en dimensiones bajas, pero en espacios de dimensión grande puede ser muy costoso e ineficiente.
- (c) Integración de Monte Carlo. Esta técnica se basa en la ley de los grandes números, como se verá más adelante. Aunque en dimensiones bajas no es mejor que los métodos numéricos, en dimensiones grandes es muy eficiente.

Comentarios sobre enfoque (b)

- Hay muchas fórmulas determinísticas de cuadratura para el cálculo de integrales cuando el integrando se comporta “bien”, como la fórmula trapezoidal:

$$\int_a^b f(x)\,dx \approx (b-a)\left[\frac{f(a)+f(b)}{2}\right]$$

  o la regla de Simpson:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right].$$

- Sin embargo, hay situaciones en donde la función tiene soporte no acotado, no es bien comportada o la integral es multivariada y las fórmulas resultan muy complicadas de aplicar.
- En esos casos, los métodos de Monte Carlo son más simples y dan buenos resultados. Estos métodos fueron inventados en 1946 por Stanislaw Ulam, un matemático polaco que trabajó junto a John von Newman en el proyecto Manhattan durante la Segunda Guerra Mundial.

Montecarlo

Método crudo de Monte Carlo I

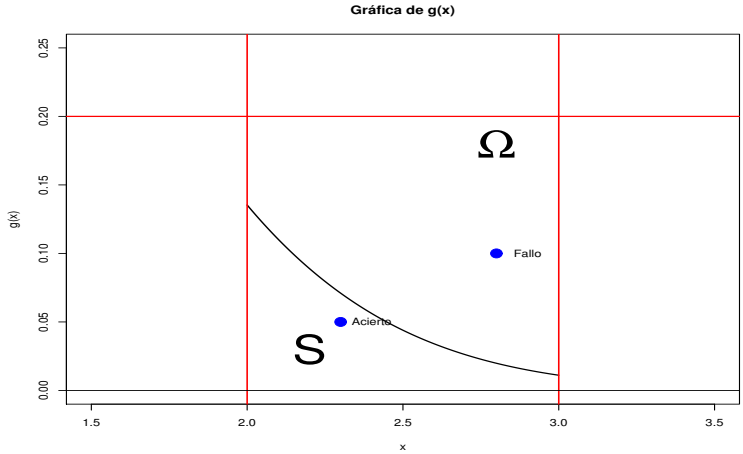
El método crudo de Monte Carlo parte de lo que hemos visto para estimar áreas a través de números uniformes.

- Comencemos con un ejercicio sencillo. Supongamos que queremos calcular el valor de una integral definida, e.g.

$$\theta = \int_2^3 e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}(\Phi(3) - \Phi(2))$$

  El valor 'exacto' de la integral es $\theta = 0.05364243$.
- Cada integral puede representarse como un *valor esperado* y el problema de estimar una integral vía Monte Carlo es equivalente al problema de estimar un parámetro desconocido θ.
- Sea Ω el rectángulo $\{(x, y)|2 \leq x \leq 3, 0 \leq y \leq 0.20\}$ y sea $S = \{(x, y)|y \leq g(x) = e^{-0.5x^2}\}$. Ahora lanzamos dardos "al azar" sobre Ω.

Método crudo de Monte Carlo II



Método crudo de Monte Carlo III

- "Al azar" significa que el vector de coordenadas (X, Y) tiene una *distribución uniforme* sobre Ω.
- El área bajo $g(x)$ es precisamente la integral que queremos, el área de S. La probabilidad de que un dardo pegue en S es:

$$p = \frac{\text{Área} S}{\text{Área} \Omega} = \frac{\theta}{0.2}$$

- Si lanzamos N dardos, y de éstos N_a son los que caen en el área S, podemos estimar a p con la proporción $\hat{p} = \frac{N_a}{N}$.
- Finalmente, una estimación de nuestra integral estaría dada por

$$\hat{\theta} = 0.2 \frac{N_a}{N}$$

Obteniendo el valor

- La siguiente tabla muestra valores obtenidos con diferentes lanzamientos.

N	$\hat{\theta}$	$\lvert\hat{\theta} - \theta\rvert$
10	0.100	0.04635757
50	0.044	0.00964243
100	0.054	0.00035757
1,000	0.0526	0.00104243
10,000	0.05268	0.00096243
100,000	0.053684	4.157e-05

- Hay dos puntos importantes a responder aquí:
  1. ¿Cuál es la variación esperada de $\hat{\theta}$?, y
  2. ¿Cuántos lanzamientos se requieren para lograr una precisión dada?

¿Cuál es la variación esperada de $\hat{\theta}$?

- Como cada lanzamiento de dardo acierta o falla, la distribución de cada lanzamiento es una variable de tipo Bernoulli, y suponemos que son independientes. Entonces:

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{N} = \frac{\theta(0.2-\theta)}{0.04N}$$

  y por lo tanto, la varianza de $\hat{\theta}$ es

$$\text{Var}(\hat{\theta}) = 0.04\text{Var}(\hat{p}) = \frac{\theta(0.2-\theta)}{N}.$$

- El error estándar es un estimador de la desviación estándar, reemplazando θ por $\hat{\theta}$:

$$se(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(0.2-\hat{\theta})}{N}}.$$

¿Cuál es la variación esperada de $\hat{\theta}$?

- Ahora podemos completar la tabla anterior:

N	$\hat{\theta}$	$\lvert\hat{\theta} - \theta\rvert$	error estándar $= \hat{\sigma}_{\theta}$
10	0.100	0.04635757	0.03162
50	0.044	0.00964243	0.01172
100	0.054	0.00035757	0.00888
1,000	0.0526	0.00104243	0.00278
10,000	0.05268	0.00096243	0.00088
100,000	0.053684	4.157e-05	0.00028

- La segunda pregunta es más útil en la práctica, pero requiere más teoría de probabilidad.

¿Cuántos lanzamientos son necesarios?

- En términos probabilísticos, queremos encontrar N tal que

$$P[|\theta - \hat{\theta}| < \epsilon] \geq \alpha,$$

  donde ϵ y α son determinadas.
- Aplicando la desigualdad de Chebyshev, $P[|\theta - \hat{\theta}| < \epsilon] \geq 1 - \frac{\text{Var}(\hat{\theta})}{\epsilon^2}$, obtenemos que

$$\alpha \leq 1 - \frac{\text{Var}(\hat{\theta})}{\epsilon^2}$$

- En $\text{Var}(\hat{\theta})$ aparece N. Despejando obtenemos que:

$$N \geq \frac{0.04p(1-p)}{(1-\alpha)\epsilon^2}$$

¿Cuántos lanzamientos son necesarios?

- La siguiente tabla muestra los resultados para diferentes combinaciones de ϵ, α y p.

ϵ	α	p	N
0.001	0.90	0.5	100
0.00001	0.95	0.01	792
0.00001	0.95	0.6	19,200
0.0001	0.999	0.5	100,000

- La desigualdad de Chebyshev da una estimación conservadora de N. Usualmente es mejor intentar valores más grandes.
- Con el poder computacional con el que se cuenta actualmente, ya no es limitación importante encontrar valores grandes de N.
- Este método muestra una aplicación más de la técnica de *aceptación-rechazo*.

Método Mejorado de Monte Carlo

- El Método Mejorado de Monte Carlo es una generalización del método crudo de Monte Carlo, para evaluar la integral sobre muestras de variables aleatorias con otras distribuciones y no sólo de la distribución uniforme.
- Introducir otras distribuciones puede ayudar a acelerar la convergencia y requerir menores tamaños de muestra.
- El soporte de MC sigue siendo la Ley de los grandes números.

Método Mejorado de Monte Carlo

Para evaluar la integral

$$\mathsf{E}_f[h(X)] = \int_\chi h(x) f(x)\, dx$$

- Se obtiene una muestra $x_1, \ldots, x_n \sim f$
- Calcula $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n h(x_i)$
- La varianza asintótica de la aproximación está dada por:

$$\mathsf{Var}(\hat{\theta}_n) = \frac{1}{n} \int_\chi (h(x) - \theta)^2 f(x)\, dx,$$

con estimador:

$$\nu_n = \widehat{\mathsf{Var}}(\hat{\theta}_n) = \frac{1}{n^2} \sum_{i=1}^n (h(x_i) - \hat{\theta}_n)^2$$

- Además, por el TLC:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\nu_n}} \dot{\sim} \mathcal{N}(0, 1)$$

Justificación

El fundamento para el algoritmo anterior es la Ley fuerte de los grandes números:

Ley (fuerte) de los grandes números

Si X es una variable aleatoria con la misma distribución que X_i y suponiendo que $h : \mathbb{R} \to \mathbb{R}$ es una función acotada, entonces $h(X_1), h(X_2), \ldots$ es una sucesión de variables independientes e idénticamente distribuidas con media finita y

$$P\left(\lim_{n\to\infty} \frac{h(X_1) + \cdots + h(X_n)}{n} = \mathsf{E}[h(X)]\right) = 1$$

La ley débil establece que la convergencia se da en probabilidad: para cualquier $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{h(X_1) + \cdots + h(X_n)}{n} - \mathsf{E}[h(X)]\right| < \epsilon\right) = 1$$

Ejemplo MC I

- Usando el mismo ejemplo que vimos en Monte Carlo crudo:

$$\theta = \int_2^3 e^{-\frac{x^2}{2}}\, dx = \sqrt{2\pi}(\Phi(3) - \Phi(2))$$

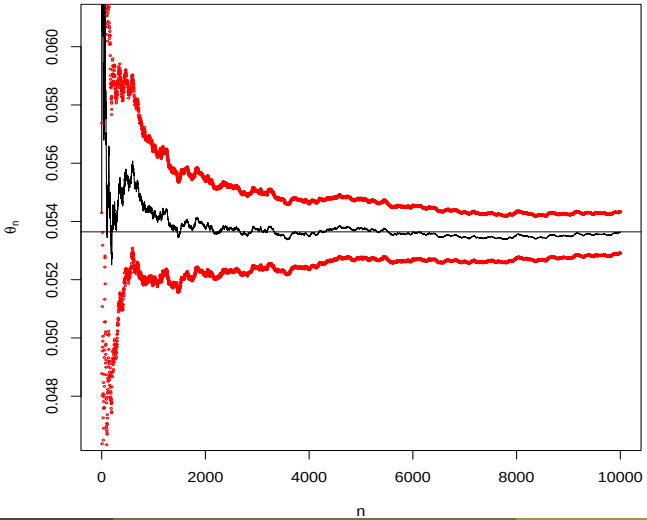
- Podemos considerar generar una muestra aleatoria de valores de una $\mathcal{U}(2,3)$,

```
n <- 10000                  #máximo tamaño de muestra
h <- function(x)exp(-x^2/2)
x <- h(runif(n,2,3))        #Genera una muestra grande de observaciones
theta_n <- cumsum(x)/(1:n)  #estimadores de la integral para diferentes n
vn <- sqrt(cumsum((x-theta_n)^2))/(1:n)  #error estándar estimado para cada n

plot(theta_n, type = "l",ylim = mean(theta_n) + c(-1,1)*20*vn[n],
                         ylab = expression(theta[n]),
                         xlab = "n")

#Agrega líneas correspondientes a nivel de confianza del 95% para cada n
points(theta_n + c(-1,1)*2*vn, col = "red",cex = 0.3)
abline(h = sqrt(2*pi)*(pnorm(3)-pnorm(2)))
```

Ejemplo MC II



Ejemplo MC I

- Alternativamente, podemos generar muestras de una $\mathcal{N}(0,1)$ (escalada por la constante) y considerar la función indicadora en el intervalo $(2,3)$:

$$\theta = \int_2^3 e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi}\int_{-\infty}^{\infty}\phi(x)I(2<x<3)\,dx$$

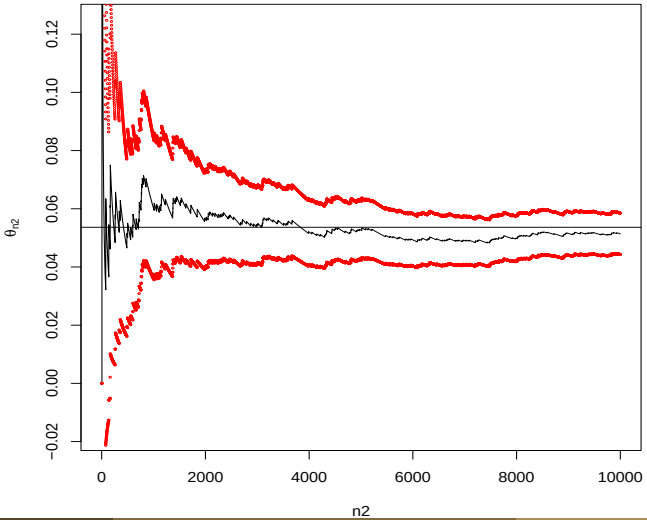
- ¿Converge a la misma velocidad? Lo veremos más adelante, está relacionado al tamaño de la varianza del estimador

```
#Mismo ejercicio, invirtiendo las distribuciones
g <- function(x)ifelse(x>=2 & x<=3,1,0) #función indicadora
x <- sqrt(2*pi)*g(rnorm(n))
theta_n2 <- cumsum(x)/(1:n)
vn2      <- sqrt(cumsum((x-theta_n2)^2))/(1:n)  #error estándar estimado para cada n

plot(theta_n2,type = "l", ylim = mean(theta_n2)+c(-1,1)*20*vn2[n],
              ylab = expression(theta[n2]),
              xlab = "n2")

#Agrega lineas correspondientes a nivel de confianza del 95% para cada n
points(theta_n2+c(-1,1)*2*vn2,col = "red", cex = 0.3)
abline(h = sqrt(2*pi)*(pnorm(3)-pnorm(2)))
```

Ejemplo MC II



Ejemplo 2 I

Ejercicio 3.1, Casella y Robert

El siguiente ejercicio muestra que puede haber algunos problemas cuando ν_n no es un estimador adecuado de la varianza de $\hat{\theta}_n$ o cuando no converge o converge de manera muy lenta. En estos casos, el estimador y la región de confianza asociada no será de confianza.
Supongan que tenemos una observación $X \sim \mathcal{N}(\theta, 1)$ y una distribución inicial para $\theta \sim Cauchy(0, 1)$. Queremos actualizar la información de θ basada en la información que provee X. En este contexto, la verosimilitud es:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\theta)^2\right]$$

con distribución inicial

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}$$

Ejemplo 2 II

Ejercicio 3.1, Casella y Robert

Usando el teorema de Bayes, la distribución posterior es proporcional a la verosimilitud por la inicial, esto es:

$$\pi(\theta|x) \propto \exp\left[-\frac{1}{2}(x-\theta)^2\right] \frac{1}{(1+\theta^2)},$$

donde la constante de proporcionalidad es el recíproco de C con

$$C = \int \exp\left[-\frac{1}{2}(x-\theta)^2\right] \frac{1}{(1+\theta^2)}\, d\theta.$$

El estimador puntual para θ (también conocido como estimador de Bayes para el modelo normal-Cauchy) es la media posterior, dada por

$$\delta(x) = \mathsf{E}(\theta|x) = \frac{\int \theta \exp\left[-\frac{1}{2}(x-\theta)^2\right] \frac{1}{(1+\theta^2)}\, d\theta}{C}$$

El ejercicio pide resolver la ecuación para $x = 0, 2, 4$.

Ejemplo 2 I

Ejercicio 3.1, Casella y Robert

- Grafica los integrandos, y usa integración de Monte Carlo basada en la simulación Cauchy para calcular las integrales.

Noten que para C tenemos a la Cauchy multiplicada por *algo*, donde *algo* es

$$h(\theta) = \pi \exp\left[-\frac{1}{2}(x-\theta)^2\right]$$

asi que podemos aplicar MC con el siguiente algoritmo:
Dado el valor observado $X = x$

1. simular una muestra de Cauchys: $\theta_1, \theta_2, \ldots, \theta_n$.
2. Estimar la integral en el denominador con $\frac{1}{n}\sum_{i=1}^{n} \pi \exp\left[-\frac{(x-\theta_i)^2}{2}\right]$.
3. Estimar la integral en el numerador con $\frac{1}{n}\sum_{i=1}^{n} \pi\theta_i \exp\left[-\frac{(x-\theta_i)^2}{2}\right]$.
4. Define la razón $\delta(x)$.

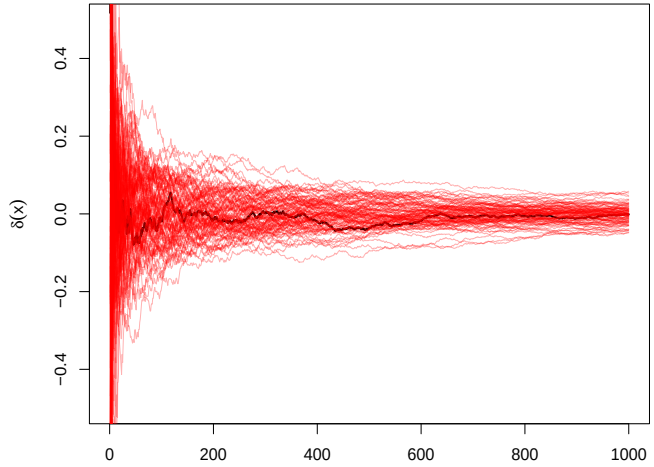
Ejemplo 2 II

Ejercicio 3.1, Casella y Robert

```
delta <- function(x){
        #estimador de Bayes
        n <- 1000 #número de valores simulados
        x <- 0  #valor de la variable observada
        th <- rcauchy(n)
        h <- function(x){pi*exp(-0.5*(x-th)^2)}
        cumsum(th*h(x))/cumsum(h(x))
}
rojo=rgb(1,0,0,alpha=0.3)
plot(delta(1),type="l",ylim=c(-0.5,0.5),ylab=expression(delta(x)),lwd=2)
for(i in 1:100)lines(delta(1),type="l",col=rojo)
```

Ejemplo 2 III

Ejercicio 3.1, Casella y Robert



Monte Carlo multidimensional

- Podemos escribir en el caso multivariado

$$\begin{aligned} \theta &= \int_\Omega f dV \\ &\approx \text{Volumen de } \Omega \times \text{promedio de } f \text{ en } \Omega \end{aligned}$$

- Para integración unidimensional, los métodos de MC son muy ineficientes. Pero para integración multidimensional es muy eficiente.
- Como el error es proporcional a $1/\sqrt{n}$, no depende de la dimensión.
- Se pueden manejar fácilmente regiones con fronteras irregulares.

Ejemplo 3

Monte Carlo multidimensional

Evaluar la integral

$$\theta = \int \int_{\Omega} sen(\sqrt{log(x+y+1)})dxdy$$

donde Ω es el disco definido por la condición $(x-0.5)^2 + (y-0.5)^2 \leq 0.25$.

Solución.

Considerando que la región de integración está en el cuadro unitario, es posible simular de uniformes independientes y quedarnos con los valores que estén dentro de la región de integración

```
N <- 1e6 #numero de simulaciones
X <- cbind(runif(N),runif(N)) #matriz de uniformes
      #solo nos quedamos con las observaciones que están en la región considerada
X1 <- X[(X[,1]-0.5)^2+(X[,2]-0.5)^2 <=0.25,]
N1 <- dim(X1)[1] # dimensión que queda con sólo aceptados
thetahat <- sin( sqrt( log(X1[,1] + X1[,2] + 1)))
a <- (pi/4)*cumsum(thetahat)/(1:N1) #se ajusta por el volumen del disco
a[N1]

[1] 0.5678335

sd(thetahat)/sqrt(N1) #error estándar

[1] 9.171498e-05
```

Ejemplo 4 I

En la aproximación

$$\theta \approx \text{Volumen de } \Omega \times \text{promedio de } f \text{ en } \Omega$$

se puede estimar el volumen de la región Ω al mismo tiempo que se estima la función f.
Se generan puntos en un volumen V que puede ser usualmente rectangular, tal que $\Omega \subseteq V$, y se generan n puntos en V de los cuales k están también en Ω, entonces

$$Vol(\Omega) \approx \frac{k}{n} Vol(V)$$

Como el promedio de f en Ω es aproximado a $\frac{1}{k}\sum f$, entonces las k se cancelan y se tiene:

$$\theta \approx Vol(V) \times \frac{1}{n} \sum_{p_i \in \Omega} f(p_i)$$

Por ejemplo, evaluar: $\theta = \int \int_{\Omega} y dx dy$ con Ω la semi-elípse Ω dada por

$$x^2 + 4y^2 \leq 1, \quad y \geq 0$$

Ejemplo 4 II

Se consideramos V como el rectángulo $[-1,1] \times [0,0.5]$, podemos generar $X \sim \mathcal{U}(-1,1)$ y $y \sim \mathcal{U}(0,0.5)$ El área del rectángulo es 1. Entonces:

```
N <- 1e6 #numero de simulaciones
X <- cbind(runif(N,-1,1),runif(N,0,0.5)) #matriz de uniformes
#solo nos quedamos con las observaciones que están en la región considerada
X1 <- X[X[,1]^2 + 4*X[,2]^2 <= 1,]
N1 <- dim(X1)[1] # dimensión que queda con sólo aceptados
thetahat <- X1[,2]
a <- sum(thetahat)/N
a

[1] 0.166667

sd(thetahat)/sqrt(N) #error estándar

[1] 0.0001321896
```

El valor exacto de la integral es $1/6$. (tarea)

Ejemplo 1 - I

Dagpunar 1.1, Casella-Robert. 3.1

Consideremos la función gamma:

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x}\, dx$$

y calculemos su valor de dos formas diferentes:

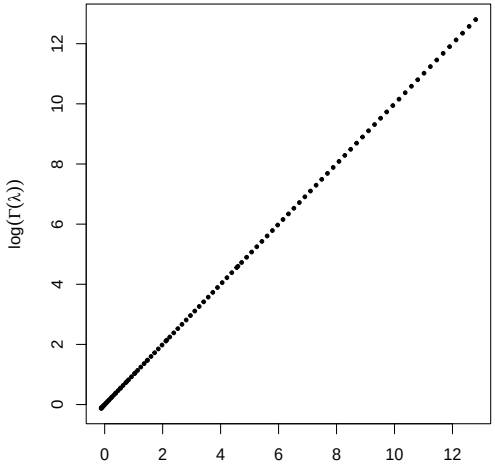
- usando la función `integrate`
- usando el método crudo de Monte Carlo.

Forma 1:
En este ejemplo, la función `integrate` funciona bien.

```
# Usando la función integrate y comparando contra la función (log) gamma:
h <- function(lambda){integrate(function(x){x^{lambda-1}*exp(-x)},0,Inf)$val}
x <- seq(0.01,10,length=100) #soporte
par(pty="s") #hacemos el plot cuadrado
plot(lgamma(x),log(apply(as.matrix(x),1,h)), xlab = "log(integrate(h(lambda)))",
                          ylab = expression(log(Gamma(lambda))),
                          pch=19, cex=0.5)
```

Ejemplo 1 - II

Dagpunar 1.1, Casella-Robert. 3.1



Ejemplo I

Dagpunar 1.1, Casella-Robert. 3.1

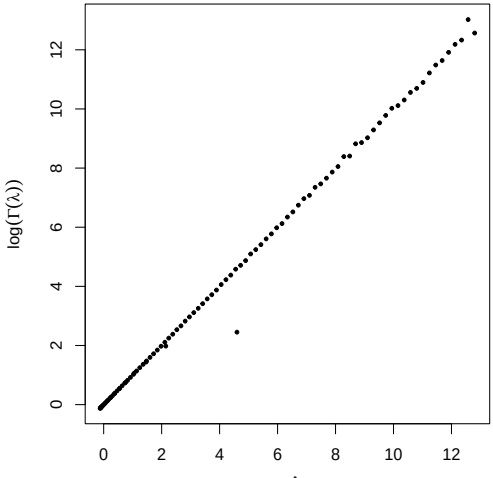
Forma 2:
Ahora consideremos la estimación usando Monte Carlo Crudo.

- Podemos usar una variable aleatoria X exponencial: $X \sim \exp(1)$ donde $f(x) = e^{-x}$ en $[0, \infty)$.
- Entonces podemos escribir a $\theta = \mathsf{E}_f(X^{\lambda-1})$.
- Extraemos una muestra de la distribución exponencial con parámetro 1 y calculamos el promedio $\hat{\theta}_n = \frac{1}{n}\sum_{I=1}^{n} X_i^{\lambda-1}$. Aquí usamos una muestra de tamaño $n = 1,000,000$.

```
# Usando Monte Carlo crudo y comparando contra la función (log) gamma:
h <- function(lambda){mean(rexp(1e6,1)^(lambda-1))}
x <- seq(0.01,10,length=100) #soporte
par(pty="s") #hacemos el plot cuadrado
plot(lgamma(x),log(apply(as.matrix(x),1,h)),
             xlab = expression(log(hat(theta)[lambda])),
             ylab =  expression(log(Gamma(lambda))),
             pch=19, cex=0.5)
```

Ejemplo II

Dagpunar 1.1, Casella-Robert. 3.1



Variabilidad del estimador $\hat{\theta}_n$

- Como vemos, la estimación en general no es mala, pero tenemos variabilidad que depende de la muestra. El estimador $\hat{\theta}_n$ es un estimador insesgado de θ:

$$\mathsf{E}(\hat{\theta}_n) = \mathsf{E}\left(\frac{1}{n}\sum_{I=1}^{n} X_i^{\lambda-1}\right) = \mathsf{E}(X_1^{\lambda-1}) = \theta$$

- Si la muestra es independiente, la varianza de $\hat{\theta}_n$ está dada por:

$$\mathsf{Var}_f(\hat{\theta}_n) = \frac{1}{n}\mathsf{Var}_f(X^{\lambda-1})$$

  y su desviación estándar es: $\sigma_f(\hat{\theta}_n) = \sigma_f(X^{\lambda-1})/\sqrt{n}$. El error estándar es el estimador de la desviación estándar:

$$se(\hat{\theta}_n) = \hat{\sigma}_f(X^{\lambda-1})/\sqrt{n}.$$

- Noten que para cambiar el error estándar en un factor de K, se requiere que la muestra cambie por un factor de $1/K^2$, lo que hace ineficiente el proceso:

$$K \cdot se(\hat{\theta}_n) = K\hat{\sigma}_f(X^{\lambda-1})/\sqrt{n} = \hat{\sigma}_f(X^{\lambda-1})/\sqrt{n/K^2}$$

Ejemplo 2

Consideremos un caso particular de la integral anterior, con $\lambda = 1.9$:

$$\theta = \int_0^\infty x^{0.9} e^{-x}$$

Considerando una muestra de $n = 100$ observaciones:

```
x <- rexp(100,1) #genera exponenciales con parámetro 1:
theta <- mean(x^{0.9})
theta

[1] 0.9992371

sd(x) #desviación estándar muestral

[1] 0.8666727

sd(x)/sqrt(100) #error estándar

[1] 0.08666727
```

¿De qué tamaño tiene que ser la muestra para reducir el error estándar a 0.0001? Como $K = 0.08638555/0.0001 = 863.8556$, entonces el tamaño de muestra tiene que ser del orden de $100 \times 863.86^2 = 74,625,410$

Sobre la función `integrate` y `area`

- Ambas funciones sólo son para integrales unidimensionales.
- La función `integrate` utiliza un método de cuadratura. Si la función a integrar es casi constante (o cero) en su rango, es posible que el resultado de la estimación y su error puedan ser muy equivocados. Esta función es muy frágil.
- La función `area` es parte del paquete `MASS`, no acepta límites infinitos, por lo que se requiere conocer de antemano el comportamiento de la función en la región de integración.
- Más adelante veremos un ejemplo de problemas que pueden surgir con ambas funciones, pero antes tenemos que revisar temas Bayesianos, para introducir el contexto de los ejemplos.

Integración en el contexto Bayesiano

Contexto Bayesiano I

- El paradigma Bayesiano se basa en el teorema de Bayes:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{\int P(B|A)P(A)\,dA} \propto P(B|A)P(A)$$

- Este teorema se aplica para desarrollar un sistema de aprendizaje: Una persona modifica su afirmación o creencia inicial de probabilidad sobre los parámetros antes de observar datos $y = (y_1, \ldots, y_n)$ a un conocimiento posterior o actualizado que combina el conocimiento inicial y los datos que se observan.
- Sea $\boldsymbol{\theta}$ un vector de parámetros. El conocimiento inicial se $\boldsymbol{\theta}$ se resume en una distribución inicial $\pi(\boldsymbol{\theta})$. La verosimilitud es $f(y|\boldsymbol{\theta})$ y el conocimiento actualizado está contenido en la distribución posterior $\pi(\boldsymbol{\theta}|y)$. Aplicando el teorema de Bayes,

$$\pi(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(y)} \propto f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

Contexto Bayesiano II

donde $m(y)$ es la verosimilitud marginal. Este valor se puede expresar como una integral: $m(y) = \int_\Theta f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, pero al no depender de $\boldsymbol{\theta}$ se puede considerar como una constante para normalizar $\pi(\boldsymbol{\theta}|y)$ y garantizar que sea una densidad.

- Noten que las integrales que se menciona en este modelo ya no son integrales unidimensionales, sino de la dimensión del vector de parámetros $\boldsymbol{\theta}$.
- Con la distribución posterior surgen varias cantidades de interés a estimar. A partir de la distribución posterior, se pueden estimar funciones de $\boldsymbol{\theta}$, usualmente de la forma $\mathrm{E}[h(\boldsymbol{\theta}|y)]$.
  - Probabilidad posterior de que $h(\theta)$ esté en un cierto conjunto A:

$$P(h(\boldsymbol{\theta}) \in A|y) = \frac{\int_{h(\boldsymbol{\theta})\in A} \pi(\boldsymbol{\theta}) f(y|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int \pi(\boldsymbol{\theta}) f(y|\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

  - Distribuciones marginales del vector $\boldsymbol{\theta}$:

$$\pi(\theta_1) \propto \int \pi(\boldsymbol{\theta}_{-j}, \theta_j)\, d\boldsymbol{\theta}_{-j}$$

Contexto Bayesiano III

- Densidades predictivas:

$$m(\tilde{y}) = \int f(\tilde{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Queda claro que todos estos casos generan problemas de integración.

El siguiente ejemplo parte de un contexto Bayesiano, y podemos ver el comportamiento y problemas de las funciones `integrate` y `area`.

- Ahora se tiene una muestra de tamaño $n = 10$ de una distribución Cauchy con parámetro de localizacion $\theta = 350$. La marginal de la muestra bajo una distribución inicial $\pi(\theta)$ queda como:

$$m(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)\pi(\theta)\, d\theta = \int_{-\infty}^{\infty} \prod_{i=1}^{10} \frac{1}{\pi[1 + (x_i - \theta)^2]}\, d\theta,$$

  considerando que la inicial es plana. En este caso, la función `integrate` no funciona bien, comparado con la función `area`:

Otro Ejemplo II

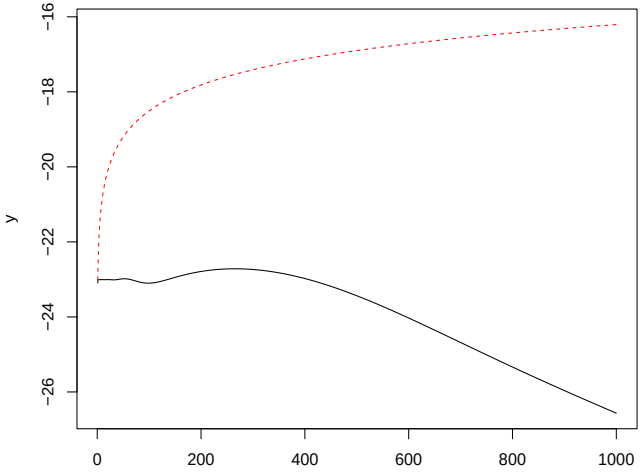
Casella-Robert 3.2

```
library(MASS) #carga para la función area
cac <- rcauchy(10)
verosimilitud <- function(th){
        #define la función de verosimilitud, a partir de la muestra dada, como función de theta
                u <- dcauchy(cac[1]-th)
                for(i in 2:10){
                        u <- u*dcauchy(cac[i]-th)
                }
                return(u)
}

f1 <- function(a){integrate(verosimilitud,-a,a)$val}
f2 <- function(a){area(verosimilitud,-a,a)}
x <- seq(1,1000,length=10^4) #partición de intervalo de 1 a 1000 muy fina.
y <- log(apply(as.matrix(x),1,f1))
z <- log(apply(as.matrix(x),1,f2))
plot(x,y, type = "l", ylim = range(cbind(y,z)))
lines(x, z, lty = 2, col = "red")
```

Otro Ejemplo III

Casella-Robert 3.2



Técnicas de Reducción de varianza

Introducción

- El objeto de las técnicas de reducción de varianza en los métodos de Montecarlo es mejorar la velocidad y eficiencia estadística de un estudio de simulación.
- Un estudio de simulación que utiliza entradas aleatorias genera salidas aleatorias, y por lo tanto tienen variabilidad que es necesario comprender.
- Por otra parte un estudio puede consumir muchos recursos como grandes cantidades de números aleatorios, y múltiples replicaciones para obtener un número puede ser demasiado costoso, por lo que es importante tratar de minimizar su variabilidad para obtener mayor precisión.

Introducción

- Usualmente en simulación estocástica se desea estimar $\theta = \mathrm{E}h(\mathbf{X})$. Como hemos visto, el algoritmo de simulación estándar dice:
  1. Genera $\mathbf{X}_1, \ldots, \mathbf{X}_n$
  2. Estima θ con $\hat{\theta}_n = \sum_{i=1}^{n} Y_i$ donde $Y_i = h(\mathbf{X}_i)$
- Un intervalo de confianza aproximado del $100(1-\alpha)\,\%$ está dado por:

$$[\hat{\theta}_n - z_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\theta}_n + z_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}]$$

  donde $\hat{\sigma}_n^2 = \hat{\mathrm{Var}}(\hat{\theta}_n) = \frac{\sum(Y_j - \bar{Y})^2}{n-1}$
- Una forma de medir la calidad de un estimador es con la *longitud media* HW del intervalo de confianza,

$$HW = z_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}$$

- Usualmente se desea que HW sea lo más pequeño posible, pero esto es a veces difícil de lograr.

Introducción

- Una forma de reducir la longitud media puede ser a través de las Técnicas de reducción de varianza. Éstas incluyen:
  - Variadas antitéticas
  - Variadas de control
  - Condicionamiento
  - Números seudo-aleatorios comunes
  - Importance sampling
  - Muestreo estratificado
- Hay muchas otras técnicas sofisticadas o generalizaciones de las que se mencionan aquí.
- En la mayoría de los casos, consideraremos la siguiente situación: X es una variable aleatoria *de salida*, es decir, es una variable relevante del estudio de simulación. En muchas situaciones lo que se desea es estimar $E(X) = \mu$, la media del proceso. Típicamente, entonces, este número es una integral.

Variadas antitéticas: Idea básica

- Este método trata de inducir correlación negativa entre series de números pseudo-aleatorios.
- La idea básica es generar pares de corridas de un modelo tal que una observación pequeña en la primera corrida tiende a compensarse por una observación grande en la otra corrida, para que se obtenga una correlación negativa.
- Por ejemplo, si $u_k \sim \mathcal{U}(0,1)$ fue generado para obtener un parámetro particular de la primera corrida (por ejemplo, un tiempo de espera) entonces $1-u_k$ se usa para obtener el mismo parámetro en la segunda corrida. En este caso se dice que u_k y $1-u_k$ están *sincronizados* en el sentido de que fueron utilizados para el mismo propósito (en este caso generar el mismo parámetro del modelo).

Fundamento matemático I

- Supóngase que $(X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)})$ y $(X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)})$ son 2 corridas de una simulación, en donde $X_j^{(1)}$ se estimó con u y $X_j^{(2)}$ se estimó con $1-u$. Noten que:
  - $E(X_j^{(1)}) = E(X_j^{(2)}) = \mu$
  - Se tienen pares independientes: $(X_{j_1}^{(1)}, X_{j_1}^{(2)}) \perp\!\!\!\perp (X_{j_2}^{(1)}, X_{j_2}^{(2)})$.
  - Definan $X_j = \frac{X_j^{(1)} + X_j^{(2)}}{2}$ y $\bar{X} = \sum_{j=1}^{n} X_j$.

  Entonces:
  - $\bar{X}$ es un estimador insesgado de μ, y
  - $\mathsf{Var}(\bar{X}) = \mathsf{Var}(X_1)/n = \frac{\mathsf{Var}(X_j^{(1)}) + \mathsf{Var}(X_j^{(2)}) + 2\mathsf{Cov}(X_j^{(1)}, X_j^{(2)})}{4n}$
- De esta forma, si se induce correlación negativa entre $X_j^{(1)}$ y $X_j^{(2)}$, entonces se puede reducir la varianza del estimador $\bar{X}$.
- Sin embargo, *no siempre* se puede garantizar que se logre el objetivo, depende del modelo. A veces se puede elaborar un estudio piloto para medir la magnitud de la reducción.

Ejemplo

Si queremos estimar $\theta = \int_a^b f(x)dx$ por el método de Montecarlo crudo, entonces

$$\hat{\theta} = \frac{(b-a)}{n} \sum_{i=1}^{n} f(x_i),$$

donde $x_i \sim \mathcal{U}(a,b)$. Si por cada x_i se usa su variable antitética $\tilde{x}_i = a + (b - x_i)$, entonces el estimador se convierte en

$$\hat{\theta} = \frac{(b-a)}{n} \sum_{i=1}^{n/2} (f(x_i) + f(\tilde{x}_i))$$

Como probamos antes, como la varianza de la suma es la suma de las covarianzas mas dos veces la covarianza y la covarianza es negativa para variables antitéticas, entonces se reduce la varianza de la suma.

Variadas de control: Idea básica I

- Supóngase que queremos estimar $\theta = \mathsf{E}(X)$ donde X es una variable aleatoria de salida, como indicamos previamente.
- En lugar de estimar μ directamente, se considera la diferencia entre el problema de interés y un modelo analítico: Y es otra variable relacionada con θ, y que está correlacionada con X, pero además se conoce $\nu = E(Y)$. A Y se le llama variada de control para X.
- Sea $X_c = X - a(Y - \nu)$ una nueva variable. Entonces
  - $E(X_c) = E(X) = \theta$, por lo que X_c es un estimador insesgado de θ.
  - $\mathsf{Var}(X_c) = \mathsf{Var}(X - a(Y - \nu)) = \mathsf{Var}(X) + a^2\mathsf{Var}(Y) - 2a\mathsf{Cov}(X, Y)$ Entonces:

$$\mathsf{Var}(X_c) \leq \mathsf{Var}(X) \text{ si } 2a\mathsf{Cov}(X, Y) > a^2\mathsf{Var}(Y).$$

La varianza mínima se alcanza si

$$a^* = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Var}(Y)}$$

En este caso $\mathsf{Var}(X_c) = (1 - \rho^2_{X,Y})\mathsf{Var}(Y)$.

Variadas de control: Idea básica II

- En la práctica, se puede conocer o no el valor de Var(Y) y muy difícilmente conocemos Cov(X, Y), por lo que es difícil conocer el valor de a.
- Una alternativa es estimar a a través de un estudio piloto (Lavenberg, Moeller y Welch, 1982):

$$\hat{a} = \frac{\hat{C_{X,Y}}}{S_Y^2}$$

  y entonces $\bar{X}_c^* = \bar{X} - \hat{a}(\bar{Y} - \nu)$. Noten que $\bar{X}_c^*$ ya no es un estimador insesgado de θ. Para reducir el sesgo se puede utilizar, por ejemplo el *jacknife*.
- Se puede mostrar (ejercicio) que el método de variadas antitéticas es un caso particular del método de variadas de control.

Ejemplo (a)

Supongan que $X \sim \mathcal{N}(0,1)$ y se requiere estimar $E(\frac{X^6}{1+X^2})$. Como

$$\frac{x^6}{1+x^2} = x^4 - x^2 + 1 - \frac{1}{1+x^2}$$

entonces podemos aproximar $\frac{x^6}{1+x^2}$ con $Y = g(x) = x^4 - x^2 + 1$. Para esta Y, $E(Y) = E(X^4) - E(X^2) + 1 = 3 - 1 + 1 = 3$, ya que en una normal estándar $E(X^4) = 3$ (curtosis).
De este modo,

$$\theta = E(\frac{X^6}{1+X^2} - (X^4 - X^2 + 1)) + 3 = 3 - E(\frac{1}{1+X^2})$$

Así que podemos aplicar Montecarlo crudo sólo a la función $h(x) = \frac{1}{1+x^2}$ muestreando de una normal estándar.

Ejemplo (b)

Se desea estimar $\theta = E(e^{(U+W)^2})$ donde $U, W \sim \mathcal{U}(0,1)$ y son independientes. Sea $X = e^{(U+W)^2}$. Elegimos una variable de control Y.
Una posible variable de control es $Y_1 = U + W$. Por la distribución de U y W, sabemos que $\nu_1 = E(Y_1) = 1$ y se puede ver que $\text{Cov}(X, Y_1) > 0$. Otra posibilidad es usar $Y_2 = (U+W)^2$, y $E(Y_2) = 7/6$.
El siguiente código muestra como se puede hacer un pequeño piloto en R.

Código R

```r
#Primero hacemos un pequeño piloto
p <- 100;n <- 1000
u <- runif(p);w <- runif(p)
x <- exp((u+w)^2)
y <- (u+w)^2
covest <- cov(cbind(x,y))
a <- -covest[1,2]/covest[2,2]

# Ahora hacemos la simulación
u <- runif(n);w <- runif(n)
x <- exp((u+w)^2); y <- (u+w)^2
v <- x + a*(y-7/6)
estimadorusual <- c(mean(x),sd(x))
estimadorcont <- c(mean(v),sd(v))
CI <- c(mean(v)-1.96*sd(v)/sqrt(n),mean(v)+1.96*sd(v)/sqrt(n))
estimadorusual

[1] 4.811075 5.590200

estimadorcont

[1] 4.826285 2.661287

CI

[1] 4.661336 4.991233
```

Condicionamiento: idea básica I

- Supongamos que se quiere resolver $\theta = \mathsf{E}(h(X))$. En algunos casos es posible evaluar parte de la integral analíticamente, y entonces estimar el resto de la integral por simulación, reduciendo parte de la variabilidad. Esto se puede hacer a través de condicionamiento.

Ejemplo

Encontrar

$$\theta = \int_0^1 \int_0^1 f(x,y)\, dxdy$$

donde $f(x,y) = e^{g(x)y}$ y $g(x) = \sqrt{5/4 + \cos(2\pi x)}$. Para x fija, se puede resolver la integral con respecto a y, para obtener que

$$\theta = \int_0^1 \frac{e^{g(x)} - 1}{g(x)}\, dx$$

que ya es un problema unidimensional. Entonces, se podría condicionar con respecto a X. Noten que para y fija, no es fácil resolver la integral.

Condicionamiento: idea básica II

- Si Z es otra variable aleatoria tal que se conoce analíticamente la esperanza condicional $E(X|Z=z)$, entonces

$$\theta = E(X) = E(E(X|Z)).$$

  Además,utilizando la relación entre la varianza y la varianza condicional, $\text{Var}(X) = \text{E}(\text{Var}(X|Z)) + \text{Var}(\text{E}(X|Z))$, se tiene que:

$$\text{Var}(E(X|Z)) = \text{Var}(X) - E(\text{Var}(X|Z)) \leq \text{Var}(X).$$

- Entonces, conviene que Z tenga las siguientes propiedades:
  - X y Z tienen que ser dependientes para que el procedimiento tenga sentido.
  - Z puede ser generado de manera eficiente.
  - $\text{E}(X|Z=z)$ se puede calcular analíticamente
  - $\text{E}(\text{Var}(X|Z))$ tiene un valor grande.
- Este método se basa en el Toerema de Rao-Blackwell y por eso a veces se refiere a éste como *rao-balckwellization*. Aunque en realidad el teorema no aplica tal cual en el contexto de simulación porque no se cumplen necesariamente los supuestos.

Condicionamiento: idea básica III

El algoritmo es el siguiente:

Condicionamiento

1. Repetir n veces:
   - Generar Z_i
   - Calcular $V_i = g(Z_i) = \mathsf{E}(Y|Z_i)$
2. Calcular $\hat{\theta}_n = \bar{V}$
3. Calcular $\hat{\sigma}_n^2 = \frac{\sum_{i=1}^n (V_i - \bar{V})^2}{n-1}$
4. Un intervalo de confianza es de la forma $\hat{\theta}_n \pm z_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}}$

Ejemplo

Supongan un modelo jerárquico de la siguiente forma: $W \sim \mathcal{P}(10)$, y $X|W \sim Beta(w, w^2+1)$. El problema es encontrar $\theta = E(X)$.
Como sabemos que $E(X|W=w) = \frac{w}{w^2+w+1}$, entonces basta con muestrar $W_1, W_2, \ldots, W_n$ y construir

$$\tilde{\theta} = \frac{1}{n}\sum_{j=1}^{n} E(X|W=w_j) = \frac{1}{n}\sum_{j=1}^{n} \frac{w_j}{w_j^2+w_j+1}.$$

Noten que en este ejemplo, no tuvimos que generar ningún valor de X, sólo valores de una distribución Poisson conocida.
En este caso es muy ineficiente utilizar el método crudo de Montecarlo, porque obliga a muestrar de una distribución difícil de calcular.

Importance sampling

- Las regiones del espacio muestral son ponderadas de acuerdo a su contribución a la estimación de $\theta = E(X)$. Esta ponderación se hace a través de una densidad.
- Si $p(x)$ es la densidad ponderadora o función de importancia, entonces

$$\theta = \int_D f(x)dx = \int_D \frac{f(x)}{p(x)} p(x)dx$$

Entonces, un estimador de Montecarlo es $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \frac{f(x_i)}{p(x_i)}$ sobre $x_i \sim p$. La varianza de este estimador está dada por:

$$\text{Var}(\hat{\theta}) = \frac{1}{m}\text{Var}\left(\frac{f(x)}{p(x)}\right) = \frac{1}{m}\left[E\left(\frac{f^2(x)}{p^2(x)}\right) - E^2\left(\frac{f(x)}{p(x)}\right)\right].$$

Importance sampling

- El objetivo de importance sampling es escoger una densidad p tal que la varianza se minimice.
- Como $E^2\left(\frac{f(x)}{p(x)}\right) = \left(\int_D f(x)dx\right)^2$ es lo que queremos estimar y no depende de p, la elección sólo depende de $E\left(\frac{f^2(x)}{p^2(x)}\right)$.
- Por la desigualdad de Jensen:

$$E\left(\frac{f^2(x)}{p^2(x)}\right) \geq \left(E\left(\frac{|f(x)|}{p(x)}\right)\right)^2 = \left(\int_D |f(x)|dx\right)^2$$

- La cota se alcanza cuando $p(x) = \frac{|f(x)|}{\int_D |f(x)|dx}$ pero otra vez, no conocemos esa integral, así que en la práctica se busca $p(x)$ tal que $\frac{|f(x)|}{p(x)}$ sea aproximadamente constante, es decir, que $p(x)$ se parezca mucho a $f(x)$.
- Este método es muy parecido al método de aceptación y rechazo.

Importance sampling

$$\theta = \mathsf{E}_f[h(X)] = \int_\chi h(x)f(x)\,dx = \int h(x)\frac{f(x)}{p(x)}p(x)\,dx = E_p\left[\frac{h(X)f(X)}{p(X)}\right]$$

donde $supp(p) \supset supp(hf)$ (g es estrictamente positiva cuando hf es diferente de cero). El estimador IS de θ será

$$\hat{\theta}_{n,IS} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i)h(y_i)}{p(y_i)},$$

donde $y_1, \ldots, y_n \sim p$

Entonces, el problema se cambia a estimar observaciones de una densidad g en lugar de la densidad f.

Ejemplo IS I

Problema : Estimar $\theta = P(X > 20)$ donde $X \sim \mathcal{N}(0, 1)$.

```
pnorm(20,lower.tail=F)
```

```
[1] 2.753624e-89
```

En este caso pueden comprobar que generando observaciones de la distribución normal estándar no funciona, ya que el evento es muy raro. Pero podemos escribir la integral trasladando el problema a la cola de la distribución de interés:

$$
\begin{aligned}
\theta = \mathsf{E}[I(X > 20)] &= \int_{-\infty}^{\infty} I(X > 20) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\
&= \int_{-\infty}^{\infty} I(X > 20) \frac{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}{\frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}\, dx \\
&= \int_{-\infty}^{\infty} I(X > 20) e^{-\mu x + \mu^2/2} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}\, dx \\
&= \mathsf{E}_{\mu}\left[I(X > 20) e^{-\mu x + \mu^2/2} \right]
\end{aligned}
$$

Ejemplo IS II

Ahora podemos muestrear de $\mathcal{N}(\mu, 1)$. Podemos escoger $\mu = 20$ para estar cerca del punto de interés.

```
#Usando importance sampling:
n <- 1e6
mu <- 20
y <- rnorm(n, mean = mu)
I <- rep(0, n)
I[which(y > 20)] <- 1
theta_is <- mean(I*exp(-mu*y+mu^2/2))
#intervalo de confianza: noten la precisión.
theta_is

[1] 2.748108e-89

theta_is + c(-1,1)*sd(I*exp(-mu*y+mu^2/2))/n

[1] 2.748094e-89 2.748121e-89
```

Muestreo estratificado

- Este es un caso particular de importance sampling. Si suponemos que en el problema $\theta = \int_D f(x)dx\ f(x)$ se puede descomponer como una mezcla de densidades, tal que $f(x) = \sum_{j=1}^k w_j f_j(x)$ con w_j pesos convexos, entonces si $\theta_j = \int_D f_j(x)dx$,

$$\theta = \sum_{j=1}^k w_j \theta_j.$$

- Se puede muestrear cada parte de manera separada, con n_j observaciones, lo que da $\hat{\theta}_j$ con $\text{Var}(\hat{\theta}_j) = \sigma_j^2/n_j$, donde σ_j es la varianza para el estimador $\hat{\theta}_j$. Combinando se obtiene

$$\text{Var}(\hat{\theta}) = \sum_{j=1}^k w_j^2 \frac{\sigma_j^2}{n_j}$$

- Se puede elegir n_j de manera óptima para $n = \sum_{j=1}^k n_j$, w_j y σ_j^2 fijas, obteniendo $n_j = \frac{w_j \sigma_j}{\sum_{j=1}^k w_j \sigma_j}$

Ejemplo

Consideren obtener una muestra de una *normal contaminada*:

$$X \sim \begin{cases} \mathcal{N}(0,1) & \text{con probabilidad } 1-\alpha \\ \mathcal{N}\left(0,\sigma^2\right) & \text{con probabilidad } \alpha \end{cases}$$

Si α es pequeña entonces en una muestra de tamaño n se tendrán una o dos observaciones del componente contaminado. Si M es el número de casos contaminados en una muestra de tamaño n, entonces $M \sim Bin(n, \alpha)$.
Entonces podemos estratificar por la variable M en los casos $M = 0, 1, \ldots, n$, con $w_j = \binom{n}{j}\alpha^j(1-\alpha)^{n-j}$.
Observación: usualmente la varianza de un estimador se incrementa conforme la contaminación aumenta, así que la n_j óptima dependerá de j así como de w_j.