Universidad Tecnológica de la Mixteca

Métodos de integración y simulación Monte Carlo en la teoría bayesiana

Tesis para obtener el título de:

Licenciado en Matemáticas Aplicadas

Presenta:

Adriana Zacarías Santiago

Directora de Tesis:

M. C. Norma Edith Alamilla López

Huajuapan de León, Oaxaca. Junio del 2006

Índice general

Dedicatoria

Dedico esta tesis a mis padres, Hermenegildo y Guillermina, quienes han creído en mí, me han brindado su confianza, comprensión y me han apoyado en todo momento de mi vida. Gracias por apoyarme para realizar mis estudios.

A Rafael, mi esposo, quien ha sabido entenderme y apoyarme en los momentos difíciles. Ha sabido brindarme todo su cariño y, por supuesto, su amor.

A Josué, mi hijo, quien ha venido a ser la fuente de fuerza más importante para impulsarme a lograr mis objetivos, y por todas las horas que me ha prestado para realizar tanto mi carrera como esta tesis.

Un agradecimiento especial a mi directora de tesis, la M. C. Norma Edith Alamilla López, por todo lo que me ha enseñado, por su paciencia y, claro, por la amistad que me ha brindado.

A mis maestros quienes han compartido conmigo su conocimiento en las diversas áreas de mi carrera y por su amistad.

A mis amigas, Juana, Leticia, Beatriz, Norma y Liliana, quienes han sabido brindarme su amistad y consejos en momentos difíciles de mi vida, pues con ellas he compartido mi andar en la licenciatura. A mi amiga Susana quien ha sabido escucharme y darme consejos de aliento.

Agradezco las observaciones realizadas a esta tesis al Doctor Luis Enrique Nieto Barajas, catedrático del Instituto Tecnológico Autónomo de México.

Agradezco a mis sinodales, los profesores, José del Carmen Jiménez Hernández, y la profesora Marcela Rivera Martínez, quienes participaron en la revisión de esta tesis.

Y por último agradezco a la Universidad Tecnológica de la Mixteca el haberme brindado la oportunidad de realizar mis estudios en la licenciatura en Matemáticas Aplicadas.

Capítulo 1

Introducción

La estadística se ocupa fundamentalmente del análisis de datos que presentan variabilidad con el fin de comprender el mecanismo que los genera, o para ayudar en un proceso específico de toma de decisiones. En cualquiera de los casos existe una componente de incertidumbre involucrada, por lo cual el estadístico se ocupa en reducir lo más posible esa componente, así como también en describirla en forma apropiada. Además, la probabilidad es el único lenguaje posible para describir una lógica que trata con todos los niveles de incertidumbre y no sólo con los extremos de verdad o falsedad.

Se supone que toda forma de incertidumbre debe describirse por medio de modelos de probabilidad. En el caso de la estadística bayesiana se considera al parámetro, sobre el cuál se desea inferir, como un evento incierto. Entonces, como nuestro conocimiento no es preciso y esta sujeto a incertidumbre, se puede describir mediante una distribución de probabilidad, lo que hace que el parámetro tenga el carácter de aleatorio.

Las situaciones de toma de decisiones se abordan de manera natural en este enfoque a través de la introducción de una función de utilidad, la cual esta basada en una medida de probabilidad, y se maximiza la utilidad esperada.

En la estadística bayesiana se utiliza el teorema de Bayes (publicación póstuma de Thomas Bayes en 1763), combinando la información a priori con la información de los datos (función de verosimilitud), produciendo una descripción conjunta de la incertidumbre sobre los valores de los parámetros de la función de verosimilitud a través de la distribución a posteriori. Y de ésta es de donde se derivan los estimadores (cuya interpretación difiere radicalmente de la interpretación proporcionada por la inferencia clásica).

Esta teoría está firmemente basada en fundamentos matemáticos y una metodología coherente con la que es posible incorporar información relevante.

La estadística bayesiana constituye una alternativa a la estadística clásica, en

la resolución de los problemas típicos estadísticos que son: estimación, contraste de hipótesis y predicción. Además, los métodos bayesianos pueden ser aplicados a problemas que han sido inaccesibles a la teoría frecuentista clásica.

Ejemplo 1.1. Dada la interpretación clásica de la probabilidad, si un experimento aleatorio puede producir N resultados igualmente verosímiles y mutuamente excluyentes y si N_A de estos resultados tienen una característica A entonces

$$P(A) = \frac{N_A}{N}$$

pero ¿qué pasa si N no es finito? Digamos que el espacio muestral sea el conjunto de los números naturales y A el conjunto de los números naturales pares. ¿Qué probabilidad se le debe asignar a cada uno de los números pares si todos deben tener la misma probabilidad?

En una prueba de hipótesis puede haber diferencia entre el verdadero valor del parámetro y la hipótesis nula. De la misma forma una diferencia que no es estadísticamente significativa puede sin embargo ser importante, como se observa en el siguiente ejemplo.

Ejemplo 1.2. ([2]) El efecto de una droga es medido por $X \sim N(\theta, 9)$. La hipótesis nula es que $\theta \leq 0$. Una muestra de 9 observaciones resulta que $\bar{x} = 1$. Esto no es significativo para una muestra de una cola (se dice que un valor observado del estadístico de prueba es significativo si la hipótesis nula es rechazada al ni-vel especificado de significancia) con un nivel de significación de $\alpha = 0.05$. Y es significativo con un nivel de significación de $\alpha = 0.16$ (dado que $\alpha = 0.16$ es el valor p, que representa el nivel de significación más bajo al cual el valor observado del estadístico de prueba es significativo). Esto es moderadamente convincente, pues si en realidad 1 estuviera muy cercano a 0, se podría estar interesado en el medicamento. Pero la diferencia entre 0 y 1 es en realidad significativa. Así, si se tuviera que hacer una decisión basada únicamente en los datos, se podría decidir probablemente que el medicamento fue efectivo.

Las ventajas principales del enfoque bayesiano son su generalidad y coherencia, pues todos los problemas de inferencia se resuelven con los principios del cálculo de probabilidades y teoría de decisión. Además, posee una capacidad de incorporar información a priori adicional que se tenga del parámetro, a la muestra.

Esta última también es una desventaja, pues algunos investigadores rechazan que la información inicial se incluya en un proceso de inferencia científica. Pero esta situación se puede evitar estableciendo una distribución a priori no informativa o de referencia, la cual se introduce cuando no se posee mucha información previa acerca del problema. A un problema específico se le puede asignar cualquier tipo

de distribución a priori, ya que finalmente al actualizar la información a priori que se tenga acerca del parámetro, mediante el teorema de Bayes y obtener la distribución a posteriori del parámetro, es con esta con las que se hacen las inferencias del mismo.

El objetivo de esta tesis es revisar y analizar algunos métodos de integración Monte Carlo y algunos métodos de simulación Monte Carlo vía cadenas de Markov, mostrando para ello una serie de ejemplos para cada uno de los métodos aqui estudiados, estos ejemplos fueron realizados en el lenguaje R. Así como también, se mostrará un ejemplo con datos reales y se implementarán algunos de los métodos estudiados.

Los métodos de integración y simulación Monte Carlo surgen porque cuando se realiza el cálculo de la distribución a posteriori muchas veces éste resulta muy costoso, ya que regularmente esta distribución no tiene una forma cerrada. El costo se centra en el cálculo de ciertas integrales que no pueden resolverse analíticamente y por lo tanto se necesita contar con métodos eficientes que permitan resolver integrales en varias dimensiones. Entonces, gracias al avance computacional se pueden utilizar diversos algoritmos estadísticos para aproximar las integrales que surgen del análisis bayesiano y así poder hacer inferencias del parámetro aún cuando no se conozca exactamente la distribución.

La estructura de la tesis es la siguiente.

En el capítulo 2 se presentan los conceptos básicos de la estadística bayesiana, los cuales se necesitan para abordar los temas que se presentan a lo largo de la tesis.

En el capítulo 3 se presentan métodos asintóticos para obtener aproximaciones analíticas de la distribución a posteriori con el propósito de utilizarla en los siguientes capítulos.

En el capítulo 4 se revisarán algunos métodos de integración Monte Carlo para aproximar algunos valores de interés, tal como el valor esperado, la varianza, etc. Dentro de los métodos directos de integración Monte Carlo se analizarán el método directo y el método de composición, dentro de los métodos indirectos de integración Monte Carlo se estudiarán los métodos de muestreo por importancia y muestreo de aceptación-rechazo.

En el capítulo 5, se analizarán algunos de los métodos de simulación Monte Carlo vía cadenas de Markov para aproximar las diferentes distribuciones a posteriori. Los métodos que se abordarán en este capítulo son los métodos de Metropolis-Hastings y el muestreo de Gibbs, los cuales son los de mayor resonancia en la actividad estadística, así como en diversas áreas del conocimiento. En éste capítulo también se analizá la convergencia de estos métodos.

Los métodos de Monte Carlo se basan en la siguiente observación. La integral

$$\int h(\theta)\, d\theta = \int g(\theta)\, \pi(\theta)\, d\theta,$$

se puede escribir como el valor esperado de la función $g(\theta)$ respecto a la densidad de probabilidad $\pi(\theta)$, esto es,

$$E_{\pi}(g(\theta)) = \int h(\theta)\, d\theta.$$

Así, si se simula un número grande $\theta_1, \theta_2, \ldots, \theta_n$ de variables aleatorias de la densidad $\pi(\theta)$, entonces,

$$\int h(\theta) d\theta \simeq \frac{1}{n} \sum_{i=1}^{n} g(\theta_i).$$

Lo anterior se conoce como integración Monte Carlo.

Si se supone que la densidad $p(\theta) \propto \pi(\theta)$, entonces

$$\int g(\theta)\, \pi(\theta)\, d\theta = \int g(\theta) \left(\frac{\pi(\theta)}{p(\theta)}\right) p(\theta)\, d\theta = E_{p(\theta)} \left[g(\theta) \left(\frac{\pi(\theta)}{p(\theta)}\right)\right].$$

Esto forma la base para el método de muestreo por importancia, donde

$$\int g(\theta)\, \pi(\theta)\, d\theta \simeq \frac{1}{n} \sum_{i=1}^{n} g(\theta_i) \left(\frac{\pi(\theta_i)}{p(\theta_i)}\right),$$

y las θ_i se generan de la distribución dada por $p(\theta)$.

Por otro lado, el método de muestreo de aceptación-rechazo, es un método muy común y del que se han hecho numerosas versiones y combinaciones con otros métodos (Ripley, 1987 [20]; Devroye, 1986 [9]). En lugar de aproximar $\pi(\theta)$, se intenta simularla.

Recientemente, los métodos de simulación Monte Carlo mediante cadenas de Markov, permiten la resolución de problemas que hasta entonces no eran analíticamente tratables, y que precisaban distintas aproximaciones numéricas para las integrales implicadas.

Los métodos de Monte Carlo vía cadenas de Markov son útiles para simular muestras de distribuciones multivariadas. Se basan en la generación de una cadena de Markov cuya distribución de equilibrio es $g(\boldsymbol{\theta})$, donde $\boldsymbol{\theta}$ es multidimensional. Se corre esta cadena por un tiempo suficientemente largo; entonces los valores simulados pueden ser utilizados para estimar algunas características de $g(\boldsymbol{\theta})$ que sean de interés.

Estos métodos permiten muestrear la distribución a posteriori del parámetro de interés, siempre y cuando se conozca su forma analítica, salvo por una constante de normalización. Para este efecto se construye una cadena de Markov cuya distribución estacionaria es precisamente la distribución a posteriori.

En el capítulo 6 se presentan las conclusiones y algunos comentarios.

En el apéndice A se presenta un ejemplo real aplicando algunos de los métodos analizados, tomando datos reales de accidentes aéreos fatales de los países miembros de la Organización de Aviación Civil Internacional, los cuales fueron obtenidos de la publicación técnica No. 152 de la secretaría de comunicaciones y Transportes e Instituto Mexicano del Transporte [21]. Con estos datos se implementarán tanto los métodos directos de integración Monte Carlo, como los métodos de simulación Monte Carlo vía cadenas de Markov.

Capítulo 2

Conceptos básicos de la estadística bayesiana

La estadística bayesiana puede verse como un problema de decisión estadístico. Su estructura, en un ambiente de incertidumbre, está compuesto por un espacio $\mathcal{D}$ de acciones, una $\sigma-$álgebra ξ de subconjuntos de un conjunto Θ, y un conjunto $\mathcal{C}$ de consecuencias posibles [3].

El espacio $\mathcal{D}$ de acciones debe definirse de manera que sea exhaustivo y sus elementos mutuamente excluyentes. Los elementos de la $\sigma-$álgebra ξ son eventos relevantes al problema de decisión, y el conjunto $\mathcal{C}$ de consecuencias posibles describe las consecuencias de elegir una acción del conjunto $\mathcal{D}$, cuando ocurre un evento en ξ.

Para cuantificar las preferencias del estadístico entre las distintas consecuencias, se considera una función de utilidad, donde la solución óptima del problema de decisión consiste en elegir aquella acción en $\mathcal{D}$ que la maximice.

En la estadística bayesiana el parámetro de una cierta distribución tiene el carácter de aleatorio. La inferencia de los posibles valores del parámetro, se obtiene aplicando el cálculo de probabilidades (teorema de Bayes).

Estos cálculos se fundamentan en la información previa que posea el investigador, la cual puede cuantificarse mediante una distribución de probabilidad, denominada distribución a priori.

La información de los datos se cuantifican en **la función de verosimilitud**, la cual se define como:

$$f\left(\mathbf{x}|\boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(x_i|\boldsymbol{\theta}\right)$$

Considerando la distribución a priori y la distribución de los datos (función de verosimilitud) se aplica el teorema de Bayes, mediante el cual se obtiene la

distribución a posteriori del parámetro de interés, y de esta distribución es de donde se derivan los estimadores.

Las ventajas principales de la teoría bayesiana son su generalidad y coherencia, pues todos los problemas de inferencia se resuelven utilizando elementos de la teoría de decisión y los principios del cálculo de probabilidades.

La estadística bayesiana tiene una fundamentación axiomática en los llamados axiomas de coherencia, con el fin de tener un comportamiento coherente en el momento de tomar decisiones en un ambiente de incertidumbre. Estos axiomas son útiles en la construcción formal de las definiciones de probabilidad. Los axiomas de coherencia se describen a continuación [16].

2.1. Axiomas de coherencia

En la estadística, un investigador expresa sus preferencias entre las distintas opciones que se le presentan, entendemos por opción una situación en la que se obtiene la consecuencia $c_1 \in \mathcal{C}$ si sucede la acción $A_1 \in \mathcal{D}$, la consecuencia $c_2 \in \mathcal{C}$ si sucede la acción $A_2 \in \mathcal{D}$,..., la consecuencia $c_k \in \mathcal{C}$ si sucede la acción $A_k \in \mathcal{D}$. Esto se denota por

$$l = \{c_1 A_1, c_2 A_2, ..., c_k A_k\},$$

donde l denota el conjunto de opciones posibles.

Para poder expresar tales preferencias, se define la relación de orden entre las opciones, es decir,

i) $l_1 > l_2$ si se prefiere la opción l_1 a la opción l_2.

ii) $l_1 \geq l_2$ si la opción l_1 resulta más o igualmente preferible a la opción l_2 y

iii) $l_1 \sim l_2$ si las opciones l_1 y l_2 son igualmente preferibles.

De manera análoga se interpreta $l_1 < l_2$ y $l_1 \leq l_2$.

Así, se presentan los axiomas de coherencia que se utilizan para hacer una elección racional entre opciones alternativas.

Axioma 2.1. *(Comparabilidad) Para cada par de opciones l_1 y l_2, es cierta una y sólo una de las siguientes relaciones: $l_1 > l_2, l_1 < l_2$ o $l_1 \sim l_2$. Además existen dos consecuencias c^* y c_*, tales que $c_* < c^*$ y para toda $c \in \mathcal{C}$ sucede $c_* \leq c \leq c^*$ (donde c^* denota la consecuencia suprema y c_* denota la consecuencia ínfima).*

Axioma 2.2. *(Transitividad) Si $l_1 > l_2$ y $l_2 > l_3$, entonces $l_1 > l_3$. Análogamente, si $l_1 \sim l_2$ y $l_2 \sim l_3$, entonces $l_1 \sim l_3$ y si $l_1 < l_2$ y $l_2 < l_3$, entonces $l_1 < l_3$.*

Axioma 2.3. *(Sustitución y Dominancia) Si $l_1 > l_2$ cuando sucede A y $l_1 > l_2$ cuando sucede A^c, entonces $l_1 > l_2$. De igual manera, si $l_1 \sim l_2$ cuando sucede A y $l_1 \sim l_2$ cuando sucede A^c, entonces $l_1 \sim l_2$.*

Axioma 2.4. *(Suceso de referencia, experimento auxiliar) El decisor puede imaginar un procedimiento para generar un punto aleatorio z en el cuadrado unitario contenido en $\mathbb{R}^2$, esto es, un punto $z = (x, y)$ en $I = [0, 1] \times [0, 1]$ tal que para cualquier par de regiones R_1, R_2 de I el suceso $\{z \in R_1\}$ le resulta menos verosímil que el suceso $\{z \in R_2\}$ si y sólo si, el área de R_1 es menor que la de R_2.*

Axioma 2.5. *(Axioma de la medida precisa de preferencia): Para toda opción l_i existe una región R contenida en I tal que $l_i \sim l_R$.*

Estos axiomas dan sustento a la teoría de la decisión, y son una respuesta natural a la intención de tener un comportamiento coherente en el momento de tomar decisiones en un ambiente de incertidumbre, pues hay que cuantificar los eventos inciertos y las consecuencias, con el fin de determinar una decisión que resulte óptima para el investigador.

2.2. Probabilidad subjetiva

La teoría bayesiana se basa en la interpretación personal de la probabilidad (probabilidad subjetiva), es decir, la probabilidad que una persona asigna a uno de los posibles resultados de un proceso; representa su propio juicio sobre la verosimilitud del problema, el cual se basa en la opinión e información de la persona acerca del proceso.

Si los juicios de una persona sobre la verosimilitud relativa a ciertas combinaciones de resultados satisfacen ciertas condiciones de consistencia, se puede decir que sus probabilidades subjetivas se determinan de manera única.

Un elemento importante en la inferencia bayesiana es la información a priori que se tenga acerca del parámetro de interés, la cual no involucra información muestral. Así, se cuantifica el conocimiento inicial que el investigador tiene sobre el evento aleatorio que esta estudiando, en una distribución denotada por $\pi(\boldsymbol{\theta})$. Esta distribución recibe el nombre de distribución a priori.

2.3. Información a priori

Existen diversas maneras de cuantificar la información a priori, asignándole una distribución a priori al parámetro. Sobre la clasificación de las distribuciones a priori, se trata de dos tipos de clasificación: Por un lado están las conjugadas y no conjugadas, y por otro están las informativas y no informativas.

De estas posibilidades se analizarán las siguientes:

i) Distribución a priori conjugada.

ii) Distribución a priori no informativa.

2.3.1. Distribución a priori conjugada

En general, dada una función de verosimilitud $f(\mathbf{x} \mid \boldsymbol{\theta})$ y una distribución a priori $\pi(\boldsymbol{\theta})$, calcular integrales como $m(x) = \int f(\mathbf{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ no es fácil, por lo que se necesita escoger a $\pi(\boldsymbol{\theta})$ de tal manera que facilite la tratabilidad de la integral. Además, el teorema de Bayes se puede expresar como $\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ y así se puede garantizar que tanto $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ como $\pi(\boldsymbol{\theta})$ pertenecen a la misma familia general de funciones matemáticas, siempre y cuando se escoja a $\pi(\boldsymbol{\theta})$ con la misma estructura que $f(\mathbf{x} \mid \boldsymbol{\theta})$.

Es por ello, que se habla de una familia a priori que asegura la tratabilidad y fácil interpretación de esa integral, lo cual motiva a la siguiente definición.

Definición 2.1. *Sea la clase $\mathcal{F}$ de funciones de densidad $f(x \mid \boldsymbol{\theta})$, una clase $\wp$ de distribuciones a priori se dice que es* ***una familia conjugada*** *para $\mathcal{F}$ si $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ está en la clase de $\wp$, para toda $f \in \mathcal{F}$ y $\pi \in \wp$.*

Frecuentemente para una clase de densidades $\mathcal{F}$, una familia conjugada puede ser determinada examinando la función de verosimilitud $f(\mathbf{x} \mid \boldsymbol{\theta})$. Entonces la familia conjugada puede ser escogida como la clase de distribuciones con la misma estructura funcional. A esta clase de familia de conjugada se le conoce con el nombre de **distribución a priori conjugada natural.**

Ejemplo 2.1. ([2]) Sea $X = (X_1, ..., X_n)$ una muestra aleatoria independiente idénticamente distribuidas $(i.i.d.)$ con distribución de Poisson. Entonces

$$X_i \sim Pn(x_i \mid \theta),$$

es decir,

$$f(x_i \mid \theta) = \frac{\theta^{x_i} \exp(-\theta)}{x_i!},\ \theta > 0,$$

y

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \left[\frac{\theta^{x_i} \exp(-\theta)}{x_i!}\right] = \frac{\theta^{n\bar{\mathbf{x}}} \exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]} \propto \theta^{(n\bar{\mathbf{x}}+1)-1} \exp(-n\theta).$$

Aquí $\mathcal{F}$ es la clase de todas las densidades. Obsérvese que la función de verosimilitud de tales densidades se parece a una densidad gamma.

$$Ga(\theta \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta).$$

Dado que el dominio de la densidad gamma es $\theta > 0$, una posible familia de distribuciones a priori es la clase de distribuciones gamma. Se asume que $\theta \sim Ga(\alpha, \beta)$ y se observa que

$$\begin{aligned} \pi(\theta \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta)\pi(\theta) = \frac{\theta^{n\bar{x}}\exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \\ &= \frac{\beta^{\alpha}\theta^{(n\bar{x}+\alpha-1)}\exp(-(n+\beta)\theta)}{\Gamma(\alpha)\Pi_{i=1}^{n}[x_i!]} \propto \theta^{(n\bar{x}+\alpha)-1}\exp\{-(n+\beta)\theta\}. \end{aligned}$$

El factor que envuelve a θ en esta última expresión es claramente reconocida como una distribución $Ga\left(n\bar{x}+\alpha, n+\beta\right)$, la cual puede ser $\pi(\theta \mid \mathbf{x})$, entonces, la distribución a posteriori es una distribución gamma. Por lo tanto, se tiene que la clase de distribuciones gamma es verdaderamente una familia conjugada (natural) de $\mathcal{F}$.

2.3.2. Distribución a priori no informativa

Cuando no se dispone de información a priori, o ésta es mínima, se usa una distribución a priori no informativa, la cual nos da poca, o casi nula, información acerca del parámetro $\boldsymbol{\theta}$. Dentro de las no informativas hay tres tipos: la uniforme basada en el principio de la razón insuficiente, la de Jeffreys y la de referencia.

Por ejemplo, cuando en una prueba de hipótesis entre dos hipótesis simples la distribución a priori le asigna una probabilidad de $\frac{1}{2}$ a cada una de las dos hipótesis, tenemos una distribución a priori no informativa. El siguiente es un ejemplo más complejo.

Ejemplo 2.2. ([2]) Supongase que el parámetro de interés es la media de una distribución normal θ, entonces el espacio paramétrico es $\Theta = (-\infty, \infty)$. Si se desea una distribución a priori no informativa, es razonable dar igual peso a todos los posibles valores de θ. Desafortunadamente, si se elige a

$$\pi(\theta) = c > 0, \tag{2.1}$$

π tiene masa infinita (es decir, $\int \pi(\theta)\, d\theta = \infty$) y es una distribución impropia. No obstante con π se puede trabajar satisfactoriamente, pues la distribución a posteriori si es una distribución propia. La elección de c no es importante, pues típicamente en este tipo de problemas para distribuciones a priori no informativas se escoge a $\pi(\theta) = 1$. Esta distribución es llamada **la distribución uniforme** en $\mathbb{R}$, y fue introducida por Laplace en 1812.

En problemas más generales, se han hecho algunas propuestas para determinar distribuciones a priori no informativas. El método más usado es el de Jeffreys (1961), quien propuso

$$\pi\left(\boldsymbol{\theta}\right) \propto \left[I\left(\boldsymbol{\theta}\right)\right]^{\frac{1}{2}} \tag{2.2}$$

como una distribución a priori no informativa (llamada así, **distribución a priori no informativa de Jeffreys**), donde $I\left(\boldsymbol{\theta}\right)$ es la información esperada de Fisher, dada por:

$$I\left(\boldsymbol{\theta}\right) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log f\left(\mathbf{x} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^2}\right].$$

Ejemplo 2.3. Sea la función de densidad $X \mid \theta \sim N\left(x \mid \theta, \sigma^2\right)$ con σ conocido. Entonces, para calcular la distribución a priori no informativa por el método de Jeffreys se tiene que:

$$\begin{aligned}
\log f\left(x \mid \theta, \sigma^2\right) &= \log\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{(-\frac{1}{2})\left(\frac{x-\theta}{\sigma}\right)^2\right\}\right] \\
&= \log\left[\frac{1}{\sqrt{2\pi}\sigma}\right] + \log\left[\exp\left\{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2\right\}\right] \\
&= -\log\left[\sqrt{2\pi}\sigma\right] - \frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2.
\end{aligned}$$

Luego

$$\frac{\partial \log f\left(x \mid \theta, \sigma^2\right)}{\partial \theta} = \frac{1}{\sigma}\left(\frac{x-\theta}{\sigma}\right).$$

Calculando la segunda parcial

$$\frac{\partial^2 \log f\left(x \mid \theta, \sigma^2\right)}{\partial \theta^2} = -\frac{1}{\sigma^2},$$

así

$$I\left(\theta\right) = -E_{\theta}\left[\frac{\partial^2 \log f\left(x \mid \theta, \sigma^2\right)}{\partial \theta^2}\right] = -E_{\theta}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}.$$

Entonces la distribución a priori no informativa de Jeffreys para este ejemplo es:

$$\pi\left(\theta\right) \propto \left[I\left(\theta\right)\right]^{\frac{1}{2}} = \left[\frac{1}{\sigma^2}\right]^{\frac{1}{2}} = \frac{1}{\sigma}.$$

Otra distribución a priori no informativa de mucha importancia es l**a distribución a priori de referencia** $\left(\pi_i^R\left(\theta_i\right)\right)$ [1], la cual es una generalización de la distribución a priori de Jeffreys distinguiendo entre los parámetros de ruido y los de interés (ver Bernardo, 1979. [4]).

2.4. Distribución a posteriori

El análisis bayesiano combina la información a priori ($\pi(\boldsymbol{\theta})$) y la información muestral ($f(\mathbf{x} \mid \boldsymbol{\theta})$) en lo que se llama la distribución a posteriori de $\boldsymbol{\theta}$ dado $\mathbf{x}$, que es la base de todas las decisiones e inferencias acerca del parámetro.

Definición 2.2. *Dada la función de verosimilitud $f(\mathbf{x} \mid \boldsymbol{\theta})$ y una densidad a priori $\pi(\boldsymbol{\theta})$ se puede definir, mediante el teorema de Bayes, la distribución a posteriori de $\boldsymbol{\theta}$ dado $\mathbf{x}$. Es la distribución condicional de $\boldsymbol{\theta}$ dado $\mathbf{x}$ denotada por $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ y cuya expresión es:*

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\int f(\mathbf{x} \mid \boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}, \tag{2.3}$$

la cual da una visión completa de la creencia final de $\boldsymbol{\theta}$. Cabe hacer notar que $\boldsymbol{\theta}$ y $\mathbf{X}$ tienen densidad conjunta

$$h(\mathbf{x}, \boldsymbol{\theta}) = f(\mathbf{x} \mid \boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$$

y $\mathbf{X}$ tiene densidad marginal

$$m(\mathbf{x}) = \int f(\mathbf{x} \mid \boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int h(\mathbf{x}, \boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

Así, si $m(\mathbf{x}) \neq 0$, entonces

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{h(\mathbf{x}, \boldsymbol{\theta})}{m(\mathbf{x})}.$$

La distribución a posteriori combina la creencia a priori para $\boldsymbol{\theta}$, con la información de $\boldsymbol{\theta}$ contenida en una muestra $\mathbf{x}$, para dar una composición final de la creencia acerca de $\boldsymbol{\theta}$. Es decir, la información a priori refleja nuestra creencia de $\boldsymbol{\theta}$ antes de la experimentación, y $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ refleja nuestra creencia acerca de $\boldsymbol{\theta}$ después de la experimentación.

Ejemplo 2.4. ([2]) Supóngase que $X \sim N(x \mid \theta, \sigma^2)$, donde θ es desconocido, pero σ^2 es conocido. Sea $\theta \sim N(\theta \mid \mu, \tau^2)$ donde μ y τ^2 son conocidos. Entonces

$$\begin{aligned} h(x, \theta) &= \pi(\theta)\, f\left(x \mid \theta, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2}\left[\frac{\theta - \mu}{\tau}\right]^2\right\} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left[\frac{x - \theta}{\sigma}\right]^2\right\} \\ &= \frac{1}{2\pi\sigma\tau} \exp\left\{-\frac{1}{2}\left[\frac{(\theta - \mu)^2}{\tau^2} + \frac{(x - \theta)^2}{\sigma^2}\right]\right\} \end{aligned}$$

para encontrar $m(x)$, definimos

$$\rho = \frac{1}{\tau^2} + \frac{1}{\sigma^2} = \frac{\sigma^2 + \tau^2}{\tau^2\sigma^2}.$$

Y completando cuadrados tenemos:

$$\begin{aligned}
\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2} &= \frac{\sigma^2(\theta-\mu)^2 + \tau^2(x-\theta)^2}{\tau^2\sigma^2} \\
&= \frac{\sigma^2\left(\theta^2 - 2\theta\mu + \mu^2\right) + \tau^2\left(x^2 - 2x\theta + \theta^2\right)}{\tau^2\sigma^2} \\
&= \theta^2\rho - 2\theta\frac{(\sigma^2\mu + \tau^2 x)}{\tau^2\sigma^2} + \frac{(\sigma^2\mu^2 + \tau^2 x^2)}{\tau^2\sigma^2} \\
&= \theta^2\rho - 2\theta\frac{(\sigma^2+\tau^2)}{\tau^2\sigma^2}\frac{(\sigma^2\mu + \tau^2 x)}{\sigma^2+\tau^2} + \frac{(\sigma^2+\tau^2)}{\tau^2\sigma^2}\frac{(\sigma^2\mu^2 + \tau^2 x^2)}{\sigma^2+\tau^2} \\
&= \theta^2\rho - 2\theta\rho\frac{(\sigma^2\mu + \tau^2 x)}{\sigma^2+\tau^2} + \rho\frac{(\sigma^2\mu^2 + \tau^2 x^2)}{\sigma^2+\tau^2} \\
&= \rho\left[\theta^2 - 2\theta\frac{(\sigma^2\mu + \tau^2 x)}{\sigma^2+\tau^2} + \frac{(\sigma^2\mu + \tau^2 x)^2}{(\sigma^2+\tau^2)^2}\right] \\
&\quad + \rho\frac{(\sigma^2\mu^2 + \tau^2 x^2)}{\sigma^2+\tau^2} - \rho\frac{(\sigma^2\mu + \tau^2 x)^2}{(\sigma^2+\tau^2)^2} \\
&= \rho\left[\theta - \frac{(\sigma^2\mu + \tau^2 x)}{(\sigma^2+\tau^2)}\right]^2 + \frac{(\sigma^2+\tau^2)}{\tau^2\sigma^2}\left[\frac{\sigma^2\mu^2 + \tau^2 x^2}{\sigma^2+\tau^2}\right] \\
&\quad - \frac{(\sigma^2+\tau^2)}{\tau^2\sigma^2}\frac{(\sigma^2\mu + \tau^2 x)^2}{(\sigma^2+\tau^2)(\sigma^2+\tau^2)} \\
&= \rho\left[\theta - \frac{(\sigma^2\mu + \tau^2 x)}{(\sigma^2+\tau^2)}\right]^2 + \frac{\sigma^2\mu^2 + \tau^2 x^2}{\tau^2\sigma^2} - \frac{(\sigma^2\mu + \tau^2 x)^2}{\tau^2\sigma^2(\sigma^2+\tau^2)} \\
&= \rho\left[\theta - \frac{(\tau^2\sigma^2)}{(\sigma^2+\tau^2)}\frac{(\sigma^2\mu + \tau^2 x)}{(\tau^2\sigma^2)}\right]^2 + \\
&\quad \frac{(\sigma^2\mu^2 + \tau^2 x^2)(\sigma^2+\tau^2) - \sigma^4\mu^2 - 2\sigma^2\mu\tau^2 x - \tau^4 x^2}{\tau^2\sigma^2(\sigma^2+\tau^2)} \\
&= \rho\left[\theta - \frac{(\tau^2\sigma^2)}{(\sigma^2+\tau^2)}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2 \\
&\quad + \frac{\sigma^4\mu^2 + \sigma^2\tau^2 x^2 + \sigma^2\tau^2\mu^2 - \sigma^4\mu^2 - 2\sigma^2\mu\tau^2 x - \tau^4 x^2}{\tau^2\sigma^2(\sigma^2+\tau^2)}
\end{aligned}$$

$$
\begin{aligned}
&= \rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2+\frac{\tau^2\sigma^2x^2+\tau^2\sigma^2\mu^2-2\sigma^2\tau^2\mu x}{\tau^2\sigma^2(\sigma^2+\tau^2)}\\
&= \rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2+\frac{\tau^2\sigma^2(x^2-2\mu x+\mu^2)}{\tau^2\sigma^2(\sigma^2+\tau^2)}\\
&= \rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2+\frac{(\mu-x)^2}{(\sigma^2+\tau^2)}.
\end{aligned}
$$

De aquí que la densidad conjunta es:

$$
\begin{aligned}
h(x,\theta) &= \pi(\theta)\, f\left(x \mid \theta,\sigma^2\right)\\
&= \frac{1}{2\pi\sigma\tau}\exp\left\{-\frac{1}{2}\rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}\exp\left\{-\frac{(\mu-x)^2}{2\left(\sigma^2+\tau^2\right)}\right\},
\end{aligned}
$$

y la densidad marginal es:

$$
\begin{aligned}
m(x) &= \int_{-\infty}^{\infty} h(x,\theta)\,d\theta=\frac{1}{2\pi\tau\sigma}\int\exp\left\{-\frac{1}{2}\rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}\times\\
&\quad \exp\left\{-\frac{(\mu-x)^2}{2\left(\sigma^2+\tau^2\right)}\right\}d\theta\\
&= \frac{\rho^{-\frac{1}{2}}\sqrt{2\pi}}{2\pi\tau\sigma}\int\frac{1}{\rho^{-\frac{1}{2}}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}\times\\
&\quad \exp\left\{-\frac{(\mu-x)^2}{2\left(\sigma^2+\tau^2\right)}\right\}d\theta\\
&= \frac{\rho^{-\frac{1}{2}}\sqrt{2\pi}}{2\pi\tau\sigma}\exp\left\{-\frac{(\mu-x)^2}{2\left(\sigma^2+\tau^2\right)}\right\}=\frac{\sqrt{2\pi}}{\sqrt{2\pi}\sqrt{2\pi}\tau\sigma\sqrt{\rho}}\exp\left\{-\frac{(\mu-x)^2}{2\left(\sigma^2+\tau^2\right)}\right\}\\
&= \frac{1}{\sqrt{2\pi\rho}\tau\sigma}\exp\left\{-\frac{(\mu-x)^2}{2\left(\sigma^2+\tau^2\right)}\right\},
\end{aligned}
$$

es decir, es una distribución $N\left(\mu,\left(\sigma^2+\tau^2\right)\right)$.

Entonces la distribución a posteriori (dado que $m(x)\neq 0$) queda expresada como:

$$\begin{aligned}\pi(\theta|x,\sigma^2) &= \frac{h(x,\theta)}{m(x)} = \frac{\frac{1}{2\pi\sigma\tau}\exp\left\{-\frac{1}{2}\rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}\exp\left\{-\frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}\right\}}{\frac{1}{\sqrt{2\pi\rho}\tau\sigma}\exp\left\{-\frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}\right\}} \\ &= \frac{\sqrt{\rho}}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}\end{aligned}$$

es decir, es una distribución $N(\mu(x), \rho^{-1})$, donde

$$\mu(x) = \frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right).$$

es una función que depende de x.

Un ejemplo concreto es la situación cuando se les realiza una prueba de inteligencia a niños. Se asume que los resultados de la prueba siguen una distribución $N(\theta, 100)$, donde θ es el verdadero coeficiente intelectual de los niños medido por esta prueba. Supóngase que θ sigue una distribución $N(100, 225)$. Usando el resultado del ejemplo 2.4, se tiene que la distribución marginal de X es $N(\mu, (\sigma^2+\tau^2)) = N(100, 325)$. Entonces la distribución a posteriori de θ dado x sigue una distribución $N(\mu(x), \rho^{-1})$, donde

$$\rho = \frac{\sigma^2+\tau^2}{\tau^2\sigma^2} = \frac{100+225}{100(225)} = \frac{325}{22500} = \frac{13}{900}$$

entonces

$$\rho^{-1} = \frac{900}{13} = 69.23,$$

y

$$\mu(x) = \frac{900}{13}\left(\frac{100}{225}+\frac{x}{100}\right) = \frac{400+9x}{13}, \qquad \mu(115) = 110.38.$$

Entonces, si la calificación de un niño en la prueba es 115, su verdadero coeficiente intelectual θ tiene una distribución a posteriori $N(110.38, 69.23)$.

Ejemplo 2.5. ([10]) En la siguiente tabla aparece el número de accidentes fatales y muertos en accidentes aéreos regulares durante un año a lo largo de un período

de 10 años.

Años	Accidentes fatales	Pasajeros muertos
1976	24	734
1977	25	516
1978	31	754
1979	31	877
1980	22	814
1981	21	362
1982	26	764
1983	20	809
1984	16	223
1985	22	1066

Supóngase que el número de accidentes fatales cada año es independiente al de otros años, con una distribución de Poisson.

$$X_i \sim Pn\left(x_i \mid \theta\right),$$

entonces

$$f\left(x_i \mid \theta\right) = \exp\left(-\theta\right) \frac{\theta^{x_i}}{x_i!}, \quad \theta > 0.$$

Se considera una distribución a priori conjugada para θ, es decir,

$$\theta \sim Ga\left(\theta \mid \alpha, \beta\right),$$

entonces

$$\pi\left(\theta\right) = \frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)} \theta^{\alpha-1} \exp\left(-\beta\theta\right), \quad \theta > 0.$$

Por los resultados del ejemplo 2.1, la distribución a posteriori para θ dado x es:

$$\theta \mid x \sim Ga\left(\theta \mid n\overline{\mathbf{x}} + \alpha, n + \beta\right).$$

Otro ejemplo del cálculo de la distribución a posteriori es utilizando la distribución a priori no informativa de Jeffreys (2.2) obtenida en el ejemplo 2.3. Se le asigna esta distribución a priori cuando no se quiere introducir mucha subjetividad al problema que se esta estudiando. A este problema se le puede asignar cualquier tipo de distribución a priori, ya que al actualizar la información a priori que se tenga del parámetro, mediante el teorema de Bayes, se obtiene la distribución a posteriori del parámetro, y con esta distribución es con la que se hacer las inferencias acerca del parámetro.

Ejemplo 2.6. Sea $X \sim N\left(x \mid \theta, \sigma^2\right)$, donde θ es desconocido, pero σ^2 es conocido. Sea $\pi\left(\theta\right)$ la distribución a priori no informativa de Jeffreys obtenida en el ejemplo 2.3.

Entonces la densidad conjunta es

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

y la densidad marginal es

$$m\left(x\right) = \int_{-\infty}^{\infty} h\left(x, \theta\right) d\theta = \frac{1}{\sigma} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left[\frac{x-\theta}{\sigma}\right]^2\right\} = \frac{1}{\sigma}.$$

Por lo tanto la distribución a posteriori queda expresada como:

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

es decir,

$$\theta|x, \sigma^2 \sim N\left(\theta \mid x, \sigma^2\right)$$

2.5. Inferencia bayesiana

Los problemas de inferencia concernientes a $\boldsymbol{\theta}$ pueden ser fácilmente tratados usando el análisis bayesiano. La idea es que a partir de la distribución a posteriori, se supone que contamos con toda la información disponible acerca de $\boldsymbol{\theta}$ (la información a priori y muestral). Las inferencias acerca de $\boldsymbol{\theta}$ pueden considerarse sólo de las características de esta distribución.

2.5.1. Estimación

Estimación puntual

La estimación puntual es la inferencia más simple usada de la distribución a posteriori. Para estimar $\boldsymbol{\theta}$, varias técnicas clásicas pueden ser aplicadas a la distribución a posteriori. La técnica más común es la estimación de máxima verosimilitud, que escoge al estimador de $\boldsymbol{\theta}$ como el valor $\widehat{\boldsymbol{\theta}}$ que maximiza la función de verosimilitud. El análogo de estimación bayesiana de $\boldsymbol{\theta}$ se define como a continuación se describe.

Definición 2.3. *La probabilidad máxima generalizada estimada de* $\boldsymbol{\theta}$ *es la moda más grande,* $\widehat{\boldsymbol{\theta}}$*, de* $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ *(es decir, el valor* $\widehat{\boldsymbol{\theta}}$ *con el que se maximiza* $\pi(\boldsymbol{\theta} \mid \mathbf{x})$*, considerada como una función de* $\boldsymbol{\theta}$*).*

Ejemplo 2.7. (continuación del ejemplo 2.4) Sean la función de verosimilitud y la distribución a priori usadas en ese ejemplo, la densidad a posteriori obtenida fue $N(\mu(x), \rho^{-1})$. Dado que una densidad normal alcanza su valor máximo en la media, entonces la probabilidad máxima generalizada estimada de θ es:

$$\widehat{\theta} = \mu(x) = \frac{\sigma^2 \mu}{\sigma^2 + \tau^2} + \frac{\tau^2 x}{\sigma^2 + \tau^2}.$$

Ejemplo 2.8. ([2]) Supóngase que

$$f(x \mid \theta) = \exp(-(x - \theta)) I_{(\theta,\infty)}(x)$$

y

$$\pi(\theta) = \frac{1}{\pi \cdot \left(1 + \theta^2\right)},$$

entonces la distribución a posteriori queda como:

$$\begin{aligned} \pi(\theta \mid x) &= \frac{h(x,\theta)}{m(x)} = \frac{f(x \mid \theta)\pi(\theta)}{m(x)} \\ &= \frac{\exp(-(x-\theta)) I_{(\theta,\infty)}(x)}{m(x)\pi \cdot \left(1 + \theta^2\right)}. \end{aligned}$$

Para encontrar $\widehat{\theta}$ que maximiza la cantidad, nótese que:

i) si $x < \theta$, entonces $I_{(\theta,\infty)}(x) = 0$ y $\pi(\theta \mid x) = 0$.

ii) si $\theta \leq x$, tenemos que:

$$\begin{aligned} \frac{d}{d\theta}\pi(\theta \mid x) &= \left[\frac{\exp(-x)}{m(x)\pi}\right] \frac{d}{d\theta}\left[\frac{\exp\theta}{1+\theta^2}\right] \\ &= \left[\frac{\exp(-x)}{m(x)\pi}\right] \left[\frac{\exp\theta\left(1+\theta^2\right) - \exp\theta(2\theta)}{\left(1+\theta^2\right)^2}\right] \\ &= \left[\frac{\exp(-x)}{m(x)\pi}\right] \left[\frac{\exp\theta\left(1-2\theta+\theta^2\right)}{\left(1+\theta^2\right)^2}\right] \\ &= \left[\frac{\exp(-x)}{m(x)\pi}\right] \left[\frac{\exp\theta(\theta-1)^2}{\left(1+\theta^2\right)^2}\right]. \end{aligned}$$

Dado que esta derivada es siempre positiva, $\pi(\theta \mid x)$ se incrementa para $\theta \leq x$. De aquí $\pi(\theta \mid x)$ se maximiza cuando $\widehat{\theta} = x$, que es la probabilidad máxima generalizada de θ.

Otras estimaciones bayesianas de $\boldsymbol{\theta}$ incluyen la media y la mediana de $\pi\left(\boldsymbol{\theta} \mid \mathbf{x}\right)$. La media, mediana y moda son relativamente fáciles de calcular cuando las distribuciones a priori y a posteriori son de una familia conjugada de distribuciones.

Cuando se presenta una estimación estadística, usualmente se necesita indicar la precisión de la estimación. La medida bayesiana de la precisión de un estimador (en una dimensión) es el error cuadrático medio a posteriori de la estimación, el cual se define a continuación.

Definición 2.4. *El error cuadrático medio (ECM) de un estimador $\widehat{\delta}$ del parámetro θ es la función definida por:*

$$E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\theta\right)^2\right].$$

Para propósitos de cálculo, es útil notar que

$$\begin{aligned}
E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\theta\right)^2\right] &= E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\mu_{\pi(\theta|\mathbf{x})}+\mu_{\pi(\theta|\mathbf{x})}-\theta\right)^2\right] \\
&= E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\mu_{\pi(\theta|\mathbf{x})}\right)^2\right] \\
&\quad +2E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\mu_{\pi(\theta|\mathbf{x})}\right)\left(\mu_{\pi(\theta|\mathbf{x})}-\theta\right)\right] \\
&\quad +E_{\pi(\theta|\mathbf{x})}\left[\left(\mu_{\pi(\theta|\mathbf{x})}-\theta\right)^2\right] \\
&= \left(\widehat{\delta}-\mu_{\pi(\theta|\mathbf{x})}\right)^2+E_{\pi(\theta|\mathbf{x})}\left[\left(\theta-\mu_{\pi(\theta|\mathbf{x})}\right)^2\right] \\
&= \left(\widehat{\delta}-\mu_{\pi(\theta|\mathbf{x})}\right)^2+Var_{\pi(\theta|\mathbf{x})}. \qquad (2.4)
\end{aligned}$$

Obsérvese de (2.4) que la media a posteriori, $\mu_{\pi(\theta|\mathbf{x})}$, minimiza $E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\theta\right)^2\right]$ (sobre todo $\widehat{\delta}$), y de aquí $\widehat{\delta}$ es el estimador con menor error estándar. También tenemos que $\sqrt{E_{\pi(\theta|\mathbf{x})}\left[\left(\widehat{\delta}-\theta\right)^2\right]}$ es el error estándar.

Ejemplo 2.9. (continuación del ejemplo 2.4) Es claro que la varianza a posteriori es:

$$Var_{\pi(\theta|x)}=\rho^{-1}=\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Así, en el ejemplo de la prueba de inteligencia, el niño con $x=115$ puede reportar que tiene una estimación del coeficiente intelectual de $\mu_{\pi(\theta|x)}\left(115\right)=110.39$. Note

que, la estimación $\widehat{\delta} = x = 115$ puede tener un error estándar (con respecto a $\pi(\theta|x)$) de (usando 2.4):

$$\begin{aligned}\sqrt{E_{\pi(\theta|x)}\left[\left(\widehat{\delta}-\theta\right)^2\right]} &= \sqrt{Var_{\pi(\theta|x)}+\left(\mu_{\pi(\theta|x)}-\widehat{\delta}\right)^2} \\ &= \left[69.23+(110.39-115)^2\right]^{\frac{1}{2}} = \sqrt{90.48} = 9.49.\end{aligned}$$

La estimación bayesiana de un vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^t$ es más sencilla. La probabilidad máxima generalizada estimada (la moda a posteriori) es un estimador razonable, aunque hay problemas de existencia y unicidad que son más probables de encontrar en el caso multivariado. La media a posteriori

$$\mu_{\pi(\boldsymbol{\theta}|\mathbf{x})} = \left(\mu^1_{\pi(\theta_1|\mathbf{x})}, ..., \mu^p_{\pi(\theta_p|\mathbf{x})}\right)^t = E_{\pi(\boldsymbol{\theta}|\mathbf{x})}[\boldsymbol{\theta}]$$

es un estimador bayesiano y su precisión puede ser descrita por la matriz de covarianzas a posteriori

$$V_{\pi(\boldsymbol{\theta}|\mathbf{x})} = E_{\pi(\boldsymbol{\theta}|\mathbf{x})}\left[\left(\boldsymbol{\theta}-\mu_{\pi(\boldsymbol{\theta}|\mathbf{x})}\right)\left(\boldsymbol{\theta}-\mu_{\pi(\boldsymbol{\theta}|\mathbf{x})}\right)^t\right].$$

El error estándar del estimador $\mu^i_{\pi(\theta_i|\mathbf{x})}$ de θ_i puede ser $\sqrt{V^{ii}_{\pi(\theta_i|\mathbf{x})}}$ donde $V^{ii}_{\pi(\theta_i|\mathbf{x})}$ es el elemento (i,i) de $V_{\pi(\boldsymbol{\theta}|\mathbf{x})}$.

El análogo de (2.4), para la estimación general $\widehat{\boldsymbol{\delta}}$ de $\boldsymbol{\theta}$ es

$$\begin{aligned}E\left[\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\theta}\right)^2\right] &= E\left[\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\theta}\right)^t\right] \\ &= V_{\pi(\boldsymbol{\theta}|\mathbf{x})}+\left(\mu_{\pi(\boldsymbol{\theta}|\mathbf{x})}-\widehat{\boldsymbol{\delta}}\right)\left(\mu_{\pi(\boldsymbol{\theta}|\mathbf{x})}-\widehat{\boldsymbol{\delta}}\right)^t\end{aligned}$$

por lo que la media a posteriori minimiza $E\left[\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\theta}\right)^2\right]$.

Otra forma de inferir es mediante un intervalo de confianza para $\boldsymbol{\theta}$. El análogo bayesiano de un intervalo de confianza es llamado conjunto creíble, que se define a continuación:

Conjuntos creíbles

Definición 2.5. *Un conjunto creíble de* $100(1-\alpha)$ % *para* $\boldsymbol{\theta}$ *es un subconjunto* C *de* Θ *tal que*

$$1-\alpha \leq P\left(C \mid \mathbf{x}\right)=\int_{C} dF^{\pi(\boldsymbol{\theta}|\mathbf{x})}\left(\theta\right)=\begin{cases} \int_{C} \pi\left(\boldsymbol{\theta} \mid \mathbf{x}\right) d\boldsymbol{\theta} & \textit{caso continuo} \\ \sum\limits_{\boldsymbol{\theta}\in C} \pi\left(\boldsymbol{\theta} \mid \mathbf{x}\right) & \textit{caso discreto} \end{cases}$$

Puesto que la distribución a posteriori es una distribución de probabilidad de $\boldsymbol{\theta}$, uno puede hablar del significado (usando subjetividad) de la probabilidad de que $\boldsymbol{\theta}$ esté en C. Esto está en contraste con los procedimientos de confianza clásicos que son interpretados en términos de probabilidad de cobertura (la probabilidad de que la variable aleatoria $\mathbf{X}$ sea tal que el conjunto de confianza $C(\mathbf{X})$ contenga a $\boldsymbol{\theta}$).

Al elegir un conjunto creíble para $\boldsymbol{\theta}$, se desea intentar minimizar su tamaño. Para hacer esto, se pueden incluir en el conjunto solo aquellos puntos con la densidad a posteriori más grande, es decir, los valores "mas probables" de $\boldsymbol{\theta}$.

Definición 2.6. *El conjunto creíble de máxima densidad a posteriori (MDP) a un nivel de confianza de* $100(1-\alpha)\,\%$*, es el subconjunto* C *de* Θ*, de la forma*

$$C=\left\{\boldsymbol{\theta}\in\Theta : \pi\left(\boldsymbol{\theta} \mid \mathbf{x}\right) \geq K\left(\alpha\right)\right\},$$

donde $K\left(\alpha\right)$ *es la constante más grande tal que*

$$P\left(C \mid \mathbf{x}\right) \geq 1-\alpha.$$

Ejemplo 2.10. (continuación del ejemplo 2.4) Dado que la densidad a posteriori de θ dado x es $N\left(\mu\left(x\right), \rho^{-1}\right)=N\left(110.39, 69.23\right)$, que es unimodal y simétrica respecto a $\mu\left(x\right)$, el conjunto creíble de máxima densidad a un nivel de confianza de $100\left(1-\alpha\right)\,\%$ está dado por

$$C=\left(\mu\left(x\right)-z_{\frac{\alpha}{2}}\rho^{-\frac{1}{2}}, \mu\left(x\right)+z_{\frac{\alpha}{2}}\rho^{-\frac{1}{2}}\right),$$

donde z_{α} es el α-percentil de la distribución $N\left(0,1\right)$. Si queremos un nivel de significación del $95\,\%$, $\alpha=0.05$, entonces $z_{\frac{\alpha}{2}}=z_{0.025}=1.96$.

$$\begin{aligned} C &= \left(110.39-1.96\left(\sqrt{69.23}\right), 110.39+1.96\left(\sqrt{69.23}\right)\right) \\ &= \left(94.0819, 126.6981\right). \end{aligned}$$

Los conjuntos creíbles bayesianos son usualmente mucho más fáciles de calcular que los intervalos de confianza clásicos, principalmente en situaciones donde un estadístico suficiente no existe.

Cuando $\boldsymbol{\theta}$ es multivariado el uso de la aproximación normal para la distribución a posteriori es valioso, pues los cálculos se vuelven difíciles de otra manera.

2.5.2. Constraste de hipótesis

En una prueba de hipótesis clásica, la hipótesis nula $H_0 : \theta \in \Theta_0$ y la hipóteis alternativa $H_1 : \theta \in \Theta_1$ están especificadas. Un procedimiento de la prueba es evaluar en términos de probabilidades los errores de tipo I y tipo II. Estas probabilidades de los errores representan la posibilidad de que una muestra sea observada en el proceso de prueba y sea aceptada la hipótesis.

En el análisis bayesiano, decidir entre H_0 y H_1 es conceptualmente más fácil. Sólo se calculan las probabilidades a posteriori $\alpha_0 = P(\Theta_0|\mathbf{x})$ y $\alpha_1 = P(\Theta_1|\mathbf{x})$ y se decide entre H_0 y H_1 respectivamente. La ventaja conceptual es que α_0 y α_1 son probabilidades actuales (subjetivas) de la hipótesis en base a los datos y la opinión a priori.

Aunque las probabilidades a posteriori de las hipótesis son las medidas primarias bayesianas en problemas de hipótesis, los siguientes conceptos relacionados son de interés. Aquí usaremos π_0 y π_1 para denotar las probabilidades a priori de Θ_0 y Θ_1 respectivamente.

Definición 2.7. *La razón* $\frac{\alpha_0}{\alpha_1}$ *es llamada la razón a posteriori de probabilidades de* H_0 *contra* H_1, *y* $\frac{\pi_0}{\pi_1}$ *es llamada la razón a priori de probabilidades. La cantidad*

$$B = \frac{\textit{razón a posteriori de probabilidades}}{\textit{razón a priori de probabilidades}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

es llamada ***el factor de Bayes*** *a favor de* Θ_0.

El interés en el factor de Bayes es que algunas veces puede ser interpretado como "la razón de H_0 contra H_1 dada por los datos". Esta interpretación es claramente válida cuando las hipótesis son simples, esto es, cuando $\Theta_0 = \{\theta_0\}$ y $\Theta_1 = \{\theta_1\}$. Entonces

$$\alpha_0 = \frac{\pi_0 f(\mathbf{x}|\theta_0)}{\pi_0 f(\mathbf{x}|\theta_0)+\pi_1 f(\mathbf{x}|\theta_1)}, \quad \alpha_1 = \frac{\pi_1 f(\mathbf{x}|\theta_1)}{\pi_0 f(\mathbf{x}|\theta_0)+\pi_1 f(\mathbf{x}|\theta_1)}$$

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 f(\mathbf{x}|\theta_0)}{\pi_1 f(\mathbf{x}|\theta_1)} \qquad \text{y} \qquad B = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} = \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} \ .$$

En otras palabras, B es la razón de verosimilitud de H_0 contra H_1, que es comunmente vista como la probabilidad dada por los datos de H_0 contra H_1.

En general, B depende de la distribución a priori que se asigne. Para explicar esta dependencia, es conveniente escribir la distribución a priori como

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & si\ \theta \in \Theta_0 \\ \pi_1 g_1(\theta) & si\ \theta \in \Theta_1 \end{cases} ,$$

donde g_0 y g_1 son densidades propias que describen como la masa a priori se extiende sobre las dos hipótesis. Con esta representación se escribe

$$\frac{\alpha_0}{\alpha_1} = \frac{\int_{\Theta_0} dF^{\pi(\theta|\mathbf{x})}(\theta)}{\int_{\Theta_1} dF^{\pi(\theta|\mathbf{x})}(\theta)} = \frac{\int_{\Theta_0} f(\mathbf{x}|\theta)\,\pi_0 dF^{g_0}(\theta)/m(x)}{\int_{\Theta_1} f(\mathbf{x}|\theta)\,\pi_1 dF^{g_1}(\theta)/m(x)} = \frac{\pi_0 \int_{\Theta_0} f(\mathbf{x}|\theta)\,dF^{g_0}(\theta)}{\pi_1 \int_{\Theta_1} f(\mathbf{x}|\theta)\,dF^{g_1}(\theta)},$$

de aquí

$$B = \frac{\int_{\Theta_0} f(\mathbf{x}|\theta)\,dF^{g_0}(\theta)}{\int_{\Theta_1} f(\mathbf{x}|\theta)\,dF^{g_1}(\theta)}$$

que es la razón de verosimilitud (a inclinarse por g_0 y g_1) de Θ_0 contra Θ_1. Lo que implica que g_0 y g_1 no pueden ser interpretadas como una medida del soporte relativo para la hipótesis proporcionando solamente datos.

Sin embargo, B puede ser relativamente insesgado para elecciones razonables de g_0 y g_1, y entonces tal interpretación es razonable.

Ahora, se explicará en que consiste la prueba unilateral y la prueba de una hipótesis nula puntual, señalando los aspectos generales de las pruebas bayesianas y comparando los enfoques clásicos y bayesianos.

Prueba unilateral

Una prueba de hipótesis unilateral ocurre cuando

$$\begin{aligned} H_0 &: \theta \leq \theta_0, \\ H_1 &: \theta > \theta_0, \end{aligned}$$

ó

$$\begin{aligned} H_0 &: \theta \geq \theta_0, \\ H_1 &: \theta < \theta_0. \end{aligned}$$

Esta prueba no tiene una característica especial. El interés está en que en el contraste clásico el uso del $valor-p$, tiene una justificación bayesiana. Considérese el siguiente ejemplo.

Ejemplo 2.11. Cuando $X \sim N(\theta, \sigma^2)$ y θ tiene una distribución a priori no informativa (2.1) $\pi(\theta) = 1$. Se tiene

$$h(x,\theta) = \pi(\theta)\, f\left(x|\theta,\sigma^2\right) = f\left(x|\theta,\sigma^2\right)$$

y

$$m(x) = \int_{-\infty}^{\infty} h(x,\theta)\,d\theta = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) d\theta = 1,$$

entonces

$$\pi(\theta|x,\sigma^2) = \frac{h\,(x,\theta)}{m\,(x)} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right).$$

Es decir, la distribución a posteriori de θ dado x tiene una distribución $N\,(x,\sigma^2)$.

Considérese ahora la situación

$$\begin{aligned} H_0 &: \theta \leq \theta_0, \\ H_1 &: \theta > \theta_0. \end{aligned}$$

Entonces

$$\alpha_0 = P\,(\Theta_0|x) = P\,(\theta \leq \theta_0|x) = \Phi\left(\frac{\theta_0 - x}{\sigma}\right),$$

donde Φ es la función de distribución de una normal estándar.

El clásico $valor - p$ contra H_0 es la probabilidad, cuando $\theta = \theta_0$ de observar un X "más extremo" que el dato x actual. Aquí el $valor - p$ puede ser

$$valor - p = P\,(X \geq x) = 1 - \Phi\left(\frac{x-\theta_0}{\sigma}\right).$$

Por la simetría de la distribución normal, se sigue que α_0 es igual al $valor - p$ contra H_0.

Contraste de hipótesis nula puntual

Es muy común en la estadística clásica plantear hipótesis de la forma

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta \neq \theta_0. \end{aligned}$$

Tal prueba de hipótesis nula puntual es interesante, particularmente porque el enfoque bayesiano contiene algunas características originales, pero principalmente porque las respuestas bayesianas difieren radicalmente de las respuestas clásicas.

Se deben hacer algunos comentarios acerca de este tipo de hipótesis. Primero, son comúnmente realizadas en situaciones inapropiadas. En realidad, nunca se da el caso que se considere la posibilidad que $\theta = \theta_0$ exactamente. Más razonable sería la hipótesis nula $\theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$, donde $b > 0$ es alguna constante elegida tal que todo θ en Θ_0 pueda considerarse indistinguible de θ_0.

Hay algunos problemas de decisión que conducirán a intervalos de este tipo pero con b grande. Tales problemas raramente serán bien aproximados por una prueba de hipótesis nula puntual.

Dado que uno debe contrastar $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, necesitamos conocer cuando es adecuada la aproximación de H_0 por $H_0 : \theta = \theta_0$. Desde la perspectiva bayesiana la aproximación es razonable si la probabilidad a posteriori de H_0 está cerca de la igualdad en ambos contrastes. Otra condición es que la función de probabilidad observada sea aproximadamente constante en $(\theta_0 - b, \theta_0 + b)$.

Ejemplo 2.12. ([2]) Sea $X_1, ..., X_n$ una muestra observada con una distribución $N(\theta, \sigma^2)$, donde σ^2 es conocido. La función de verosimilitud observada es entonces proporcional a $N\left(\bar{x}, \frac{\sigma^2}{n}\right)$ de θ. Esto puede ser constante en $(\theta_0 - b, \theta_0 + b)$, donde b es pequeño comparado a $\frac{\sigma}{\sqrt{n}}$. Por ejemplo, cuando es una prueba clásica $z = \frac{\sqrt{n}|\bar{x}-\theta_0|}{\sigma}$ es más grande que 1, la función de verosimilitud varía no más de 5 % en $(\theta_0 - b, \theta_0 + b)$ si $b \leq (0.024)\, z^{-1} \frac{\sigma}{\sqrt{n}}$. Cuando $z = 2$, $\sigma = 1$ y $n = 25$, esto impone la cota $b \leq 0.0024$. Note que el límite en b puede depender de $|\bar{x} - \theta_0|$, más que de $\frac{\sigma}{\sqrt{n}}$.

Para realizar un contraste bayesiano de la hipótesis nula puntual $H_0 : \theta = \theta_0$, no se puede usar una distribución a priori continua. Pues la distribución a priori y la distribución a posteriori darán una probabilidad de cero en θ_0. Una aproximación razonable es dar a θ_0 una probabilidad positiva π_0, y a $\theta \neq \theta_0$ la densidad $\pi_1 g_1(\theta)$ donde $\pi_1 = 1 - \pi_0$ y g_1 es propia. Se puede pensar a π_0 como la masa que se le asignaría a la hipótesis $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$.

La densidad marginal de $\mathbf{X}$ es

$$m(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\, dF^{\pi}(\theta) = f(\mathbf{x} \mid \theta_0)\, \pi_0 + (1 - \pi_0)\, m_1(\mathbf{x}),$$

donde

$$m_1(\mathbf{x}) = \int_{\{\theta \neq \theta_0\}} f(\mathbf{x} \mid \theta)\, dF^{g_1}(\theta)$$

es la densidad marginal de $\mathbf{X}$ con respecto a g_1. De aquí la probabilidad a posteriori de $\theta = \theta_0$ es

$$\begin{aligned} \pi(\theta_0 \mid \mathbf{x}) &= \frac{f(\mathbf{x} \mid \theta_0)\, \pi_0}{m(\mathbf{x})} = \frac{f(\mathbf{x} \mid \theta_0)\, \pi_0}{f(\mathbf{x} \mid \theta_0)\, \pi_0 + (1 - \pi_0)\, m_1(\mathbf{x})} \\ &= \left[1 + \frac{1 - \pi_0}{\pi_0} \frac{m_1(\mathbf{x})}{f(\mathbf{x} \mid \theta_0)}\right]^{-1} \end{aligned}$$

Note que esto es α_0, y que $\alpha_1 = 1 - \alpha_0$ es la probabilidad a posteriori de H_1. Así la razón a posteriori de probabilidades es

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi(\theta_0 \mid \mathbf{x})}{1 - \pi(\theta_0 \mid \mathbf{x})} = \frac{\pi_0}{\pi_1} \frac{f(\mathbf{x} \mid \theta_0)}{m_1(\mathbf{x})}$$

y el factor de Bayes de H_0 contra H_1 es

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{f(\mathbf{x} \mid \theta_0)}{m_1(\mathbf{x})}$$

2.5.3. Inferencia predictiva

Cuando se intenta predecir una variable aleatoria $Z \sim g(z \mid \boldsymbol{\theta})$ basada en las observaciones de $\mathbf{X} \sim f(\mathbf{x} \mid \boldsymbol{\theta})$. Podemos asumir que $\mathbf{X}$ y Z son independientes y que g es una densidad.

La idea de la inferencia predictiva bayesiana es que, dado que $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ es la distribución de $\boldsymbol{\theta}$ (creencia a posteriori), entonces $g(z \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{x})$ es la distribución conjunta de z y $\boldsymbol{\theta}$ dado $\mathbf{x}$, e integrando sobre todo el espacio paramétrico $\boldsymbol{\theta}$ obtenemos la distribución de z dado $\mathbf{x}$.

Definición 2.8. *La densidad predictiva a posteriori de Z dado $\mathbf{x}$, cuando la distribución a priori de $\boldsymbol{\theta}$ es π, es definida por*

$$P(z \mid \mathbf{x}) = \int_{\Theta} g(z \mid \boldsymbol{\theta}) \, dF^{\pi(\boldsymbol{\theta} \mid \mathbf{x})}(\boldsymbol{\theta}) = \int_{\Theta} g(z \mid \boldsymbol{\theta}) \, \pi(\boldsymbol{\theta} \mid \mathbf{x}) \, d\boldsymbol{\theta}.$$

Ejemplo 2.13. Para calcular una distribución predictiva a posteriori usando una distribución a priori no informativa de Jeffreys (2.2) en el ejemplo de los accidentes fatales. Tenemos dado que

$$f(x_i \mid \theta) = \frac{\exp(-\theta) \, \theta^{x_i}}{x_i!},$$

tomando el logaritmo obtenemos

$$\log(f(x_i \mid \theta)) = -\theta + x_i \log(\theta) - \log(x_i!),$$

calculando la parcial con respecto de θ

$$\frac{\partial(\log(f(x_i \mid \theta)))}{\partial \theta} = -1 + \frac{x_i}{\theta}$$

y

$$\frac{\partial^2(\log(f(x_i \mid \theta)))}{\partial \theta^2} = -\frac{x_i}{\theta^2}.$$

La matriz de información de Fisher será

$$I(\theta) = -E\left[\frac{-x_i}{\theta^2}\right] = \frac{E(x_i)}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

y por lo tanto

$$\pi(\theta) \propto \theta^{-\frac{1}{2}}.$$

La distribución a posteriori se calcula como:

$$f(\mathbf{x} \mid \theta) = \frac{\theta^{n\overline{\mathbf{x}}} \exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]},$$

$$\begin{aligned} f(\mathbf{x} \mid \theta)\pi(\theta) &= \frac{\theta^{(n\overline{\mathbf{x}}+1)-1} \exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]}\theta^{-\frac{1}{2}} \\ &= \frac{\theta^{(n\overline{\mathbf{x}}+\frac{1}{2})-1} \exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]} \end{aligned}$$

donde la distribución marginal es

$$\begin{aligned} m(\mathbf{x}) &= \int f(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta \\ &= \frac{\Gamma\left(n\overline{\mathbf{x}} + \frac{1}{2}\right)}{\Pi_{i=1}^{n}[x_i!]\, n^{(n\overline{\mathbf{x}}+\frac{1}{2})}} \int \frac{n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}{\Gamma\left(n\overline{\mathbf{x}} + \frac{1}{2}\right)}\theta^{(n\overline{\mathbf{x}}+\frac{1}{2})-1} \exp(-n\theta)\, d\theta \\ &= \frac{\Gamma\left(n\overline{\mathbf{x}} + \frac{1}{2}\right)}{\Pi_{i=1}^{n}[x_i!]\, n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}. \end{aligned}$$

Así, la distribución a posteriori queda como

$$\begin{aligned} \pi(\theta \mid \mathbf{x}) &= \frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{m(x)} \\ &= \frac{n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}{\Gamma\left(n\overline{\mathbf{x}} + \frac{1}{2}\right)}\theta^{(n\overline{\mathbf{x}}+\frac{1}{2})-1} \exp(-n\theta) \end{aligned}$$

$$\theta \mid \mathbf{x} \sim Ga\left(\theta \mid \left(n\overline{\mathbf{x}} + \frac{1}{2}\right), n\right)$$

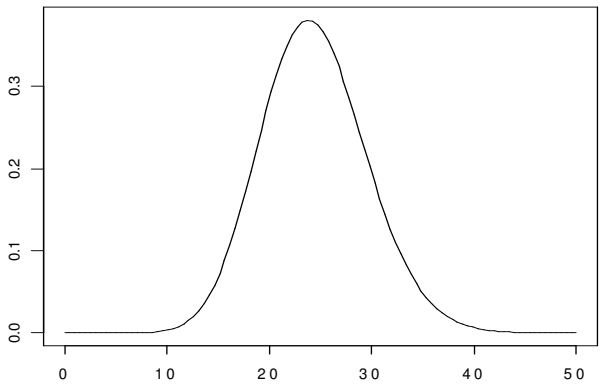

Figura 2.1: Distribución predictiva a posteriori del ejemplo 2.13.

Luego la distribución predictiva a posteriori es:

$$\begin{aligned} m\left(x^*|\mathbf{x}\right) &= \int f\left(x^*|\theta\right)\pi\left(\theta|\mathbf{x}\right)d\theta \\ &= \int \exp\left(-\theta\right)\frac{\theta^{x^*}}{x^*!}\left[\frac{(n)^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\theta^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)-1}\exp\left(-n\theta\right)\right]d\theta \\ &= \frac{(n)^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{x^*!\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\frac{\Gamma\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)}{(1+n)^{\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)}}\int\frac{(1+n)^{\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{\Gamma\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)} \\ &\quad \theta^{\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)-1}\exp\left(-\left(1+n\right)\right)d\theta \\ &= \frac{(n)^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{x^*!\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\frac{\Gamma\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)}{(1+n)^{\left(x^*+n\overline{\mathbf{x}}+\frac{1}{2}\right)}}, \end{aligned}$$

$$x^*|\mathbf{x} \sim Pg\left(x^* \mid \left(n\overline{\mathbf{x}}+\frac{1}{2}\right), n, 1\right),$$

donde

$$E\left[x^*\right] = \frac{n\overline{\mathbf{x}}+\frac{1}{2}}{n}, V\left[x^*\right] = \frac{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)(n+1)}{(n)^2}.$$

En la figura 2.1 se muestra la distribución predictiva a posteriori para este ejemplo (ver apéndice A2.1).

En la tabla 2.1 se presentan los cálculos de los momentos de la distribución predictiva a posteriori (ver apéndice A2.1), así como un intervalo creíble al 95 %

de confianza.

Esperanza	23.850
Varianza	26.235
Intervalo	$(14, 34)$

Tabla2.1. Momentos e intervalos creíbles
de la distribución predictiva a posteriori.

En el siguiente capítulo analizaremos las aproximaciones asintóticas: aproximación normal y método de Laplace.

Capítulo 3

Aproximaciones asintóticas

La solución al problema de integrar el denominador de (2.3) involucra usar métodos asintóticos para obtener aproximaciones analíticas de la densidad a posteriori. El resultado más simple es usar una aproximación normal a la distribución a posteriori, que es esencialmente una versión bayesiana del teorema del límite central. Una técnica asintótica más complicada es el método de Laplace (Tierney y Kadane, 1986. [25]), que permite obtener más precisión en las aproximaciones asimétricas a posteriori.

3.1. Aproximación normal

Cuando el número de datos es bastante grande, la verosimilitud es bastante alta, y pequeños cambios en la distribución a priori podría tener efectos en el resultado de la distribución a posteriori. El siguiente teorema muestra que la distribución a posteriori puede ser aproximadamente normal.

Teorema 3.1. *Supóngase que* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_i(x_i \mid \boldsymbol{\theta})$, *y entonces* $f(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(x_i \mid \boldsymbol{\theta})$. *Supóngase que la distribución a priori* $\pi(\boldsymbol{\theta})$ *y* $f(\mathbf{x} \mid \boldsymbol{\theta})$ *son positivas y dos veces diferenciable cerca de* $\widehat{\boldsymbol{\theta}}^{\pi}$, *la moda a posteriori de* θ, *asumiendo que existe. Entonces bajo condiciones adecuadas de regularidad, la distribución a posteriori* $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ *para* n *grande puede ser aproximada por una distribución normal teniendo media igual a la moda a posteriori, y matriz de covarianza igual a menos la inversa Hessiana (segunda derivada) del logaritmo a posteriori evaluado en la moda. Esta matriz es denotada como* $[I^{\pi}(\mathbf{x})]^{-1}$. *Dado que ésta es la generalización de la matriz de información de Fisher para* $\boldsymbol{\theta}$, *se tiene que*

$$I_{ij}^{\pi}(\mathbf{x}) = -\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\left(f(\mathbf{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\right)\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{\pi}}.$$

Demostración. Se probará para el caso unidimensional. Escribiendo

$$l(\theta) = \log f(\mathbf{x} \mid \theta)\, \pi(\theta)$$

y usando la expansión de Taylor para $l(\theta)$ tenemos

$$\begin{aligned}
\pi(\theta \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta)\, \pi(\theta) = \exp(\log(f(\mathbf{x} \mid \theta)\, \pi(\theta))) = \exp(l(\theta)) \\
&\approx \exp\left[l\left(\widehat{\theta}\right) + \frac{\partial}{\partial\theta} l(\theta) \mid_{\theta=\widehat{\theta}^{\pi}} \left(\theta - \widehat{\theta}\right) + \frac{\partial^2}{\partial\theta^2} l(\theta) \mid_{\theta=\widehat{\theta}^{\pi}} \frac{\left(\theta - \widehat{\theta}\right)^2}{2} \right] \\
&= \exp\left[l\left(\widehat{\theta}\right) - \frac{1}{2} I^{\pi}(\mathbf{x}) \left(\theta - \widehat{\theta}\right)^2 \right] \\
&\propto N\left(\widehat{\theta}^{\pi}, [I^{\pi}(\mathbf{x})]^{-1}\right). \qquad (3.1)
\end{aligned}$$

□

Ejemplo 3.1. ([5]) Supóngase que 16 consumidores han sido seleccionados para comparar dos tipos de comida en base al sabor, una comida es más cara que la otra. Después de someterlos a la prueba, los consumidores dan sus opiniones. El resultado es que 13 de los 16 consumidores prefieren la comida más cara.

Supóngase que θ es la probabilidad de que un consumidor prefiera la comida más cara.

Sea

$$Y_i = \begin{cases} 1 & \text{si el consumidor } i \text{ prefiere la comida más cara} \\ 0 & \text{en otro caso.} \end{cases}$$

Las decisiones de los consumidores son independientes y θ es constante sobre los consumidores, entonces sus decisiones forman una secuencia de pruebas de Bernoulli.

Definiendo

$$X = \sum_{i=1}^{16} Y_i$$

entonces la función de probabilidad es

$$X \mid \theta \sim Bi(16, \theta),$$

esto es, la distribución muestral para x es

$$f(x \mid \theta) = \binom{16}{x} \theta^x (1-\theta)^{16-x}$$

Como se tomó una sola muestra, la distribución de probabilidad es también la función de verosimilitud. Para seleccionar la distribución a priori, observando a la verosimilitud como una función de θ, se ve que la distribución beta ofrece una familia conjugada, pues su función de densidad es

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Ahora, la distribución a posteriori para θ es fácil de calcular como:

$$\begin{aligned}\pi(\theta \mid x) &\propto f(x \mid \theta)\pi(\theta)\\ &\propto \theta^{x+\alpha-1}(1-\theta)^{16-x+\beta-1}\\ &\propto Be(x+\alpha, 16-x+\beta).\end{aligned}$$

Si se toma a un elemento de la familia Beta con parámetros $\alpha = \beta = 1$ se obtiene una distribución a apriori uniforme. Así, se tiene

$$f(x \mid \theta)\pi(\theta) \propto \theta^{x}(1-\theta)^{n-x},$$

entonces

$$l(\theta) = \log f(x \mid \theta)\pi(\theta) = x\log\theta + (n-x)\log(1-\theta).$$

Tomando la derivada de $l(\theta)$

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{x}{\theta} - \frac{(n-x)}{(1-\theta)}$$

e igualándola a cero se obtiene

$$\begin{aligned}x(1-\theta) - \theta(n-x) &= 0,\\ x - x\theta - \theta n + x\theta &= 0,\\ \widehat{\theta} &= \frac{x}{n}.\end{aligned}$$

La segunda derivada es

$$\frac{\partial^2 l(\theta)}{\partial\theta^2} = -\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2},$$

Ahora, evaluándola en el estimador se tiene que

$$\begin{aligned}\frac{\partial^2 l(\theta)}{\partial\theta^2} \quad \Big| \quad _{\theta=\widehat{\theta}} &= -\frac{x}{\left[\frac{x}{n}\right]^2} - \frac{(n-x)}{\left(1-\left[\frac{x}{n}\right]\right)^2} = -\frac{xn^2}{x^2} - \frac{(n-x)}{\frac{(n-x)^2}{n^2}}\\ &= -\frac{n^2}{x} - \frac{n^2}{(n-x)} = -\frac{n}{\widehat{\theta}} - \frac{n}{\left(1-\widehat{\theta}\right)}.\end{aligned}$$

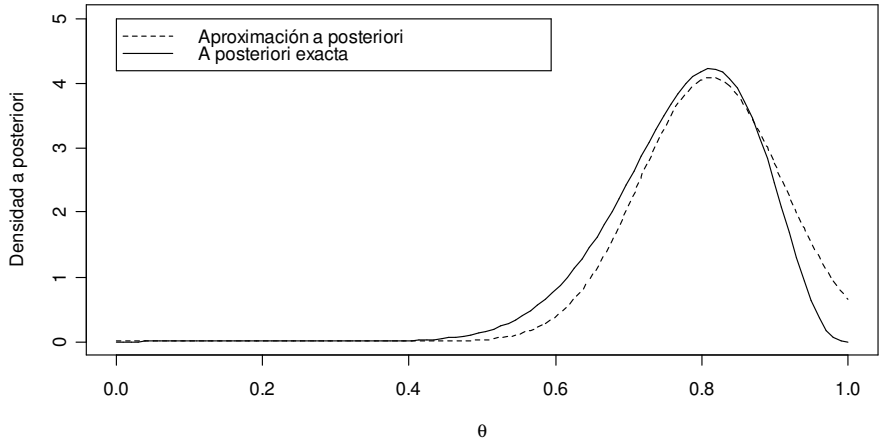

Figura 3.1: Distribución a posteriori exacta y aproximación para el ejemplo 3.1.

Entonces

$$\begin{aligned}\left[I^{\pi}(x)\right]^{-1} &= \left(\frac{n}{\widehat{\theta}}+\frac{n}{\left(1-\widehat{\theta}\right)}\right)^{-1}=\left(\frac{n-n\widehat{\theta}+n\widehat{\theta}}{\widehat{\theta}\left(1-\widehat{\theta}\right)}\right)^{-1} \\ &= \left(\frac{n}{\widehat{\theta}\left(1-\widehat{\theta}\right)}\right)^{-1}=\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{n}.\end{aligned}$$

Por lo tanto la aproximación a la distribución a posteriori es

$$\pi(\theta \mid x) \approx N\left(\widehat{\theta}, \frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{n}\right).$$

Luego, la distribución exacta a posteriori para este caso es $Beta\,(14,4)$ y la aproximación a la a posteriori es $N\left(\widehat{\theta}, \frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{n}\right)$ donde $\widehat{\theta}=\frac{13}{16}$. En el apéndice A3.1 se presenta el código de esta aproximación y en la figura 3.1 se muestra la aproximación.

La estimación de $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ puede ser pobre si la verdadera distribución a posteriori difiere significativamente de una normal. En tal caso se puede obtener

más exactitud en los estimadores a posteriori sin mucho esfuerzo en términos de derivadas de alto orden o transformaciones complicadas, con una técnica conocida como Método de Laplace. Esto se explicará a continuación.

3.2. Método de Laplace

Supóngase que f es una función suave y positiva de θ, y h es una función suave de θ donde $-h$ tiene un único máximo $\widehat{\theta}$. Tomando θ univariado por simplicidad, se desea aproximar

$$I = \int f(\theta) \exp(-nh(\theta))\, d\theta.$$

Usando la expansión en serie de Taylor para f y h cerca de $\widehat{\theta}$ ose btiene

$$\begin{aligned} I &\approx \int \left[f\left(\widehat{\theta}\right) + \frac{f'\left(\widehat{\theta}\right)}{1!}\left(\theta - \widehat{\theta}\right) + \frac{f''\left(\widehat{\theta}\right)}{2!}\left(\theta - \widehat{\theta}\right)^2 \right] \\ &\quad \times \exp\left[-nh\left(\widehat{\theta}\right) - nh'\left(\widehat{\theta}\right)\left(\theta - \widehat{\theta}\right) - nh''\left(\widehat{\theta}\right)\frac{\left(\theta - \widehat{\theta}\right)^2}{2!} \right] d\theta. \end{aligned}$$

Puesto que $h'\left(\widehat{\theta}\right) = 0$ por definición de $\widehat{\theta}$ se tiene

$$\begin{aligned} I &\approx \exp\left(-nh\left(\widehat{\theta}\right)\right) \int \left[f\left(\widehat{\theta}\right) + \frac{f'\left(\widehat{\theta}\right)}{1!}\left(\theta - \widehat{\theta}\right) + \frac{f''\left(\widehat{\theta}\right)}{2!}\left(\theta - \widehat{\theta}\right)^2 \right] \\ &\quad \exp\left[-\frac{\left(\theta - \widehat{\theta}\right)^2}{2\left(nh''\left(\widehat{\theta}\right)\right)^{-1}} \right] d\theta, \end{aligned}$$

y como el término final adentro de la integral es proporcional a $N\left(\widehat{\theta}, \left(nh''\left(\widehat{\theta}\right)\right)^{-1}\right)$

$$\begin{aligned} I &\approx \exp\left(-nh\left(\widehat{\theta}\right)\right) f\left(\widehat{\theta}\right) \int \exp\left[-\frac{\left(\theta-\widehat{\theta}\right)^2}{2\left(nh''\left(\widehat{\theta}\right)\right)^{-1}}\right] d\theta \\ &\quad + \exp\left(-nh\left(\widehat{\theta}\right)\right) f'\left(\widehat{\theta}\right) \int \left(\theta-\widehat{\theta}\right) \exp\left[-\frac{\left(\theta-\widehat{\theta}\right)^2}{2\left(nh''\left(\widehat{\theta}\right)\right)^{-1}}\right] d\theta \\ &\quad + \exp\left(-nh\left(\widehat{\theta}\right)\right) \frac{f''\left(\widehat{\theta}\right)}{2} \int \left(\theta-\widehat{\theta}\right)^2 \exp\left[-\frac{\left(\theta-\widehat{\theta}\right)^2}{2\left(nh''\left(\widehat{\theta}\right)\right)^{-1}}\right] d\theta \end{aligned}$$

$$\begin{aligned} &= \exp\left(-nh\left(\widehat{\theta}\right)\right) f\left(\widehat{\theta}\right) \sqrt{2\pi\left(nh''\left(\widehat{\theta}\right)\right)^{-1}} \\ &\quad + \exp\left(-nh\left(\widehat{\theta}\right)\right) f'\left(\widehat{\theta}\right) \sqrt{2\pi\left(nh''\left(\widehat{\theta}\right)\right)^{-1}} E\left[\left(\theta-\widehat{\theta}\right)\right] \\ &\quad + \exp\left(-nh\left(\widehat{\theta}\right)\right) \frac{f''\left(\widehat{\theta}\right)}{2} \sqrt{2\pi\left(nh''\left(\widehat{\theta}\right)\right)^{-1}} E\left[\left(\theta-\widehat{\theta}\right)^2\right] \\ &= \exp\left(-nh\left(\widehat{\theta}\right)\right) f\left(\widehat{\theta}\right) \sqrt{\frac{2\pi}{nh''\left(\widehat{\theta}\right)}} + \exp\left(-nh\left(\widehat{\theta}\right)\right) \frac{f''\left(\widehat{\theta}\right)}{2} \sqrt{\frac{2\pi}{nh''\left(\widehat{\theta}\right)}} Var\left(\theta\right) \\ &= \exp\left(-nh\left(\widehat{\theta}\right)\right) f\left(\widehat{\theta}\right) \sqrt{\frac{2\pi}{nh''\left(\widehat{\theta}\right)}} \left(1 + \frac{f''\left(\widehat{\theta}\right)}{2f\left(\widehat{\theta}\right)} Var\left(\theta\right)\right), \end{aligned}$$

entonces

$$I \approx \exp\left(-nh\left(\widehat{\theta}\right)\right) f\left(\widehat{\theta}\right) \sqrt{\frac{2\pi}{nh''\left(\widehat{\theta}\right)}} \left(1 + \frac{f''\left(\widehat{\theta}\right)}{2f\left(\widehat{\theta}\right)} Var\left(\theta\right)\right)$$

donde $Var\left(\theta\right) = \left[nh''\left(\widehat{\theta}\right)\right]^{-1} = O\left(\frac{1}{n}\right)$.

Definición 3.1. *Se dice que $f = O\left(g\right)$ cuando $x \to \infty$ si y sólo si existe un x_0 y M tal que*

$$\frac{f\left(x\right)}{g\left(x\right)} < M \qquad \text{para } x > x_0.$$

En general para $\boldsymbol{\theta}$ $m-$dimensional se tiene que

$$I = \widehat{I}\left\{1 + O\left(\frac{1}{n}\right)\right\},$$

donde

$$\widehat{I} = f\left(\widehat{\boldsymbol{\theta}}\right)\left(\frac{2\pi}{n}\right)^{\frac{m}{2}} |\widetilde{\sum}|^{\frac{1}{2}} \exp\left(-nh\left(\widehat{\boldsymbol{\theta}}\right)\right) \tag{3.2}$$

y

$$\widetilde{\sum} = \left[D^2 h\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1},$$

con

$$\left[D^2 h\left(\widehat{\boldsymbol{\theta}}\right)\right]_{ij} = \frac{\partial^2 h(\theta)}{\partial\theta_i \partial\theta_j} |_{\theta=\widehat{\theta}}$$

la matriz Hessiana de h evaluada en $\widehat{\boldsymbol{\theta}}$.

Ahora, se desea calcular la esperanza a posteriori de una función $g(\boldsymbol{\theta})$, donde $-nh(\boldsymbol{\theta}) = \log[f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})]$. Entonces

$$E[g(\boldsymbol{\theta})] = \frac{\int g(\boldsymbol{\theta}) \exp(-nh(\boldsymbol{\theta}))\, d\boldsymbol{\theta}}{\int \exp(-nh(\boldsymbol{\theta}))\, d\boldsymbol{\theta}}. \tag{3.3}$$

Aplicando el método de Laplace (3.2) para el numerador ($f = g$) y denominador ($f = 1$) se obtiene

$$\begin{aligned}
E[g(\boldsymbol{\theta})] &= \frac{g\left(\widehat{\boldsymbol{\theta}}\right)\left(\frac{2\pi}{n}\right)^{\frac{m}{2}} |\widetilde{\sum}|^{\frac{1}{2}} \exp\left(-nh\left(\widehat{\boldsymbol{\theta}}\right)\right)\left[1 + O_1\left(\frac{1}{n}\right)\right]}{\left(\frac{2\pi}{n}\right)^{\frac{m}{2}} |\widetilde{\sum}|^{\frac{1}{2}} \exp\left(-nh\left(\widehat{\boldsymbol{\theta}}\right)\right)\left[1 + O_2\left(\frac{1}{n}\right)\right]} \\
&= g\left(\widehat{\boldsymbol{\theta}}\right)\left[\frac{1 + O_2\left(\frac{1}{n}\right) - O_2\left(\frac{1}{n}\right) + O_1\left(\frac{1}{n}\right)}{1 + O_2\left(\frac{1}{n}\right)}\right] \\
&= g\left(\widehat{\boldsymbol{\theta}}\right)\left[1 + \frac{O_3\left(\frac{1}{n}\right)}{1 + O_2\left(\frac{1}{n}\right)}\right] \\
&= g\left(\widehat{\boldsymbol{\theta}}\right)\left[1 + O\left(\frac{1}{n}\right)\right].
\end{aligned}$$

Ejemplo 3.2. Considerando el ejemplo 3.1, ahora se quiere aproximar la media

a posteriori por el método de Laplace, se tiene que

$$\begin{aligned} E_{\pi(\theta|x)}[\theta] &= \int \theta\pi(\theta \mid x)\, d\theta \\ &= \frac{\int \theta f(x \mid \theta)\, \pi(\theta)\, d\theta}{\int f(x \mid \theta)\, \pi(\theta)\, d\theta} \\ &= \frac{\int \theta \cdot \theta^x (1-\theta)^{n-x}\, d\theta}{\int \theta^x (1-\theta)^{n-x}\, d\theta} \\ &= \widehat{\theta}\left[1 + O\left(\frac{1}{n}\right)\right]. \end{aligned}$$

Aquí $g\left(\widehat{\boldsymbol{\theta}}\right) = \widehat{\theta}$ es una aproximación de primer orden a la media a posteriori de θ. Esta aplicación del método de Laplace produce el mismo estimador (con el mismo orden de precisión) que la aproximación modal.

Pero si se considera el truco introducido por Tierney y Kadane (1986) [22], el cual consiste en:

Si $g(\boldsymbol{\theta}) > 0$ para todo $\boldsymbol{\theta}$, escribiendo el integrando del numerador de la ecuación (3.3) como

$$\exp(\log g(\boldsymbol{\theta}) - nh(\boldsymbol{\theta})) \equiv \exp(-nh^*(\boldsymbol{\theta})),$$

y usando el método de Laplace (3.2) con $f = 1$ en ambos numerador y denominador, el resultado es

$$E[g(\boldsymbol{\theta})] = \frac{|\sum^*|^{\frac{1}{2}} \exp(-nh^*(\boldsymbol{\theta}^*))}{|\widetilde{\sum}|^{\frac{1}{2}} \exp\left(-nh\left(\widehat{\boldsymbol{\theta}}\right)\right)}\left[1 + O\left(\frac{1}{n^2}\right)\right],$$

donde $\boldsymbol{\theta}^*$ es máximo de $-h^*$ y $\sum^* = [D^2 h^*(\boldsymbol{\theta}^*)]^{-1}$, similar a la definición de $\widehat{\boldsymbol{\theta}}$ y $\widetilde{\sum}$. Esta aplicación del método de Laplace prueba una aproximación de segundo orden a la media a posteriori de $g(\boldsymbol{\theta})$.

Para modelos multiparamétricos en $\boldsymbol{\theta}$ (digamos, de dimensión mayor que 10) los métodos de Laplace podrían raramente ser precisos, y los cálculos numéricos de la matriz Hessiana asociada podrían ser difíciles. Por esta razón, se han incrementado los métodos iterativos, especialmente los que involucran muestreo Monte Carlo, para cálculos a posteriori. Estos métodos típicamente involucran largas corridas, pero son más generales, dan una información más completa y son más fáciles de programar. De estos métodos se hablará en el siguiente capítulo.

Capítulo 4

Métodos de Integración Monte Carlo

4.1. Preliminares

La idea básica del método de Monte Carlo consiste en escribir la integral requerida como el valor esperado de alguna función con respecto a alguna distribución de probabilidad. Esto sugiere una solución estadística al problema de integración.

En la base de la generación de variables aleatorias está la generación de números pseudoaleatorios que deriva directamente en la generación de variables aleatorias uniformes en el intervalo [0,1].

Una de las más comunes aproximaciones para generar números pseudo-aleatorios inicia con un valor inicial x_0, llamado la semilla, y entonces con cálculos recursivos de sucesivos valores x_n, $n \geq 1$, se tiene

$$x_n = ax_{n-1} (\mathrm{mod}\, m) \tag{4.1}$$

donde a y m son enteros positivos, y significa que ax_{n-1} es dividida por m y el restante es tomado como el valor de x_n. Entonces cada x_n es $0, 1, \ldots, m-1$ y la cantidad $\frac{x_n}{m}$ (llamado un número pseudoaleatorio) es tomada como una aproximación al valor de una variable aleatoria uniforme en $(0, 1)$. La aproximación especificada por la ecuación (4.1) para generar números aleatorios es llamada el método del multiplicador congruencial.

Otro generador de números pseudo-aleatorios que usan recursión del tipo

$$x_n = ax_{n-1} + c (\mathrm{mod}\, m)$$

se llama generador congruente mixto (pues involucra una suma y multiplicación).

Posteriormente, estas variables uniformes son de gran utilidad en la generación de otras variables aleatorias.

Si se considera una variable aleatoria continua que tiene una distribución F, el método más general es el desarrollado en el teorema de la transformación integral de la probabilidad, el cual se enunciará y demostrará a continuación.

Teorema 4.1. *(Transformación integral de la probabilidad) Sea X una variable aleatoria con función de distribución F, entonces*

a) Si F es continua entonces $F(x)$ es uniforme en $[0,1]$.

b) Si $F^{-}(u) = \inf\{x; F(x) \geqslant u\}$ es la inversa generalizada de F y U es la distribución unifome en $[0,1]$ entonces $F^{-}(U) \sim F$. En consecuencia, para generar una $X \sim F$ basta generar $U \sim U[0,1]$ y hacer la transformación $x = F^{-}(u)$.

Demostración. a) Tenemos que ver que para todo $[a,b] \subset (0,1)$ se tiene que

$$P[a \leq F(X) \leq b] = b - a.$$

Así, por ser F continua se tiene que $F^{-1}[a,b]$ es cerrado, y además será intervalo por ser no decreciente. Es decir $F^{-1}[a,b] = [x_a, x_b]$, donde $F(x_a) = a$ y $F(x_b) = b$. Luego

$$[x_a \leq x \leq x_b] \Longleftrightarrow [a \leq F(x) \leq b]$$

de donde

$$[x_a \leq X \leq x_b] \Longleftrightarrow [a \leq F(X) \leq b]$$

por lo tanto

$$P[x_a \leq X \leq x_b] \Longleftrightarrow [a \leq F(X) \leq b] = F(x_b) = b(x_a) = b - a.$$

b) Para todo $u \in [0,1]$ y para todo $x \in F^{-}([0,1])$ la inversa generalizada satisface

$$F\left(F^{-}(u)\right) \geq u \text{ y } F^{-}(F(x)) \leq x,$$

de aquí que

$$\left\{(u,x) : F^{-}(u) \leq x\right\} = \{(u,x) : F(x) \geq u\}$$

entonces

$$P\left[F^{-}(U) \leq x\right] = P[U \leq F(x)] = F(x).$$

□

Los algoritmos correspondientes a estos métodos son expuestos en el apéndice A4.1. Una de las aplicaciones de la generación de números aleatorios fue el cálculo de integrales. Esta aproximación de integrales es llamada la aproximación Monte Carlo.

Un aspecto fundamental es saber cuántas iteraciones necesitamos en nuestra simulación. Está claro que, en términos generales, entre más iteraciones se realicen mejor será la aproximación obtenida. Sin embargo, existe un tope a partir del cual las mejoras que se obtienen en la estimación no compensan el esfuerzo realizado. En muchos casos la experiencia práctica, el sentido común o la disponibilidad de tiempo y recursos nos pueden dar una cota para n. El procedimiento teórico que se utiliza para el cálculo del tamaño de muestra (número de iteraciones) se muestra en el apéndice A4.2.

4.2. Motivación del concepto de integración Monte Carlo

Sea $\pi : \mathbb{R} \longrightarrow \mathbb{R}^{+}$. Supóngase que existe $M > 0$ tal que $0 \leq \pi(\theta) \leq M$ para todo $\theta \in [a, b]$, y que se quiere calcular la integral ([14])

$$A = \int_{a}^{b} \pi(\theta)\, d\theta. \tag{4.2}$$

El valor de ésta integral no es más que el área bajo la curva φ de la función $\pi(\theta)$ para todo $\theta \in [a, b]$. Esta gráfica queda inscrita en el rectángulo $R = [a, b] \times [0, M]$.

Si se define

$$k(\theta, \varphi) = \frac{1}{M(b-a)} I_R(\theta, \varphi),$$

entonces $k(\theta, \varphi)$ corresponde a la función de densidad de una distribución uniforme sobre el rectángulo R. La integral A puede entonces estimarse simulando una muestra $(\theta_1, \varphi_1), ..., (\theta_N, \varphi_N)$ de $k(\theta, \varphi)$ y contando cuántos de estos valores caen bajo la curva $\varphi = \pi(\theta)$. Específicamente, sea

$$N_f = \sum_{i=1}^{N} I_C(\theta_i, \varphi_i),$$

donde $C = \{(\theta, \varphi) \in R : a \leq \theta \leq b, 0 \leq \varphi \leq \pi(\theta)\}$. Entonces

$$\widehat{A}_1 = M(b-a)\frac{N_f}{N},$$

es un estimador insesgado de A. Para verificar esto, nótese que cada observación (θ_i, φ_i) corresponde a un ensayo de Bernoulli (pues está o no está en el conjunto C), con probabilidad de éxito $A/\left[M\left(b-a\right)\right]$, por lo que

$$\begin{aligned} E\left[\widehat{A}_1\right] &= E\left[M\left(b-a\right)\frac{N_f}{N}\right] \\ &= \frac{M\left(b-a\right)}{N}E\left[N_f\right] \\ &= \frac{M\left(b-a\right)}{N}E\left[\sum_{i=1}^{N} I_C\left(\theta_i,\varphi_i\right)\right] \\ &= \frac{M\left(b-a\right)}{N}\frac{NA}{\left\{M\left(b-a\right)\right\}} \\ &= A \end{aligned} \tag{4.3}$$

por lo tanto, $\widehat{A}_1$ es un estimador insesgado de A. La varianza de éste estimador es:

$$\begin{aligned} Var\left(\widehat{A}_1\right) &= E\left[\left(\widehat{A}_1 - A\right)^2\right] \\ &= E\left[\widehat{A}_1^2\right] + E^2\left[\widehat{A}_1\right] \\ &= \frac{M\left(b-a\right)}{N}A - \frac{A^2}{N} + A^2 - A^2 \\ &= \frac{A}{N}\left(M\left(b-a\right) - A\right). \end{aligned}$$

Ejemplo 4.1. Sea

$$\pi\left(\theta\right) = \frac{1}{2}\exp\left(-\frac{1}{2}\theta\right),$$

sea $M = 0.1839$ tal que $0 \leq \pi\left(\theta\right) \leq 0.1839$ para todo $\theta \in [2,5]$. Entonces se quiere estimar el valor de la integral

$$A = \int_2^5 \frac{1}{2}\exp\left(-\frac{1}{2}\theta\right)d\theta$$

que es aproximadamente

$$\widehat{A}_1 = 0.1839\left(3\right)\frac{N_f}{100000}$$

donde

$$N_f = \sum_{i=1}^{N} I_C\left(\theta_i,\varphi_i\right),$$

y

$$C = \left\{ (\theta, \varphi) \in R = [2,5] \times [0, 0.1839] : 2 \leq \theta \leq 5, 0 \leq \varphi \leq \frac{1}{2} \exp\left(-\frac{1}{2}\theta\right) \right\}.$$

Se obtiene como resultado (ver apéndice A4.3) que la aproximación $\widehat{A}_1 = 0.2859$, mientras que el valor real de la integral es 0.2857.

4.3. Métodos directos

En problemas de estadística se tiene la situación de evaluar:

$$h(\mathbf{x}) = E_\pi \left[f(\mathbf{x}|\theta) \right] = \int f(\mathbf{x}|\theta)\, \pi(\theta)\, d\theta. \tag{4.4}$$

Donde $\theta \sim \pi(\theta)$. Uno de los métodos para evaluar la integral (4.4) es el muestreo directo.

4.3.1. Muestreo directo

La definición más básica de integración Monte Carlo consiste en suponer $\theta \sim \pi(\theta \mid \mathbf{x})$, y se busca el valor de

$$\gamma \equiv E\left[g(\theta) \right] = \int g(\theta)\, \pi(\theta \mid \mathbf{x})\, d\theta. \tag{4.5}$$

Entonces, se generan posibles variables aleatorias $\theta_1, ..., \theta_m$ *i.i.d.* de $\pi(\theta \mid \mathbf{x})$ se tiene la aproximación de (4.5) como el promedio muestral:

$$\widehat{\gamma} = \frac{1}{m} \sum_{i=1}^{m} g(\theta_i) \tag{4.6}$$

que converge a $E\left[g(\theta) \right]$ con probabilidad 1, cuando $m \longrightarrow +\infty$, según la ley fuerte de los grandes números. Se demostrará la ley débil de los grandes números, lo mismo que dos teoremas auxiliares que a continuación se enuncian, y se enunciará la ley fuerte de los grandes números.

Teorema 4.2. *(**Desigualdad de Markov**) Supóngase que X es una variable aleatoria tal que $P(X \geq 0) = 1$. Entonces, para cualquier número $t > 0$*

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Demostración. Supóngase que X tiene una distribución discreta cuya función de densidad de probabilidad es f. La demostración para una distribución continua o un tipo de distribución más general es análoga. Para una distribución discreta

$$E(X) = \sum_x xf(x) = \sum_{x<t} xf(x) + \sum_{x\geq t} xf(x)$$

como X puede tomar únicamente valores no negativos todos los términos de la suma son no negativos. Luego

$$E(X) \geq \sum_{x\geq t} xf(x) \geq \sum_{x\geq t} tf(x) = tP(X \geq t).$$

Por tanto

$$\frac{E(X)}{t} \geq P(X \geq t).$$

□

Teorema 4.3. ***(Desigualdad de Chebyshev)*** *Sea X una variable aleatoria para la cual $Var(X)$ existe. Entonces para cualquier número concreto $\varepsilon > 0$*

$$P(\mid X - E(X) \mid \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

Demostración. Sea $Y = [X - E(X)]^2$. Entonces $P(Y \geq 0) = 1$ y $E(Y) = Var(X)$. Aplicando a Y la desigualdad de Markov

$$P(\mid X - E(X) \mid \geq \varepsilon) = P\left(Y \geq \varepsilon^2\right) \leq \frac{E(Y)}{\varepsilon^2} = \frac{Var(X)}{\varepsilon^2}.$$

Por lo tanto

$$P(\mid X - E(X) \mid \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}.$$

□

Teorema 4.4. ***(Ley débil de los grandes números)*** *Supóngase que $X_1, ..., X_n$ constituyen una muestra aleatoria cuya media es μ y sea $\overline{X}_n$ la media muestral. Entonces*

$$P \lim_{n\longrightarrow\infty} \overline{X}_n = \mu.$$

Es decir, la sucesión $X_1, ..., X_n$ de variables aleatorias converge en probabilidad a μ, si para cualquier número dado $\varepsilon > 0$,

$$\lim_{n\longrightarrow\infty} P\left(|\overline{X}_n - \mu| < \varepsilon\right) = 1.$$

Demostración. Supóngase que la distribución de donde se ha seleccionado la muestra aleatoria tiene una media finita μ y una varianza finita σ^2. De la desigualdad de Chebyshev, para cualquier número dado $\epsilon > 0$

$$\begin{aligned} P\left(|\overline{X}_n - \mu| \geq \epsilon\right) &\leq \frac{Var\left(\overline{X}\right)}{\epsilon^2} = \frac{Var\left(X\right)}{n\epsilon^2}, \\ 1 - P\left(|\overline{X}_n - \mu| < \epsilon\right) &\leq \frac{Var\left(X\right)}{n\epsilon^2}, \\ 1 - \frac{Var\left(X\right)}{n\epsilon^2} &\leq P\left(|\overline{X}_n - \mu| < \epsilon\right). \end{aligned}$$

Entonces

$$\lim_{n\to\infty} P\left(|\ \overline{X_n} - \mu\ | < \varepsilon\right) = 1,$$

por lo tanto

$$P \lim_{n\to\infty} \overline{X_n} = \mu.$$

□

Es decir, la media muestral siempre converge en probabilidad a la media de la distribución de la que se seleccionó la muestra aleatoria.

Teorema 4.5. ***(Ley fuerte de los grandes números)*** *Sea $X_1, ..., X_n$ una muestra aleatoria cuya media es μ y sea $\overline{X}_n$ la media muestral. Entonces*

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

Es decir la sucesión $X_1, ..., X_n$ converge a μ con probabilidad 1.

En este caso, $\pi\left(\theta \mid \mathbf{x}\right)$ es una distribución a posteriori y γ es la media a posteriori de $g\left(\theta\right)$. Aquí el cálculo de la esperanza a posteriori solo requiere una muestra de tamaño m de la distribución a posteriori.

Note que, en contraste con los métodos de la sección anterior, la calidad de la aproximación en (4.6) mejora cuando se incrementa m, el tamaño de la muestra Monte Carlo (que nosotros elegimos) y se aproxima a n, el tamaño del conjunto de datos (que está típicamente más allá de nuestro control).

Otro contraste con los métodos asintóticos es que la estructura de (4.6) también nos permite evaluar su precisión para cualquier m fijo. Dado que $\widehat{\gamma}$ es en si mismo una media muestral de observaciones independientes, se tiene que

$$\begin{aligned} Var\left(\widehat{\gamma}\right) &= Var\left(\frac{1}{m}\sum_{i=1}^{m} g\left(\theta_i\right)\right) = \frac{1}{m^2} Var\left(\sum_{i=1}^{m} g\left(\theta_i\right)\right) \\ &= \frac{m}{m^2} Var\left(g\left(\theta\right)\right) = \frac{1}{m} Var\left[g\left(\theta\right)\right], \end{aligned}$$

pero $Var\left[g\left(\theta\right)\right]$ puede ser estimada por la varianza muestral de los $g\left(\theta_i\right)$ valores,

$$\begin{aligned} Var\left[g\left(\theta\right)\right] &= E\left[\left(g\left(\theta\right)-\widehat{\gamma}\right)^2\right] \\ &= \int \left(g\left(\theta\right)-\widehat{\gamma}\right)^2 \pi\left(\theta \mid x\right) d\theta \\ &\approx \frac{1}{m}\sum_{i=1}^{m}\left(g\left(\theta_i\right)-\widehat{\gamma}\right)^2, \end{aligned}$$

entonces el error estandar estimado de $\widehat{\gamma}$ esta dado por

$$\widehat{se}\left(\widehat{\gamma}\right)=\sqrt{\frac{1}{m\left(m-1\right)}\sum_{i=1}^{m}\left(g\left(\theta_i\right)-\widehat{\gamma}\right)^2}.$$

Finalmente, el teorema del límite central implica que $\widehat{\gamma}\pm 2\widehat{se}\left(\widehat{\gamma}\right)$ proporciona una aproximación del 95 % de confianza para el intervalo del verdadero valor de la media a posteriori γ. De nuevo, m puede ser elegido tan grande como sea necesario para proporcionar cualquier nivel de confianza deseado del intervalo.

Ejemplo 4.2. Sea $X_i \sim N\left(x_i|\theta,\sigma^2\right)$, con θ desconocida y σ^2 conocida, entonces $\overline{x} \sim N\left(\overline{\mathbf{x}}|\theta,\frac{\sigma^2}{n}\right)$, con distribución a priori $\pi\left(\theta\right)\propto 1$. Queremos calcular

$$\pi\left(\theta|\mathbf{x},\sigma^2\right)\propto f\left(\mathbf{x}|\theta,\sigma^2\right)\pi\left(\theta\right),$$

donde cada X_i tiene función de densidad

$$f\left(x_i|\theta,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(x_i-\theta\right)^2\right).$$

Así,

$$\begin{aligned}
\prod_{i=1}^{n} f\left(x_i|\theta,\sigma^2\right) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(x_i-\theta\right)^2\right)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\theta\right)^2\right)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\overline{\mathbf{x}})-(\theta-\overline{\mathbf{x}})\right)^2\right)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\times\\
&\quad \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left[(x_i-\overline{\mathbf{x}})^2-2(x_i-\overline{\mathbf{x}})(\theta-\overline{\mathbf{x}})+(\theta-\overline{\mathbf{x}})^2\right]\right]\right)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\times\\
&\quad \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})^2-2(\theta-\overline{\mathbf{x}})\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})+n(\theta-\overline{\mathbf{x}})^2\right]\right)\\
&\propto \exp\left(-\frac{n}{2\sigma^2}(\overline{\mathbf{x}}-\theta)^2\right),
\end{aligned}$$

luego,

$$\begin{aligned}
\int f\left(\mathbf{x}\,|\theta,\sigma^2\right)\pi\left(\theta\right)d\theta &\propto \int\exp\left(-\frac{1}{2\frac{\sigma^2}{n}}(\overline{\mathbf{x}}-\theta)^2\right)d\theta\\
&= \sqrt{2\pi}\frac{\sigma}{\sqrt{n}}\int\frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}}\exp\left(-\frac{1}{2\frac{\sigma^2}{n}}(\overline{\mathbf{x}}-\theta)^2\right)d\theta=\sqrt{2\pi}\frac{\sigma}{\sqrt{n}},
\end{aligned}$$

por lo tanto,

$$\theta\mid\mathbf{x},\sigma^2\sim N\left(\theta\mid\overline{\mathbf{x}},\frac{\sigma^2}{n}\right).$$

La gráfica de la distribución a posteriori se muestra en la figura 4.1.

Para aproximar $E\left(\theta\mid\mathbf{x}\right)=\overline{\mathbf{x}}$, se generan $\theta^1,\theta^2,...,\theta^m\sim N\left(\theta\mid\overline{\mathbf{x}},\frac{\sigma^2}{n}\right)$. Luego

$$E\left(\theta\mid\sigma^2,\mathbf{x}\right)=\int\theta\pi\left(\theta\mid\sigma^2,\mathbf{x}\right)d\theta\simeq\frac{1}{m}\sum_{i=1}^{m}\theta^i.$$

Con $m=1000$, y $X_i\sim N\left(x_i\mid\theta,1\right)$. Aplicando el muestreo directo de Monte Carlo se obtiene (ver apéndice A4.4):

$$\widehat{E}\left(\theta\mid\sigma^2,\mathbf{x}\right)=0.01943,$$

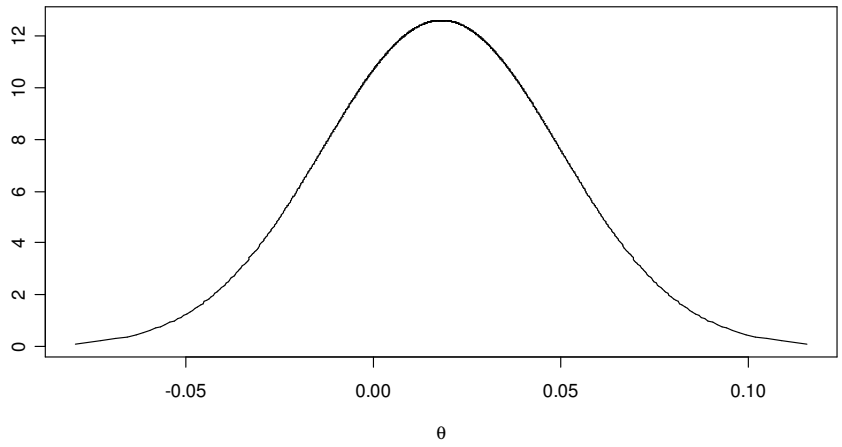

Figura 4.1: Distribución a posteriori del ejemplo 4.2.

la media obtenida de manera analítica es

$$\overline{\mathbf{x}} = 0.01941,$$

mientras que la varianza es

$$Var\left(\theta \mid \sigma^2, \mathbf{x}\right) = E\left(\theta^2 \mid \sigma^2, \mathbf{x}\right) - E\left(\theta \mid \sigma^2, \mathbf{x}\right)^2.$$

Por muestreo directo se tiene que

$$E\left(\theta^2 \mid \sigma^2, \mathbf{x}\right) = \int \theta^2 \pi\left(\theta \mid \sigma^2, \mathbf{x}\right) d\theta \simeq \frac{1}{m}\sum_{i=1}^{m}\left(\theta^i\right)^2 = 0.00137,$$

entonces

$$\begin{aligned} \widehat{Var}\left(\theta \mid \sigma^2, \mathbf{x}\right) &= 0.00137 - 0.01943^2 \\ &= 0.00099. \end{aligned}$$

La varianza de la distribución a posteriori es

$$\frac{\sigma^2}{n} = 0.001.$$

Así, el error estándar estimado de $\widehat{E}\left(\theta \mid \sigma^2, \mathbf{x}\right)$ es

$$\begin{aligned} \widehat{se}\left(\widehat{E}\left(\theta \mid \sigma^2, \mathbf{x}\right)\right) &= \sqrt{\frac{1}{m-1}\widehat{Var}\left(\theta \mid \sigma^2, \mathbf{x}\right)} \\ &= 0.00099. \end{aligned}$$

La simulación Monte Carlo proporciona un ejemplo donde la simulación es claramente apropiada. Note que

$$p \equiv P\left\{a < g\left(\theta\right) < b \mid \mathbf{x}\right\} = E\left\{I_{(a,b)}\left[g\left(\theta\right)\right] \mid \mathbf{x}\right\},$$

entonces un estimador de p es simplemente

$$\widehat{p} = \frac{\text{número de } \theta_{is} \in (a,b)}{m},$$

que se le asocia una distribución binomial con error estándar estimado

$$\widehat{se}(\widehat{p}) = \sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{m}}.$$

4.3.2. Método de composición

A veces no es posible obtener simulaciones de una distribución marginal, pero sí resulta fácil hacerlo de una distribución condicional.

Si $f\left(\mathbf{x} \mid \theta\right)$ es una densidad fácil de simular, para obtener una muestra

$$X_1, ..., X_m \overset{i.i.d.}{\sim} h\left(\mathbf{x}\right) = \int f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right) d\theta.$$

El algoritmo del método de composición dentro de los métodos directos de Monte Carlo se describe a continuación.

Entrada: m número de simulación deseada.
$\pi\left(\theta\right)$ Una distribución de probabilidad.
$f\left(\mathbf{x}|\theta\right)$ Distribución de probabilidad.
Salida: (X_j, θ_j) Muestra $i.i.d.$ de la densidad conjunta $h\left(\mathbf{x}, \theta\right) = \pi\left(\theta\right) f\left(\mathbf{x}|\theta\right)$
Paso 1.- Para $j = 1$ hasta m
Generar $\theta_j^* \sim \pi\left(\theta\right)$
Generar $X_j^* \sim f\left(\mathbf{x}|\theta_j^*\right)$
Paso 2.- **Salida** (X_j, θ_j) para $j = 1, ..., m$

Las cantidades $X_1, ..., X_m$ son una muestra $i.i.d.$ de la distribución marginal $h\left(\mathbf{x}\right)$.

Ejemplo 4.3. Sea una muestra $X_i \overset{iid}{\sim} N\left(x_i \mid \theta, \sigma^2\right)$, para $i = 1, ..., n$, con θ y σ^2 desconocidos, y supóngase una distribución a priori no informativa de Jeffreys para los parámetros (ver ejemplo 5) $\pi\left(\theta, \sigma\right) = \frac{1}{\sigma}$. Se tiene que la distribución marginal a posteriori con respecto a θ es

$$\pi\left(\theta|\mathbf{x}\right) = \int \pi\left(\theta, \sigma^2|\mathbf{x}\right) d\sigma^2 \tag{4.7}$$

y

$$\pi\left(\theta, \sigma^2|\mathbf{x}\right) = \pi\left(\theta|\sigma^2, \mathbf{x}\right) \pi\left(\sigma^2|\mathbf{x}\right). \tag{4.8}$$

Así, si

$$f\left(x_i \mid \theta, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x_i - \theta\right)^2\right],$$

entonces la función de verosimilitud es

$$\begin{aligned}
f\left(\mathbf{x} \mid \theta, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x_i - \theta\right)^2\right] \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \theta\right)^2\right] \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\left(x_i - \overline{\mathbf{x}}\right) - \left(\theta - \overline{\mathbf{x}}\right)\right]^2\right] \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \times \\
&\quad \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(x_i - \overline{\mathbf{x}}\right)^2 - 2\left(\theta - \overline{\mathbf{x}}\right)\sum_{i=1}^{n}\left(x_i - \overline{\mathbf{x}}\right) + n\left(\theta - \overline{\mathbf{x}}\right)^2\right]\right\} \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(x_i - \overline{\mathbf{x}}\right)^2 + n\left(\theta - \overline{\mathbf{x}}\right)^2\right]\right\}.
\end{aligned}$$

Se trabaja en términos de la precisión

$$\tau = \frac{1}{\sigma^2},$$

de donde

$$\sigma^2 = \frac{1}{\tau} \qquad \text{y} \qquad \sigma = \frac{1}{\sqrt{\tau}}$$

$$\begin{aligned} f(\mathbf{x} \mid \theta, \tau) &= \frac{1}{(2\pi)^{\frac{n}{2}} \left\{ \left[\frac{1}{\sqrt{\tau}} \right]^2 \right\}^{\frac{n}{2}}} \exp \left\{ -\frac{1}{2 \left[\frac{1}{\sqrt{\tau}} \right]^2} \left[\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 + n (\theta - \overline{\mathbf{x}})^2 \right] \right\} \\ &= \frac{\tau^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}} \exp \left\{ -\frac{\tau}{2} \left[\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 + n (\theta - \overline{\mathbf{x}})^2 \right] \right\}, \end{aligned}$$

luego

$$f(\mathbf{x} \mid \theta, \tau) \propto \tau^{\frac{n}{2}} \exp \left\{ -\frac{\tau}{2} \left[n (\theta - \overline{\mathbf{x}})^2 + \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 \right] \right\}.$$

Como

$$\pi(\theta, \tau) = \sqrt{\tau}$$

la distribución a posteriori esta dada por

$$\begin{aligned} \pi(\theta, \tau \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta, \tau) \pi(\theta, \tau) \\ &\propto \tau^{\frac{n}{2}} \tau^{\frac{1}{2}} \exp \left\{ -\frac{\tau}{2} \left[n (\theta - \overline{\mathbf{x}})^2 + \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 \right] \right\} \\ &\propto \tau^{\frac{n+1}{2}} \exp \left\{ -\frac{\tau}{2} \left[n (\theta - \overline{\mathbf{x}})^2 + \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 \right] \right\}. \end{aligned}$$

Ahora la marginal de θ tomando un valor fijo de τ esta dada por

$$\begin{aligned} \pi(\theta \mid \mathbf{x}, \tau) &\propto \exp \left\{ -\frac{\tau n}{2} \left[(\theta - \overline{\mathbf{x}})^2 \right] \right\} \propto \exp \left\{ -\frac{1}{2\frac{1}{\tau n}} \left[(\theta - \overline{\mathbf{x}})^2 \right] \right\} \\ &\propto N \left(\overline{\mathbf{x}}, \frac{1}{\tau n} \right). \end{aligned}$$

Entonces la distribución marginal a posteriori de θ dado $\mathbf{x}$ esta dada por

$$\theta \mid \mathbf{x}, \sigma^2 \sim N \left(\overline{\mathbf{x}}, \frac{\sigma^2}{n} \right).$$

Ahora

$$\pi(\tau \mid \mathbf{x}, \theta) \propto \tau^{\frac{n+1}{2}} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 \right\},$$

$$\begin{aligned}\pi\left(\sigma^2 \mid \mathbf{x}, \theta\right) &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)\sum_{i=1}^{n}\left(x_i-\overline{\mathbf{x}}\right)^2\right\} \\ &= \frac{1}{\left(\sigma^2\right)^{\frac{n+1}{2}}} \exp\left\{\left(\frac{\sum_{i=1}^{n}\left(x_i-\overline{\mathbf{x}}\right)^2}{2}\right)\left(\frac{1}{\sigma^2}\right)\right\} \\ &= \frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}+1}} \exp\left\{\left(\frac{\sum_{i=1}^{n}\left(x_i-\overline{\mathbf{x}}\right)^2}{2}\right)\left(\frac{1}{\sigma^2}\right)\right\}.\end{aligned}$$

Así la distribución marginal a posteriori de σ^2 dado $\mathbf{x}$ está dada por

$$\sigma^2|\mathbf{x} \sim IG\left(\frac{n-1}{2}, \frac{\sum_{i=1}^{n}\left(x_i-\overline{\mathbf{x}}\right)^2}{2}\right).$$

Así (4.7) se puede rescribir como:

$$\pi\left(\theta|\mathbf{x}\right) = \int \pi\left(\theta|\sigma^2, \mathbf{x}\right)\pi\left(\sigma^2|\mathbf{x}\right)d\sigma^2.$$

Para este caso, el algoritmo de composición debe realizarse de la siguiente manera:

Entrada: n número de simulación deseada.
$IG\left(\frac{n-1}{2}, \frac{\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})^2}{2}\right)$ Distribución gamma-inversa.
$N\left(\overline{\mathbf{x}}, \frac{\sigma^2}{n}\right)$ Distribución de probabilidad.
Salida: $\left\{\left(\theta_j, \sigma_j^2\right), j = 1, \ldots, n\right\}$ un conjunto de la densidad conjunta (4.8)
Paso 1.- Para $j = 1$ hasta n
Generar $\sigma_j^2 \sim IG\left(\frac{n-1}{2}, \frac{\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})^2}{2}\right)$.
Generar $\theta_j \sim N\left(\overline{\mathbf{x}}, \frac{\sigma_j^2}{n}\right)$.
Paso 2.- **Salida** $\left\{\left(\theta_j, \sigma_j^2\right), j = 1, \ldots, n\right\}$.

Las cantidades θ_j son una muestra *i.i.d.* de la distribución marginal a posteriori $\pi\left(\theta \mid \mathbf{x}\right)$, con las cuales se pueden hacer los cálculos deseados.

Así, si se quiere estimar la media a posteriori de θ

$$E\left(\theta \mid \mathbf{x}\right) = \int \theta\pi\left(\theta \mid \mathbf{x}\right)d\theta \tag{4.9}$$

usando muestreo directo de Monte Carlo, se obtiene una estimación de (4.9) mediante

$$\widehat{E}\left(\theta \mid \mathbf{x}\right) = \frac{1}{M}\sum_{j=1}^{M}\theta_j.$$

Los dos ejemplos anteriores sugieren que, dada una muestra de la distribución a posteriori, casi cualquier cantidad de interés puede ser estimada. Pero, ¿que pasa si no se puede muestrear de la distribución directamente? Este es un viejo problema que precede el interés de los estadísticos bayesianos por muchos años, debido a que hay muchas distribuciones que son difíciles, o casi imposibles de simular directamente, al no disponer de la representación de la distribución en su forma usual, es necesario implementar otro tipo de simulaciones. A continuación se analizaran los métodos indirectos de Monte Carlo.

4.4. Métodos Indirectos

Como resultado al problema de muestrear directamente de la distribución a posteriori, hay diversas aproximaciones que uno podría intentar, de las cuales se analizarán sólo dos:

·Muestreo por importancia.
·Muestreo aceptación-rechazo.

4.4.1. Muestreo por importancia

Se quiere aproximar la esperanza a posteriori dada por

$$\begin{aligned} E\left(g\left(\theta\right) \mid \mathbf{x}\right) &= \int g\left(\theta\right) \pi\left(\theta \mid \mathbf{x}\right) d\theta \\ &= \int g\left(\theta\right) \left[\frac{f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right)}{\int f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right) d\theta}\right] d\theta \\ &= \frac{\int g\left(\theta\right) f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right) d\theta}{\int f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right) d\theta}. \end{aligned}$$

Donde por conveniencia notacional no se asume ninguna dependencia de la función de interés $g\left(\theta\right)$ y la función de verosimilitud $f\left(\mathbf{x} \mid \theta\right)$ en los datos $\mathbf{x}$. En este caso el parámetro θ es unidimensional.

Supóngase que se puede aproximar a la verosimilitud normalizada en términos de la distribución a priori, $cf\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right)$, por alguna densidad $\pi^{*}\left(\theta\right)$ de la cual se puede muestrear fácilmente.

Se define la función de pesos como

$$w\left(\theta\right) = \frac{f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right)}{\pi^{*}\left(\theta\right)}.$$

Teniendo así,

$$\begin{aligned} E\left(g\left(\theta\right) \mid \mathbf{x}\right) &= \frac{\int g\left(\theta\right) w\left(\theta\right) \pi^{*}\left(\theta\right) d\theta}{\int w\left(\theta\right) \pi^{*}\left(\theta\right) d\theta} \\ E\left(g\left(\theta\right) \mid \mathbf{x}\right) &\approx \frac{\frac{1}{M}\sum_{j=1}^{M} g\left(\theta_j\right) w\left(\theta_j\right)}{\frac{1}{M}\sum_{j=1}^{M} w\left(\theta_j\right)}, \end{aligned} \tag{4.10}$$

donde $\theta_j \sim \pi^{*}\left(\theta\right)$. Aquí, $\pi^{*}\left(\theta\right)$ es llamada la **función importante**.

La pregunta ahora es, ¿cómo hacer que la función importante se parezca mucho a $cf\left(\mathbf{x} \mid \theta\right)\pi\left(\theta\right)$ obteniendo una buena aproximación en (4.10)? Para ver esto, note que si $\pi^{*}\left(\theta\right)$ es una buena aproximación, los pesos podrían ser todos más o menos iguales, los cuales deberán minimizar la varianza del numerador y denominador (ver Ripley, 1987 [20]). Si, por otro lado, $\pi^{*}\left(\theta\right)$ no es una buena aproximación, muchos de los pesos podrían acercarse a cero, y entonces pocos θ_{js} podrían dominar la suma, produciendo una aproximación incorrecta. Acontinuación se define el soporte de una distribución, esta definición se necesita en los temas siguientes.

Definición 4.1. *El soporte de una distribución es el conjunto cerrado más pequeño cuyo complemento tiene probabilidad cero.*

Una de las técnicas más sencillas de reducción de la varianza del estimador consiste en elegir una distribución de muestreo por importancia adecuada. Generalmente $\pi^{*}\left(\theta\right)$ debe:

a) Ser fácil de simular.

b) Tener una forma similar a la de $f\left(\mathbf{x} \mid \theta\right)\pi\left(\theta\right)$, la función que se desea integrar.

c) Tener las colas más pesadas que $f\left(\mathbf{x} \mid \theta\right)\pi\left(\theta\right)$, pues de otra forma la varianza del estimador podría llegar a ser muy grande o incluso infinita. Además debe satisfacer que $sop\left(f\left(\theta \mid \mathbf{x}\right)\pi\left(\theta\right)\right) \subset \pi^{*}\left(\theta\right)$.

Ejemplo 4.4. Sea $X_i \sim N\left(x_i \mid \theta, 1\right)$. La función de verosimilitud es una distribución $N\left(\overline{\mathbf{x}} \mid \theta, \frac{1}{n}\right)$. Y tomando una distribución a priori $\pi\left(\theta\right) \propto 1$, se quiere aproximar

$$E\left(\theta \mid \mathbf{x}, \sigma^2\right) = \frac{\int \theta w\left(\theta\right) \pi^{*}\left(\theta\right) d\theta}{\int w\left(\theta\right) \pi^{*}\left(\theta\right) d\theta},$$

$$\widehat{E}\left(\theta \mid \mathbf{x},\sigma^2\right) \approx \frac{\sum_{j=1}^{m} \theta_j w\left(\theta_j\right)}{\sum_{j=1}^{m} w\left(\theta_j\right)},$$

donde los pesos, tomando como función importante una distribución t de Student con n grados de libertad, son

$$w\left(\theta_i\right) = \frac{N\left(\overline{\mathbf{x}} \mid \theta_i, \frac{1}{n}\right) \cdot 1}{t_n\left(\theta_i\right)}.$$

Ahora aplicando muestreo por importancia con $m = 1000$ (ver apéndice A4.5) se obtiene que

$$\widehat{E}\left(\theta \mid \mathbf{x},\sigma^2\right) = 0.0220,$$

mientras que la verdadera media es

$$\overline{\mathbf{x}} = 0.0224.$$

El siguiente método, sólo requiere del conocimiento de la forma funcional de la densidad que se quiere aproximar, y una constante multiplicativa. Además se usa una densidad $\pi^*\left(\theta\right)$ de la que es fácil simular.

4.4.2. Muestreo aceptación-rechazo

En este método, en lugar de intentar aproximar a la distribución a posteriori

$$\pi\left(\theta \mid \mathbf{x}\right) = \frac{\pi\left(\theta\right) f\left(\mathbf{x} \mid \theta\right)}{\int \pi\left(\theta\right) f\left(\mathbf{x} \mid \theta\right) d\theta},$$

se intenta envolver ésta.

Esto es, supóngase que existe una constante $k > 0$ y una densidad suave $g\left(\theta\right)$, llamada la **función envoltura**, tal que

$$f\left(\mathbf{x} \mid \theta\right) \pi\left(\theta\right) < k g\left(\theta\right),$$

para todo θ. El método de aceptación-rechazo es:

Entrada: m tamaño deseado de la muestra.
$g(\theta)$ función envoltura.
$f(\mathbf{x} \mid \theta)$ función de verosimilitud.
$\pi(\theta)$ distribución a priori.
Salida: $\{\theta_i, i = 1, ..., m\}$ variables aleatorias con distribución $\pi(\theta \mid \mathbf{x})$
Paso 1.- $j = 1$.
Paso 2.- Mientras$j \leq m$
Paso 3.- Generar $\theta^* \sim g(\theta)$
Paso 4.- Generar $U \sim U(0, 1)$
Paso 5.- Si $kUg(\theta^*) < f(\mathbf{x} \mid \theta^*)\pi(\theta^*)$,
aceptar $\theta_j = \theta^*$
$j = j + 1$
en otro caso
rechazar θ^*.
Paso 6.- **Salida** $\{\theta_j, j = 1, ..., m\}$

Los elementos de esta muestra $\{\theta_j, j = 1, ..., m\}$ son variables aleatorias que se distribuyen con $\pi(\theta \mid \mathbf{x})$.

Esto permite muestrear indirectamente de $\pi(\theta \mid \mathbf{x})$ si su forma funcional se conoce y es fácil obtener una muestra aleatoria de una distribución que se aproxima a ella, $g(\theta)$. Una sugerencia intuitiva es tomar k tan pequeña como sea posible, para no perder muestras innecesariamente.

Si la distribución a priori es propia, podríamos utilizarla como función pivote, $\pi^*(\theta) = \pi(\theta)$ y como constante $k^* = sop_\theta f(\mathbf{x}|\theta)$.

Ejemplo 4.5. Tomando el ejemplo 3.1, donde la función de verosimilitud es una distribución Binomial, la distribución a priori es una distribución Beta. En el ejemplo 3.1 se toma a un elemento de la familia Beta con parámetros $\alpha = \beta = 1$, teniendo así que

$$f(\mathbf{x} \mid \theta)\pi(\theta) \propto \theta^x (1 - \theta)^{n-x}.$$

Se tiene que encontrar un $k > 0$ y una densidad $g(\theta)$ tal que

$$\theta^x (1 - \theta)^{n-x} < kg(\theta).$$

Tomando como función envoltura la aproximación normal obtenida en el ejemplo

3.1 se tiene que la distribución $g(\theta)$ es

$$g(\theta) = N\left(\widehat{\theta}, \frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{n}\right).$$

Para calcular el valor de k que envuelva a $f(\mathbf{x} \mid \theta)\pi(\theta)$ se tiene que acotar a

$$\frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{g(\theta)},$$

para ello se maximiza el siguiente cociente:

$$\begin{aligned}\frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{g(\theta)} &\propto \frac{\theta^x(1-\theta)^{n-x}}{\exp\left(-\frac{1}{2\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\left(\theta-\widehat{\theta}\right)^2\right)} \\ &= \theta^{13}(1-\theta)^3\exp\left(\frac{16^3}{2\cdot 3\cdot 13}\left(\theta-\frac{13}{16}\right)^2\right) \qquad (4.11)\end{aligned}$$

derivando e igualando a cero se obtiene que $\widehat{\theta} = \frac{13}{16}$, ahora se evalúa en (4.11) y se obtiene que $k = 0.00044$.

Se implementa el método de aceptación-rechazo para este ejemplo, obteniendo la aproximación mostrada en la tabla 4.1 (ver apéndice A4.6):

	A posteriori	Aceptación-rechazo
Media	0.7777	0.7802
Varianza	0.0090	0.0079

Tabla 4.1. Comparación de la media y varianza por aceptación rechazo.

En la figura 4.2 se muestra en un histograma la aproximación por el método aceptación-rechazo y la distribución a posteriori exacta con una curva continua.

Cuando el parámetro $\boldsymbol{\theta}$ es un vector los métodos de simulación Monte Carlo vía cadenas de Markov.son más fáciles de implementar, además se obtiene una simulación de la distribución a posteriori y con ella se pueden hacer las estimaciones deseadas. De éstos métodos hablaremos en el siguiente capítulo.

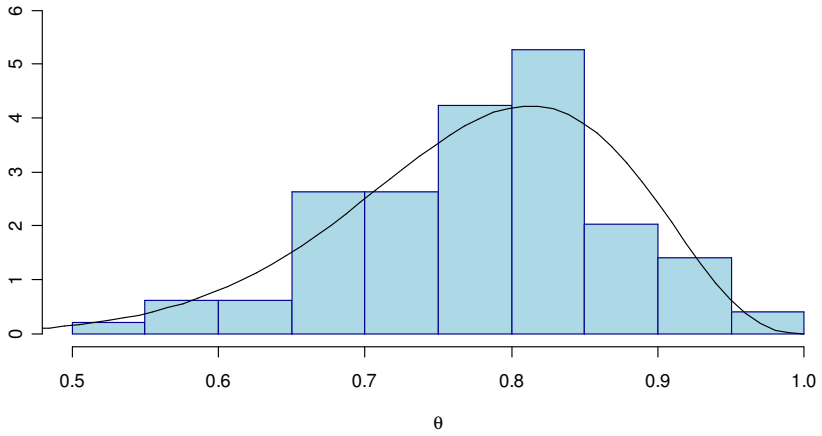

Figura 4.2: Aproximación a la distribución a posteriori por el método aceptación-rechazo para el ejemplo 4.5.

Capítulo 5

Simulación Monte Carlo vía Cadenas de Markov

5.1. Preliminares

En la práctica suceden fenómenos que evolucionan con el tiempo, de una manera aleatoria. Estos son llamados procesos estocásticos.

Definición 5.1. *Un **proceso estocástico** es una familia de variables aleatorias* $\{X_t\}_{t\in T}$ *indexada por un conjunto* T *que representa el tiempo y para cada tiempo* $t_0 \in T$, X_{t_0} *es el estado del proceso en el tiempo* t_0. *A todo el conjunto de estados posibles se le llama **espacio de estados**.*

Definición 5.2. *Una **cadena de Markov** es un tipo especial de proceso estocástico, que se describe como sigue. En cualquier instante de tiempo* t *dado, cuando el estado actual* X_t *y todos los estados previos* $X_1, \ldots, X_{t-1}$ *del proceso son conocidos, las probabilidades de los estados futuros* $X_j\,(j > t)$ *dependen solamente del estado actual* X_t *y no dependen de los estados anteriores* $X_1, \ldots, X_{t-1}$. *Es decir, una cadena de Markov es un proceso estocástico tal que para* $t = 1, 2, \ldots$ *y para cualquier sucesión posible de estados* $x_1, x_2, \ldots, x_{t+1}$,

$$P\left(X_{t+1} = x_{t+1} \mid X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t\right) = P\left(X_{t+1} = x_{t+1} \mid X_t = x_t\right).$$

Definición 5.3. *Una cadena de Markov para la cual existe sólo un número finito* k *de estados posibles* $\{s_1, \ldots, s_k\} = S$ *(*S *espacio de estados) y en cualquier instante de tiempo la cadena debe estar en uno de estos* k *estados, se denomina **cadena de Markov finita.***

Definición 5.4. *La probabilidad condicional* $P(X_{t+1} = s_j | X_t = s_i) = p_{ij}$ *de que la cadena de Markov esté en el estado* s_j *en el instante de tiempo* $t+1$ *si está en el estado* s_i *en el instante de tiempo* t *se denomina* ***probabilidad de transición.***

Definición 5.5. *Si para una cadena de Markov esta probabilidad de transición tiene el mismo valor para todos los instantes de tiempo* t $(t = 1, 2, \ldots)$*, se dice que la cadena de Markov tiene* ***probabilidades de transición estacionarias.***

Se define

$$\mu_j(t) = P(X_t = s_j),$$

como la probabilidad de que la cadena esté en el estado j al tiempo t. Y $\mu(t)$ denota el vector fila del espacio de estados de probabilidades al paso t. Se inicia la cadena por especificación de un vector inicial $\mu(0)$, frecuentemente todos los elementos de $\mu(0)$ son cero excepto un elemento que es 1, correspondiente al inicio del proceso en un estado particular.

La probabilidad de que la cadena tenga un valor en el estado s_i al tiempo (o paso) $t+1$ está dado por la igualdad de Chapman-Kolmogorov.

Teorema 5.1. *(**Igualdad de Chapman-Kolmogorov**) La suma sobre las probabilidades de estar en un estado particular en el actual paso y la probabilidad de transición de que el estado esté en el estado* s_i*, están relacionadas por la fórmula recursiva*

$$\mu_i(t+1) = \sum_k p_{ki}\mu_k(t).$$

Demostración. Para verificar la fórmula aplicamos la probabilidad total

$$\begin{aligned}
\mu_i(t+1) &= P(X_{t+1} = s_i) \\
&= \sum_k P(X_{t+1} = s_i | X_t = s_k) P(X_t = s_k) \\
&= \sum_k p_{ki}\mu_k(t)
\end{aligned}$$

□

Sucesivas iteraciones de la igualdad de Chapman-Kolmogorov describen la evolución de la cadena.

Se puede escribir más compactamente la igualdad de Chapman Kolomogrov en forma de matriz.

Definición 5.6. *La **matriz de probabilidad de transición P** se define como la matriz cuyo elemento i, j es p_{ij}, la probabilidad de moverse del estado i al estado j, (esto implica que las filas suman uno, $\sum_j p_{ij} = 1$).*

La igualdad de Chapman-Kolmogorov se reescribe como:

$$\mu(t+1) = \mu(t)\,\mathbf{P},$$

usando la forma de la matriz, se ve como se relaciona la igualdad de Chapman-Kolmogorov

$$\mu(t) = \mu(t-1)\,\mathbf{P} = (\mu(t-2)\,\mathbf{P})\,\mathbf{P} = \mu(t-2)\,\mathbf{P}^2,$$

continuando, se tiene que

$$\mu(t) = \mu(0)\,\mathbf{P}^t.$$

Definiendo para el $n-$ésimo paso la probabilidad de transición $p_{ij}^{(n)}$ como la probabilidad que el proceso esté en el estado j dado que éste inició en el estado i hace n pasos, esto es,

$$p_{ij}^{(n)} = P(X_{t+n} = s_j | X_t = s_i),$$

que es justamente el elemento $ij-$ésimo de $\mathbf{P}^n$.

La evolución de una cadena es descrita por sus probabilidades de transición, $P(X_{t+1} = s_j | X_t = s_i)$, lo cual puede ser un poco complicado, dado que sus probabilidades dependen de tres cantidades t, j e i. Cuando las probabilidades no dependen de t sólo de i y j se tiene la siguiente definición.

Definición 5.7. *La cadena* $\mathbf{X}$ *es llamada **homogénea** si*

$$P(X_{t+1} = s_j | X_t = s_i) = P(X_1 = s_j | X_0 = s_i).$$

Definición 5.8. *Una cadena de Markov se dice que es **irreducible** si existe un número positivo tal que $p_{ij}^{(n_{ij})} > 0$ para todo i, j. Esto es, todos los estados se comunican con los otros, es decir, uno puede ir de un estado a cualquier otro estado.*

Definición 5.9. *El período $d(i)$ del estado $i \in S$ es definido por*

$$d(i) = mcd\{n \geq 0 : p_{ii}^n > 0\},$$

el máximo común divisor del estado para el cual regresar es posible. Se le llama al estado i periódico si $d(i) > 1$, y aperiódico si $d(i) = 1$. [11].

Definición 5.10. *Una cadena de Markov es aperiódica si $\forall\, i \in S$, i es aperiódico.*

Definición 5.11. *Una cadena de Markov alcanza una* ***distribución estacionaria*** π^**, cuando el vector de probabilidades es, en particular, un estado independientemente de la condición inicial. La distribución estacionaria satisface*

$$\pi^* = \pi^* \mathbf{P}.$$

Las condiciones para una distribución estacionaria es que la cadena sea ***irreducible*** *y* ***aperiódica****. Una condición suficiente para una única distribución estacionaria es que la cadena de Markov sea* ***reversible****, esto es*

$$P(j \to k)\, \pi_j^* = P(k \to j)\, \pi_k^*,$$

que es la condición de reversibilidad.

El rango de aplicación de las cadenas de Markov es bastante vasto, incluyendo cualquier sistema dinámico que evoluciona sobre tiempos que involucran incertidumbre. Estos sistemas se aplican en variedad de campos, tal como comunicación, control automático, manufactura, economía, asignación de recursos, etc. Para entender más el concepto de cadena de Markov observe el siguiente ejemplo.

Ejemplo 5.1. ([26]) Supóngase que el espacio de estados consta de tres posibilidades: lluvioso, soleado y nublado, y que el clima sigue un proceso de Markov. Esto es, la probabilidad del clima de mañana depende del clima de hoy, y no de cualquier otro día previo. supóngase además que la probabilidad de transición dado que hoy está lloviendo es:

$$\begin{aligned} P(\textit{mañana llueva}|\textit{hoy llueve}) &= 0.5, \\ P(\textit{mañana soleado}|\textit{hoy llueve}) &= 0.25, \\ P(\textit{mañana soleado}|\textit{hoy llueve}) &= 0.25. \end{aligned}$$

La primera fila de la matriz de transición de probabilidad es $(0.5,\ 0.25,\ 0.25)$. supóngase que el resto de la matriz de transición está dada por

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix},$$

esta cadena de Markov es irreducible, pues todo estado se comunican con los otros estados.

Si se supone que hoy es soleado. ¿Cuál es la probabilidad de que el clima se encuentre en alguno de los estados dentro de dos días? ¿Y en siete días? Aquí

$\mu(0) = (0\ 1\ 0)$, entonces

$$\begin{aligned}\mu(2) &= \mu(0)\,\mathbf{P}^2 \\ &= (0\ 1\ 0)\begin{pmatrix} 0.4375 & 0.1875 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.1875 & 0.4375 \end{pmatrix} \\ &= (0.375\ 0.25\ 0.375)\end{aligned}$$

y

$$\begin{aligned}\mu(7) &= \mu(0)\,\mathbf{P}^7 \\ &= (0\ 1\ 0)\begin{pmatrix} 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.4 \end{pmatrix} \\ &= (0.4\ 0.2\ 0.4)\,.\end{aligned}$$

Ahora si se supone que hoy está lloviendo, entonces $\mu(0) = (1\ 0\ 0)$. La probabilidad de que el clima se encuentre en alguno de los estados está dado por

$$\mu(2) = \begin{pmatrix} 0.4375 & 0.1875 & 0.375 \end{pmatrix} \quad \text{y} \quad \mu(7) = \begin{pmatrix} 0.4 & 0.2 & 0.4 \end{pmatrix}.$$

Note que después de un tiempo suficientemente largo el clima esperado es independiente del valor inicial. En otras palabras la cadena alcanza una distribución estacionaria.

La idea básica de una cadena de Markov en estado discreto puede generalizarse para un proceso de Markov en estado continuo, teniendo una probabilidad que satisface

$$\int P(x,y)\,dy = 1,$$

y la igualdad de Chapman-Kolomogrov esta dada por

$$\mu_t(y) = \int \mu_{t-1}(x)\,P(x,y)\,dy,$$

para equilibrar, la distribución estacionaria satisface

$$\mu^*(y) = \int \mu^*(x)\,P(x,y)\,dy.$$

Ahora se analizarán los métodos de simulación Monte Carlo vía cadenas de Markov.

5.2. Métodos de simulación Monte Carlo

Nos enfrentamos a un problema cuando queremos calcular

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\pi(\boldsymbol{\theta})\, f(\mathbf{x}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})\, f(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}},$$

pues el denominador no adopta una forma funcional conocida y la evaluación generalmente no es posible de forma analítica. Se hace necesario el tratamiento numérico, que se agrava en muchos casos porque la dimensión del espacio paramétrico es mayor que uno, es decir $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.

Si fuera posible generar directamente muestras independientes de $\pi(\boldsymbol{\theta}|\mathbf{x})$ mediante algún método aleatorio de simulación, conduciría a la obtención de la cantidad a posteriori de interés. Pero en muchos casos no es posible simular directamente de muestras independientes para $\pi(\boldsymbol{\theta}|\mathbf{x})$. Sin embargo, puede ser factible simular muestras con algún tipo de dependencia, que converjan (bajo ciertas condiciones de regularidad) a la distribución de interés $\pi(\boldsymbol{\theta}|\mathbf{x})$.

Nos enfrentamos a un problema al aplicar integración Monte Carlo para obtener muestras de la misma complejidad que la distribución objetivo. Intentar resolver este problema es la raíz de los métodos Monte Carlo vía cadenas de Markov.

En muchos problemas, especialmente de alta dimensión, puede ser un poco difícil o imposible encontrar una densidad que sea fácil de muestrear.

Así, los métodos basados en simulación Monte Carlo mediante cadenas de Markov (MCMC) permiten muestrear la distribución a posteriori, siempre y cuando se conozca la forma funcional de la distribución aunque la constante de normalización sea desconocida, gracias a la construción de una cadena de Markov cuya distribución estacionaria es precisamente la distribución a posteriori. A la distribución a posteriori se le llama distribución objetivo.

La clave de la simulación con cadenas de Markov es crear una cadena de Markov cuya distribución estacionaria es $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. Se corre la simulación un tiempo suficientemente largo tal que la distribución generada en la iteración actual sea lo suficientemente cerca de la distribución estacionaria. Una discusión general de estos métodos puede verse en Smith y Roberts (1993) [22].

Proposición 5.1. *Sea $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ una cadena de Markov homogénea, irreducible y aperiódica, con espacio de estados Θ y distribución de equilibrio $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. Entonces conforme $t \to \infty$,*

(i) $\boldsymbol{\theta}_t \xrightarrow{D} \boldsymbol{\theta}$, *donde* $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta} \mid \mathbf{x})$;

(ii) $\frac{1}{t} \sum_{i=1}^{t} g(\boldsymbol{\theta}_i) \to E(g(\boldsymbol{\theta}) \mid \mathbf{x})$.

Un método de similación Monte Carlo vía cadenas de Markov es el algoritmo de Metropolis-Hastings, el cual se analiza a continuación.

5.2.1. Algoritmo de Metropolis-Hastings

El algoritmo de Metropolis-Hastings es un método Monte Carlo vía cadenas de Markov que ha sido ampliamente usado en la física matemática y restauración de images por muchos años, pero recientemente descubierto por los estadísticos.

El método en su forma más simple lo publica Metropolis (1953), el cual se explicará a continuación. Supóngase que la verdadera distribución a posteriori conjunta de un parámetro $\boldsymbol{\theta}$ (posiblemente vector de valores) tiene densidad $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. Se escoge una función auxiliar $q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$, tal que $q(\cdot \mid \boldsymbol{\theta}_t)$ es una función de densidad de probabilidad para toda $\boldsymbol{\theta}_t$, además es simétrica, es decir:

$$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*).$$

La función q es llamada distribución de densidad propuesta o candidato-generadora. Entonces, el algoritmo de Metropolis genera una cadena de Markov como sigue:

Entrada: Conocer la forma funcional de $\pi(\boldsymbol{\theta} \mid \mathbf{x})$.
Dar un valor arbitrario $\boldsymbol{\theta}_0 \in sop(\pi(\boldsymbol{\theta} \mid \mathbf{x}))$, tal que $\pi(\boldsymbol{\theta}_0 \mid \mathbf{x}) > 0$.
T tiempo al cual se alcanza la convergencia.
$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ distribución de densidad candidato-generadora.
Salida: $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k, ...)$.
Paso 1.- Para $t = 0$ hasta T
Paso 2.- Generar una observación $\boldsymbol{\theta}^*$ de $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_t)$
Paso 3.- Generar una variable $u \sim U(0,1)$
Paso 4.- Dado el punto candidato $\boldsymbol{\theta}^*$, calcular

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t) = \text{mín}\left(\frac{\pi(\boldsymbol{\theta}^* \mid \mathbf{x})}{\pi(\boldsymbol{\theta}_t \mid \mathbf{x})}, 1\right)$$

Paso 5.- Si $u \leq \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t)$,
hacer $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^*$,
en caso contrario
hacer $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$.
Paso 6.- **Salida** $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k, ...)$

Esto genera una cadena de Markov $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k, ...)$, donde la probabilidad de transición de $\boldsymbol{\theta}_t$ a $\boldsymbol{\theta}_{t+1}$ depende sólo de $\boldsymbol{\theta}_t$ y no de $\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_{t-1}$. Y cuya distribución de transición es

$$P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) = \alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t)\, q(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t).$$

La cadena se aproxima a una distribución estacionaria y las muestras del vector $(\boldsymbol{\theta}_{k+1}, ..., \boldsymbol{\theta}_{k+n})$ son muestras de la densidad objetivo $\pi(\boldsymbol{\theta} \mid \mathbf{x})$.

Note que la densidad a posteriori conjunta π es necesaria sólo en una constante de proporcionalidad. Dado que en aplicaciones bayesianas

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) \propto f(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}),$$

una forma que es típicamente disponible.

Hastings (1970) generaliza el algoritmo de Metropolis usando una función arbitraria de probabilidad de transición (candidato-generadora) y aceptando la probabilidad de un punto candidato como

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t) = \text{mín}\left(\frac{\pi(\boldsymbol{\theta}^* \mid \mathbf{x}) q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_t \mid \mathbf{x}) q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_t)}, 1\right),$$

este es el **algoritmo de Metropolis-Hastings**. Si la distribución propuesta es simétrica, se recupera el algoritmo de Metropolis original.

A continuación se muestra un ejemplo para aproximar una distribución a posteriori por el algoritmo de Metropolis.

Ejemplo 5.2. Supóngase que $X_i \sim f(x_i \mid \theta)$ para $i = 1, \ldots, n, \theta > 0$ y $x_i \geq 0$. Entonces

$$f(x_i \mid \theta) = \theta \exp(-\theta x_i).$$

La función de verosimilitud está dada por

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i) = \theta^n \exp\left[-\theta\left(\sum_{i=1}^{n} x_i\right)\right].$$

Se toma una distribución a priori

$$\begin{aligned} \pi(\theta) &= Ga(\theta \mid \alpha, \beta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta), \end{aligned}$$

entonces

$$\begin{aligned} f(\mathbf{x} \mid \theta) \pi(\theta) &= \theta^n \exp\left[-\theta\left(\sum_{i=1}^{n} x_i\right)\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta), \\ f(\mathbf{x} \mid \theta) \pi(\theta) &\propto \theta^{(n+\alpha)-1} \exp\left[-\theta\left(\sum_{i=1}^{n} x_i + \beta\right)\right], \end{aligned}$$

dado que

$$\pi\left(\boldsymbol{\theta} \mid \mathbf{x}\right) \propto f\left(\mathbf{x} \mid \boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right),$$

$$\pi\left(\boldsymbol{\theta} \mid \mathbf{x}\right) \propto \theta^{(n+\alpha)-1} \exp\left[-\theta\left(\sum_{i=1}^{n} x_i + \beta\right)\right].$$

Se toma como distribución candidato-generadora una distribución uniforme en $(0, 100)$ (esta distribución es simétrica). Claramente, los valores arriba de 100 tienen una probabilidad, pero se supone, para este caso, que ésta es suficientemente pequeña, entonces se puede ignorar.

Se corre el algoritmo:

Entrada. Iniciar con $\theta_0 = 1$ como valor inicial,

1.- Para $t = 1$ hasta 500

2. Se genera un valor candidato θ^* de la uniforme, la distribución candidato-generadora.

3. Se genera una variable U de una distribución uniforme en $(0, 1)$.

4. Dado el punto candidato θ^*, se calcula

$$\alpha\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t\right) = \text{mín}\left(\frac{\left((\theta^*)^{(n+\alpha)-1} \exp\left[-\theta^*\left(\sum_{i=1}^{n} x_i + \beta\right)\right]\right)}{\left(\theta_t^{(n+\alpha)-1} \exp\left[-\theta_t\left(\sum_{i=1}^{n} x_i + \beta\right)\right]\right)}, 1\right)$$

5. Si se cumple que $u \leq \alpha\left(\theta^*, \theta_t\right)$, se hace $\theta_1 = \theta^*$. En caso contrario se hace $\theta_1 = \theta_0$, y regresar al paso 1.

6. 500 valores de θ.

El resultado de los primeros 500 valores de θ son graficados en la figura 5.1, ver apéndice A5.1.

En la figura 5.1 se observan largos períodos planos (correspondientes a todos los valores de θ^* que son rechazados), tal cadena es llamada **mala mezcla**.

En contraste con el ejemplo anterior se supone que la distribución propuesta es una distribución χ^2, y se aplica el algoritmo de Metropolis-Hastings.

Ejemplo 5.3. Supóngase que se usa una distribución χ^2 como la densidad candidato. Si $\theta \sim \chi^2\left(\theta \mid \nu\right)$, entonces

$$\chi^2\left(\theta \mid \nu\right) \propto \theta^{\frac{\nu}{2}-1} \exp\left(-\frac{\theta}{2}\right).$$

Se tiene que $q\left(\theta^2 \mid \theta^1\right) = C\left(\theta^2\right)^{\frac{\theta^1}{2}-1} \exp\left(-\frac{\theta^2}{2}\right)$. Note que $q\left(\theta^2 \mid \theta^1\right)$ no es simétrica, esto es, $q\left(\theta^1 \mid \theta^2\right) \neq q\left(\theta^2 \mid \theta^1\right)$. Aquí se usa el algoritmo de Metropolis-

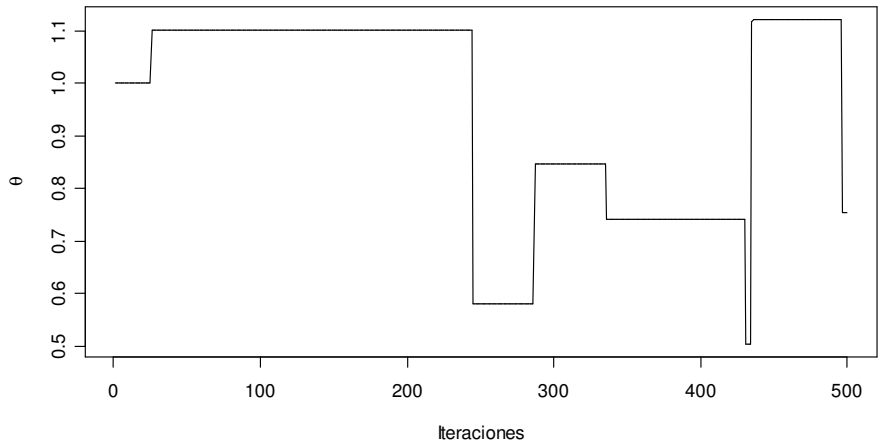

Figura 5.1: Valores de theta obtenidos del ejemplo 5.2 aplicando el algoritmo de Metropolis.

Hastings, con probabilidad de aceptación

$$\alpha(y,x) = \text{mín}\left[\frac{\pi(\theta^* \mid \mathbf{x})\, q(\theta_t \mid \theta^*)}{\pi(\theta_t \mid \mathbf{x})\, q(\theta^* \mid \theta_t)}, 1\right] = \text{mín}\left[\frac{\pi(\theta^* \mid \mathbf{x})\,(\theta_t)^{\frac{\theta^*}{2}-1}\exp\left(-\frac{\theta_t}{2}\right)}{\pi(\theta_t \mid \mathbf{x})\,(\theta^*)^{\frac{\theta_t}{2}-1}\exp\left(-\frac{\theta^*}{2}\right)}, 1\right].$$

Usando la misma distribución objetivo del ejemplo anterior,

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = C\theta^{(n+\alpha)-1}\exp\left[-\theta\left(\sum_{i=1}^{n} x_i + \beta\right)\right]$$

se tiene que la probabilidad de rechazo es

$$\alpha(\theta^*, \theta_t) = \text{mín}\left[\frac{\left((\theta^*)^{(n+\alpha)-1}\exp\left[-\theta^*\left(\sum_{i=1}^{n} x_i + \beta\right)\right]\right)\left((\theta_t)^{\frac{\theta^*}{2}-1}\exp\left(-\frac{\theta_t}{2}\right)\right)}{\left(\theta_t^{(n+\alpha)-1}\exp\left[-\theta_t\left(\sum_{i=1}^{n} x_i + \beta\right)\right]\right)\left((\theta^*)^{\frac{\theta_t}{2}-1}\exp\left(-\frac{\theta^*}{2}\right)\right)}, 1\right],$$

usando la distribución ji-cuadrada con dos grados de libertad, la cual tiene una varianza mínima y una probabilidad de aceptación alta. En la figura 5.2 se muestran los 1000 primeros valores de θ (ver apéndice A5.2).

En la figura 5.2 se observa una serie de tiempo, la cual da la apariencia de ser un ruido blanco, esto es, el proceso consiste de una sucesión de variables aleatorias las cuales son independientes e idénticamente distribuidas; así se dice que la cadena es una **buena mezcla**.

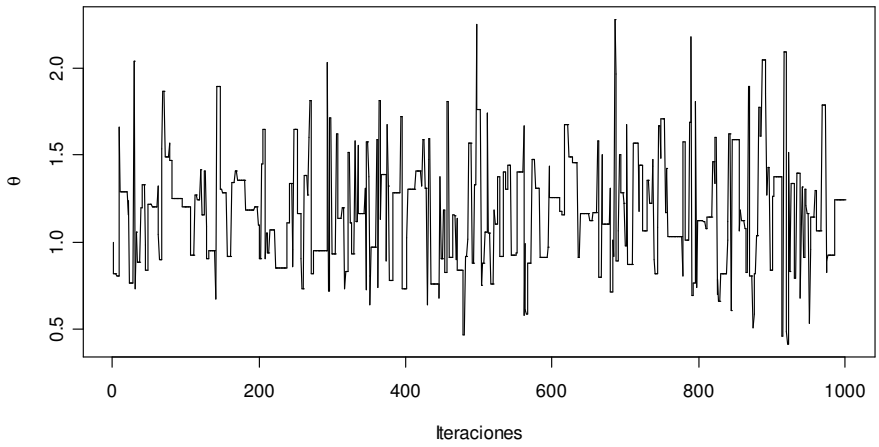

Figura 5.2: Valores de theta para el ejemplo 5.3 aplicando el algoritmo de Metropolis-Hastings.

La idea de una distribución candidato en el algoritmo Metropolis-Hastings es muestrear puntos candidatos de la distribución objetivo, esto es $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) \equiv \pi(\boldsymbol{\theta}^* \mid \mathbf{x})$ para todo $\boldsymbol{\theta}$. Pero como las simulaciones se aplican a problemas que no se puede muestrear directamente de $\pi(\boldsymbol{\theta}^* \mid \mathbf{x})$, entonces una buena distribución candidato tiene las siguientes propiedades.

- Para cualquier $\boldsymbol{\theta}$, es fácil muestrear de $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$.
- Es fácil calcular el cociente

$$\frac{\pi(\boldsymbol{\theta}^* \mid \mathbf{x})\, q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_t \mid \mathbf{x})\, q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_t)}.$$

- Cada salto va a distancias razonables en el espacio de parámetros.
- Los saltos no son rechazados con frecuencia.
- En el algoritmo de Metropolis-Hastings se debe cumplir que $sop(\pi(\boldsymbol{\theta} \mid \mathbf{x})) \subset \bigcup_{\theta \in sop(\pi(\boldsymbol{\theta}|\mathbf{x}))} sop(q(\cdot \mid \boldsymbol{\theta}))$.

Un algoritmo particular con cadenas de Markov que ha sido útil en muchos problemas multidimensionales es el muestreo de Gibbs, el cual se analiza a continuación.

5.2.2. Muestreo de Gibbs

El muestreo de Gibbs (Geman y Geman, 1984 [11]; Gelfand y Smith, 1990 [12]) es una técnica para generar variables aleatorias de distribuciones marginales indirectas sin tener que calcular la densidad y que en la mayoría de los casos tienen una forma más sencilla que la de $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. El muestreo de Gibbs está basado en propiedades elementales de cadenas de Markov y se considera el más simple de los métodos de muestreo de cadenas de Markov.

En éste método se supone que el vector de parámetros $\boldsymbol{\theta}$ ha sido dividido en d componentes, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ y se supone que las distribuciones condicionales completas $\{\pi(\theta_i \mid \theta_{1,t+1}, \ldots, \theta_{i-1,t+1}, \theta_{i+1,t}, \ldots, \theta_{d,t}, \mathbf{x}), i = 1, \ldots, d\}$ (donde π denota una función de densidad) estan disponibles (esto significa que las muestras pueden ser generadas por algún método dando valores aproximados de variables aleatorias condicionales) para ser muestreadas.

Estas distribuciones condicionales completas únicamente determinan la distribución conjunta completa $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ y de aquí todas las distribuciones marginales $\pi(\theta_i \mid \mathbf{x})$. Entonces el método básico de muestreo de Gibbs simula una cadena de Markov en la que $\boldsymbol{\theta}_{t+1}$ se obtiene a partir de $\boldsymbol{\theta}_t$ de la siguiente manera:

Entrada: Valor inicial arbitrario $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{d,0})'$
T tiempo al cual se alcanza la convergencia.
$\pi(\theta_i \mid \theta_{1,t+1}, \ldots, \theta_{i-1,t+1}, \theta_{i+1,t}, \ldots, \theta_{d,t}, \mathbf{x})$ distribuciones condicionales completas.

Salida: $\{\boldsymbol{\theta}_t, t = 1, 2, \ldots\}$

Paso 1.-Para $t = 0$ hasta T

Paso 2.- Generar $\boldsymbol{\theta}_{t+1} = (\theta_{1,t+1}, \theta_{2,t+1}, \ldots, \theta_{d,t+1})'$ como sigue:

- Generar $\theta_{1,t+1} \sim \pi(\theta_1 \mid \theta_{2,t}, \ldots, \theta_{d,t}, \mathbf{x})$
- Generar $\theta_{2,t+1} \sim \pi(\theta_2 \mid \theta_{1,t+1}, \theta_{3,t}, \ldots, \theta_{d,t}, \mathbf{x})$

$\ldots \quad \ldots \quad \ldots$

- Generar $\theta_{d,t+1} \sim \pi(\theta_d \mid \theta_{1,t+1}, \theta_{2,t+1}, \ldots, \theta_{d-1,t+1}, \mathbf{x})$

Paso 3.- **Salida** $\{\boldsymbol{\theta}_t, t = 1, 2, \ldots\}$

Entonces, cada componente de $\boldsymbol{\theta}$ es visitado en orden natural y un ciclo de este método requiere generar d variables aleatorias. Así, la sucesión de vectores $\{\boldsymbol{\theta}_t, t = 1, 2, \ldots\}$ es una realización de una cadena de Markov cuya distribución de transición está dada por

$$P(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) = \prod_{i=1}^{d} \pi(\theta_{i,t+1} \mid \theta_{1,t+1}, \theta_{2,t+1}, \ldots, \theta_{i-1,t+1}, \theta_{i+1,t}, \ldots, \theta_{d,t}, \mathbf{x})$$

y tiene distribución estacionaria $\pi(\boldsymbol{\theta} \mid \mathbf{x})$.

Lo que se requie para obtener muestras de la distribución conjunta de $(\theta_1, \ldots, \theta_d)$ es la capacidad de muestrear de las correspondientes d distribuciones condicionales completas. En el contexto del análisis bayesiano, la distribución conjunta de interés es la distribución conjunta a posteriori $\pi(\theta_1, \ldots, \theta_d \mid \mathbf{x})$, o quizás una o más de las distribuciones marginales a posteriori, $\pi(\theta_i \mid \mathbf{x})$.

Para entender mejor el trabajo del muestreo de Gibbs, a continuación se explorá el caso de dos variables. Iniciando con un par de variables aleatorias (X, Y).

Ejemplo 5.4. ([7]). Para la siguiente distribución conjunta de X y Y,

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \quad x = 0, 1, \ldots, n \quad 0 \leq y \leq 1, \tag{5.1}$$

supóngase que interesa calcular alguna característica de la distribución marginal $f(x)$ de X. El muestreo de Gibbs permite generar una muestra de esta distribución marginal como sigue: de (5.1) se tiene (omitiendo toda dependencia en n, α, y β) que

$$\begin{aligned} x &\mid y \sim Bi(n, y) \\ y &\mid x \sim Be(x+\alpha, n-x+\beta) \end{aligned}$$

Aplicando el muestreo de Gibbs (tomando a $\theta_1 = X$ y $\theta_2 = Y$) se puede generar una muestra de $X_1, \ldots, X_m$ de $f(x)$ de la siguiente manera.

Entrada: Valor inicial arbitrario $\theta_{2,0}$
T simulación deseada
$Bi(n, \theta_{2,t})$
$Be(\theta_{1,t}+\alpha, n-\theta_{1,t}+\beta)$
Salida: $\{(\theta_{1,t}, \theta_{2,t})\}$ $t = 0, ..., T$.
Paso 1.- Para $t = 0$ hasta T
Paso 2. Generar $\boldsymbol{\theta}_{t+1} = (\theta_{1,t+1}, \theta_{2,t+1})'$
- Generar $\theta_{1,t+1} \sim Bi(n, \theta_{2,t})$
- Generar $\theta_{2,t+1} \sim Be(\theta_{1,t+1}+\alpha, n-\theta_{1,t+1}+\beta)$

Paso 3.- **Salida** $\{(\theta_{1,t}, \theta_{2,t})\}$ $t = 0, ..., T$.

Se usa esta muestra para estimar cualquier característica deseada. Tomando $n = 16$, $\alpha = 2$ y $\beta = 4$, así se tiene una muestra de tamaño 500 la cual se obtuvo al tomar los valores finales del algoritmo de Gibbs con $j = 10$ iteraciones cada cadena. En la figura 5.3 se muestran los valores de θ_1 (ver apéndice A5.3).

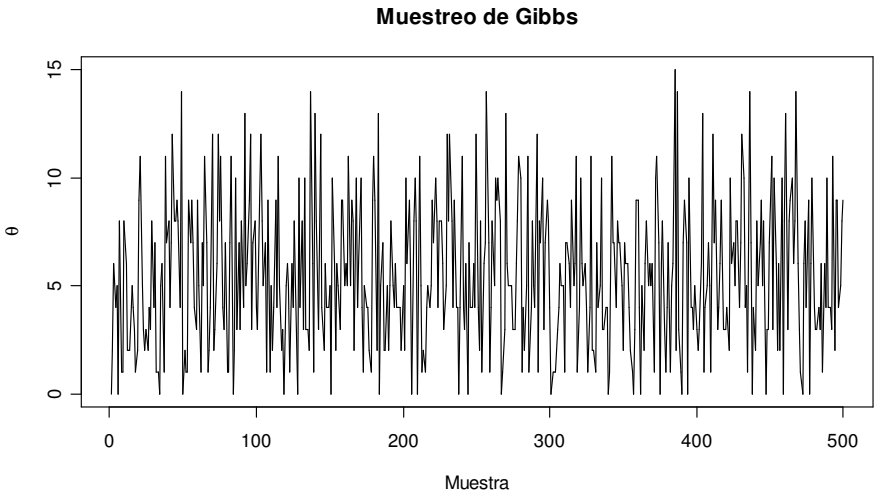

Figura 5.3: Valores de theta obtenidos del ejemplo 5.4 al aplicarle muestreo de Gibbs.

En la figura 5.3 se observa que los valores obtenidos para θ_1 son una muy buena mezcla, pues presentan un ruido blanco, es decir, son una sucesión de variables aleatorias independientes e idénticamente distribuidas.

El muestreo de Gibbs en este ejemplo no es necesario, puesto que $f(x)$ puede ser obtenida analíticamente de (5.1)

$$\begin{aligned} f(x) &= \int f(x,y)\,dy \\ &= \int \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}\,dy \\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+n+\beta)} \times \\ &\quad \int \frac{\Gamma(x+\alpha+n-x+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}\,dy \\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+n+\beta)}. \end{aligned}$$

La distribución beta-binomial, sin embargo, es útil para ilustrar el proceso. En la figura 5.4 se muestra el histograma con los datos generados del muestreo de Gibbs.

El muestreo de Gibbs puede ser indispensable en situaciones donde la dis-

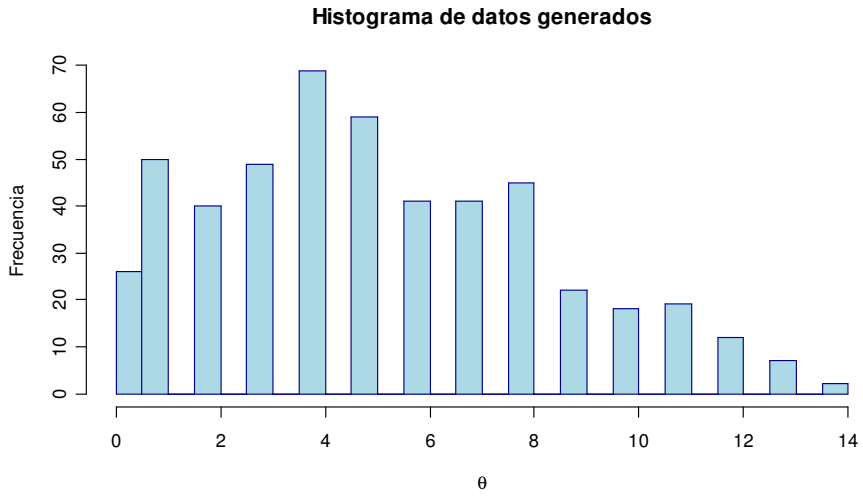

Figura 5.4: Histrograma del ejemplo 5.4 generado por el muestreo de Gibbs.

tribución conjunta o sus marginales no puedan ser calculadas, lo cual se muestra a continuación.

Ejemplo 5.5. ([7]) Supóngase que X y Y tienen distribuciones condicionales que son distribuciones exponenciales restringidas al intervalo $(0, B)$, esto es,

$$f(x \mid y) \propto y \exp(-yx), \quad 0 < x < B < \infty \tag{5.2}$$

$$f(y \mid x) \propto x \exp(-xy), \quad 0 < y < B < \infty \tag{5.3}$$

donde B es una constante positiva conocida ($B = 5$). La restricción del intervalo $(0, B)$ asegura que la marginal $f(x)$ existe. Aunque la forma de esta marginal no es fácil de calcular, aplicando el muestreo de Gibbs a las condicionales (5.2 y 5.3) cualquier característica de $f(x)$ puede ser obtenida. En la figura 5.5 se muestra los valores graficados de $\theta = X$ (ver apéndice A5.4).

De la misma forma que el ejemplo anterior en la figura 5.5 se observa que los valores obtenidos son una buena mezcla, pues presentan un ruido blanco, es decir, una sucesión de variables aleatorias independientes idénticamente distribuidas.

En la figura 5.6 se muestra un histograma de una muestra de tamaño 500 de $f(x)$, obtenida al usar los valores finales de las sucesiones de Gibbs de longitud 15 (ver apéndice A5.4).

En la figura 5.6 se observa que el histograma obtenido se aproxina a una función exponencial.

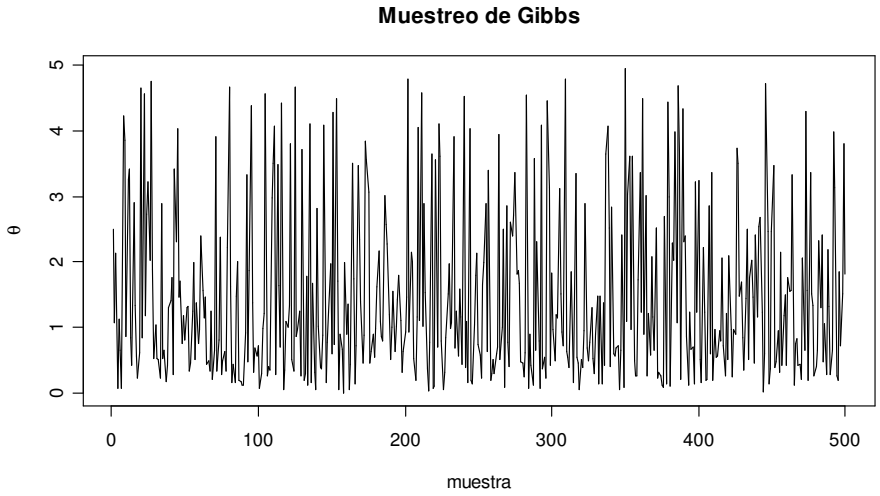

Figura 5.5: Valores de theta obtenidos del ejemplo 5.5 al aplicarle muestreo de Gibbs.

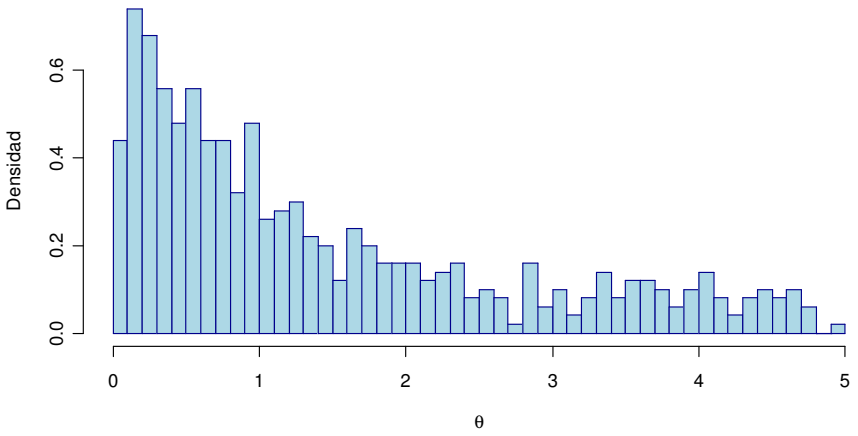

Figura 5.6: Histograma de datos usando muestreo de Gibbs con B=5.

5.2.3. Forma híbrida de los algoritmos

Una buena característica de los algoritmos de Monte Carlo vía cadenas de Markov es que pueden ser combinados en un mismo problema, con el fin de tomar las ventajas que cada uno ofrece. Un método interesante es el siguiente.

Supóngase que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, y que todas las distribuciones condicionales completas son disponibles en forma cerrada excepto

$$\pi\left(\theta_{i,t+1} \mid \theta_{1,t+1}, \theta_{2,t+1}, \ldots, \theta_{i-1,t+1}, \theta_{i+1,t}, \ldots, \theta_{d,t}, \mathbf{x}\right),$$

quizás porque no hay conjugada en la a priori para θ_i. Puesto que el algoritmo de Metropolis-Hastings no requiere densidades normalizadas de las cuales muestrear, se puede pensar en correr un muestreo de Gibbs con un subalgoritmo Metropolis para obtener la muestra necesaria $\theta_i^{(t)}$ en la iteración t. Esto es, se puede escoger una densidad candidato condicional simétrica

$$q\left(\theta_i^{(t-1)}, \theta^* \mid \theta_{j<i}^{(t)}, \theta_{j>i}^{(t-1)}, \mathbf{x}\right),$$

y correr un subalgoritmo Metropolis para T iteraciones, aceptando o rechazando los candidatos como apropiados. Entonces hacemos $\theta_i^{(t)}$ igual al resultado final de este subalgoritmo, y procedemos con el paso externo de Gibbs. Nótese entonces que esta aproximación puede ser aplicada a cualquier parámetro del cual se desconosca una forma cerrada condicional completa. Esta aproximación es frecuentemente referida como Metropolis dentro de Gibbs.

Esta mezcla de los dos métodos y los dos métodos en si, son los métodos Monte Carlo vía cadenas de Markov más comunes actualmente en uso, pero hay una variedad de extensiones y otros algoritmos MCMC que son útiles en el análisis de ciertos modelos específicos.

5.3. Inferencia

El método básico de inferencia para los algoritmos iterativos es usar la colección de toda la simulación generada de $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ para resumir la densidad a posteriori y calcular cuantiles, momentos, y otras cantidades de interés. Para disminuir el efecto de la distribución inicial, generalmente se descarta la primera mitad de cada sucesión y se enfoca la atención en la segunda mitad, entonces la inferencia se basa en el supuesto de que las distribuciones de valores simulados $\boldsymbol{\theta}_t$, para t suficientemente grande, son cercanos a la distribución objetivo.

Para verificar si la convergencia ha sido alcanzada se pueden usar k simulaciones generadas, con el propósito de tener aproximadamente generaciones independientes de la distribución objetivo, aunque aquí se tiene el problema de

almacenar los cálculos. La recomendación para inferir de la simulación iterativa se basa en comparar diferentes sucesiones simuladas, con puntos iniciales generados de una distribución dispersa. Se pueden estimar todos los parámetros de interés en el modelo y cualquier cuantil de interés o el valor de una observación futura.

Si se quiere obtener la estimación de la varianza de las medias a posteriori obtenidas de las salidas MCMC se procede como sigue: supóngase que para un parámetro dado θ (unidimensional) se tiene una única cadena de muestras MCMC, es decir $\left\{\theta^{(t)}\right\}_{t=1}^{N}$. La estimación más simple de $E(\theta \mid \mathbf{x})$ está dada por

$$E(\theta \mid \mathbf{x}) = \widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N}\theta^{(t)}.$$

Se puede estimar $Var\left(\widehat{\theta}_N\right)$ usando la varianza muestral $s_\theta^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\theta^{(t)} - \widehat{\theta}_N\right)^2$, es decir,

$$\widehat{Var}\left(\widehat{\theta}_N\right) = \frac{s_\theta^2}{N} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\theta^{(t)} - \widehat{\theta}_N\right)^2.$$

Otra forma de estimar un sólo parámetro con J cadenas es: si para cada parámetro θ, se etiqueta la generación de J sucesiones paralelas de longitud n como θ_{ij} $(i = 1, ..., n; j = 1, ..., J)$, y se calculan las varianzas (B) entre y (W) dentro de cada serie,

$$B = \frac{n}{J-1}\sum_{j=1}^{J}\left(\overline{\theta}_{\cdot j} - \overline{\theta}_{\cdot\cdot}\right)^2, \text{ donde } \overline{\theta}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij},\ \overline{\theta}_{\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J}\overline{\theta}_{\cdot j},$$

$$W = \frac{1}{J}\sum_{j=1}^{J}s_j^2, \text{ donde } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\theta_{ij} - \overline{\theta}_{\cdot j}\right)^2.$$

Se puede estimar $Var(\theta \mid x)$, la varianza a posteriori del estimador como

$$\widehat{Var}^{+}(\theta \mid x) = \frac{n-1}{n}W + \frac{1}{n}B.$$

Los algoritmos estudiados en este capítulo son algoritmos iterativos, donde se requiere muestrear hasta que la convergencia se haya obtenido.

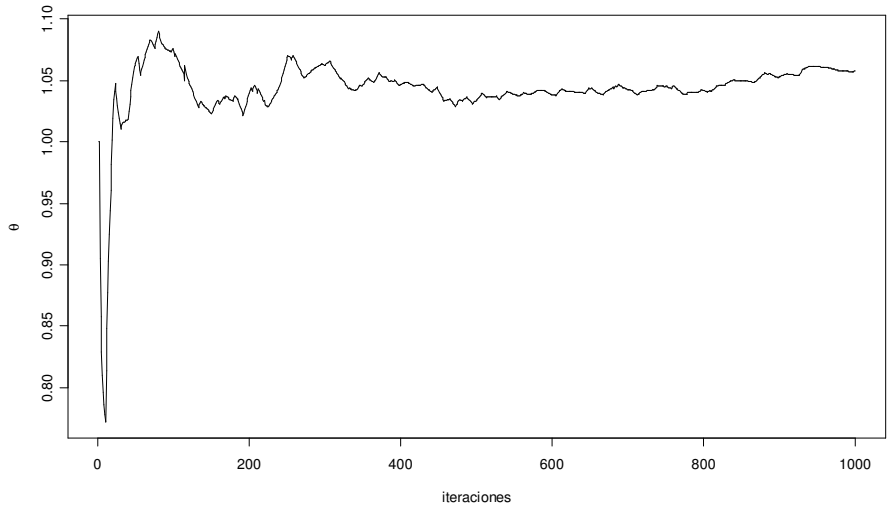

Figura 5.7: Promedio ergódico del ejemplo 5.6 (Metropolis-Hastings).

5.4. Determinación de la convergencia

Se concluye la presentación de los algoritmos Monte Carlo vía cadenas de Markov con una discusión de su convergencia.

Cuando se dice que un algoritmo alcanza la convergencia al tiempo T, significa que sus salidas provienen de la verdadera distribución estacionaria de la cadena de Markov para toda $t > T$.

Si se quiere generar una muestra de tamaño N de la distribución $\pi(\boldsymbol{\theta} \mid \mathbf{x})$, se corre alguno de los dos algoritmos anteriores para cada uno de los N valores iniciales $\boldsymbol{\theta}_0^1, \ldots, \boldsymbol{\theta}_0^N$, después de un cierto número de iteraciones T suficientemente grande los valores $\boldsymbol{\theta}_T^1, \ldots, \boldsymbol{\theta}_T^N$ pueden considerarse una muestra de la distribución final de $\boldsymbol{\theta}$.

Un método basado en la proposición 5.1 (ii) para determinar en que momento la cadena alcanza la convergencia es graficar los promedios ergódicos de algunas funciones de $\boldsymbol{\theta}$ contra el número de iteraciones y elegir el valor T a partir del cual las gráficas se estabilizan.

Ejemplo 5.6. Continuando con el ejemplo 5.3, en la figura 5.7 se observa el promedio ergódico para los primeros 500 valores de θ, del ejemplo del algoritmo de Metropolis-Hastings, en la figura se observa que la convergencia aún no se alcanza en esa iteración (ver apéndice A5.5).

Por lo que se vuelve a generar otra cadena con 10000 iteraciones. En la figura

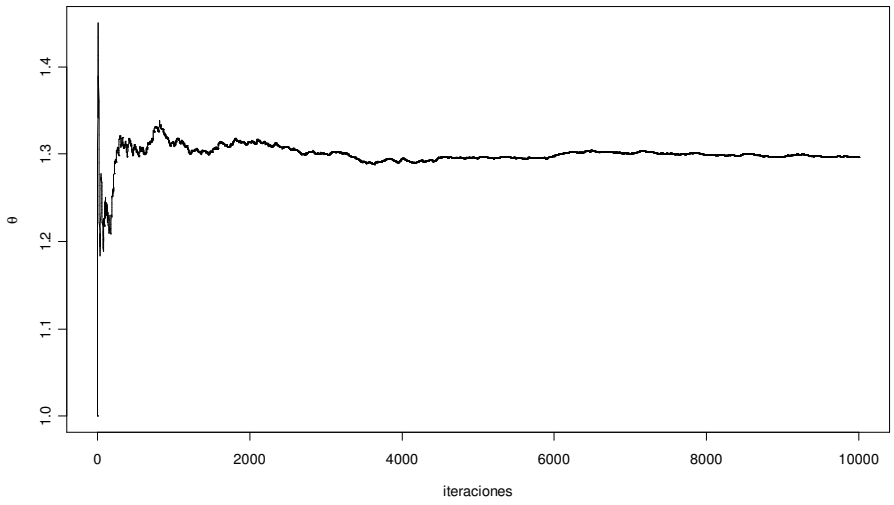

Figura 5.8: Promedio ergódico con 10000 iteraciones (Metropolis-Hastings)

5.8 se observa el promedio ergódico y se observa que la convergencia se alcanza en la iteración 6500.

Esto es debido a que la función candidato generadora que se utiliza en ese ejemplo no se parece a la función objetivo para ese ejemplo.

Otro ejemplo donde la convergencia se alcanza con mayor rapidez es el siguiente.

Ejemplo 5.7. Continuando con el ejemplo 5.5, se analiza la convergencia del método graficando el promedio ergódico de los valores generados para θ, donde se observa en la figura 5.9 que después de 250 iteraciones la convergencia se ha alcanzado (ver apéndice A5.6).

Es común eliminar los primeros valores de la cadena, con el fin de permitir que la cadena salga de una primera fase de inestabilidad, a este período se le llama período de calentamiento.

Otra manera de determinar la convergencia es graficando algunas sucesiones cortas de diferentes cadenas las cuales inician con puntos dispersos y observar cuando los puntos convergen a un punto estacionario.

Hay un debate general acerca de una única corrida y múltiples corridas. Pues una única sucesión tiene dificultad para dejar la vecindad de un modelo atractivo que exhibe aceptable conocimiento sin pensar que éste falla al explorar todo el soporte de la distribución objetivo. Múltiples sucesiones podrían tener mejor fuerza

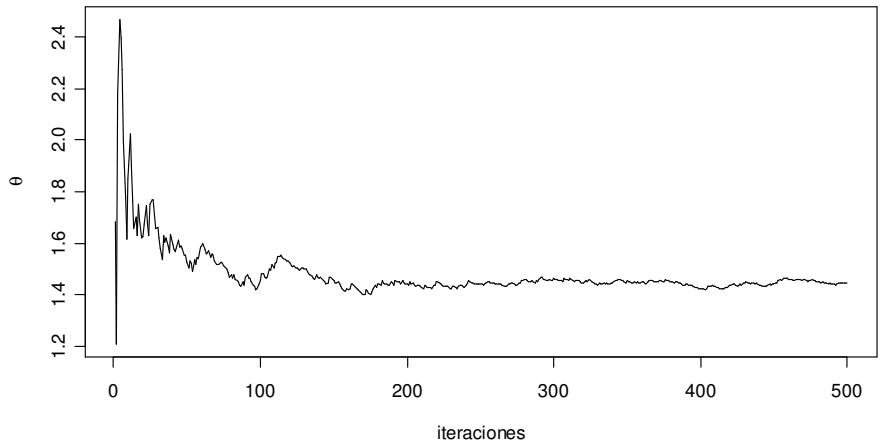

Figura 5.9: Promedio ergódico del ejemplo 5.7 (muestreo de Gibbs).

de exploración, pero dependen altamente de la elección de los puntos iniciales, además se pueden tener problemas al almacenar los cálculos. Como una guía se sugiere usar ambos tipos de gráficas al explorar la convergencia.

Para muchos problemas prácticos, la dimensión del espacio paramétrico es alto y no puede ser factible examinar las gráficas de todos los parámetros, en tal caso se pueden construir gráficas de algunos parámetros seleccionados.

Para el muestreo de Gibbs la velocidad de convergencia depende de la correlación entre los componentes del vector $\boldsymbol{\theta}$ bajo la distribución final $\pi(\boldsymbol{\theta} \mid \mathbf{x})$, pues entre más alta sea la correlación más lenta será la convergencia.

En el apéndice A se muestra un ejemplo real, en el que se implementan algunos algoritmos de los que aquí se han analizado.

Capítulo 6

Conclusiones

La estadística contribuye al desarrollo de técnicas apropiadas para hacer inferencia bajo condiciones de incertidumbre. Por eso, la estadística bayesiana constituye una eficiente alternativa al enfoque clásico, ya que el enfoque bayesiano proporciona una forma natural de plantear los problemas estadísticos como un problema de decisión. Además, tiene una base axiomática en los llamados axiomas de coherencia, dando resultados consistentes entre sí. Se debe mencionar que en esta tesis no se planteó ningún problema estadístico como un problema de decisión, pues sólo nos enfocamos a analizar algunos métodos de integración y simulación Monte Carlo, vistos desde un enfoque bayesiano.

En la estadística clásica sólo se utiliza la información muestral, mientras que en la estadística bayesiana, además de la información muestral representada en la función de verosimilitud, también se utiliza la información que el investigador tiene acerca del problema, representada ésta en la función de distribución a priori. Combinando la información muestral y la información a priori en el teorema de Bayes se obtiene un conocimiento final de las probabilidades, representado éste en una distribución a posteriori.

Al utilizar el teorema de Bayes surgen integrales que muchas veces no pueden ser calculadas analíticamente, pero gracias al avance computacional se pueden utilizar diversos algoritmos estadísticos para aproximar las integrales que surgen en el análisis bayesiano.

En esta tesis se han analizado diversos métodos de integración Monte Carlo, tales como muestreo directo, muestreo por composición, muestreo por importancia y muestreo aceptación-rechazo.

Por otra parte se analizaron dos métodos de simulación Monte Carlo vía cadenas de Markov, los cuales son el algoritmo de Metropolis-Hastings y el muestreo de Gibbs.

Es difícil decidir cual de los métodos aquí analizados es mejor, pues todos son

muy buenos. Estos se utilizan dependiendo de la información que uno requiera en algún problema en particular.

En cuanto a los métodos de integración Monte Carlo el más utilizado es el muestreo por importancia, pues en este método no es necesario conocer la distribución a posteriori sino que se toma una distribución pivote la cual aproxima a la distribución a posteriori y es de la distribución pivote de la que se toma una muestra para calcular los valores deseados.

A diferencia del muestreo directo en el que se necesita conocer la distribución a posteriori. El método de composición es útil cuando no se puede obtener una muestra de una distribución marginal, entonces se obtiene la muestra de distribuciones condicionales y con estas se inplementa el método de composición.

En el método de aceptación-rechazo no es necesario conocer la distribución a posteriori, solo se necesita conocer la forma funcional de esta distribución a posteriori, además de una constante, la cual se obtiene maximizando un cociente que involucra a la distribución a priori, la verosimilitud y la función de probabilidad envolvente. Luego se trata de envolver la forma funcional de la distribución a posteriori para aplicar el muestreo aceptación rechazo. Estos métodos regularmente se utilizan para calcular valores esperados, intervalos creíbles y estimaciones puntuales.

Análogamente, los métodos de simulación aquí estudiados son buenos, teniendo sus ventajas cada uno. El algoritmo de Metropolis-Hastings se utiliza cuando se conoce la forma funcional de la función objetivo, en cambio, el muestreo de Gibbs se utiliza cuando no se conoce la función objetivo, pero la distribuciones marginales a posteriori se conocen y son fáciles de muestrear. Aunque estos métodos pueden llegar a tener un costo computacional muy alto, pues el almacenamiento de los datos aumenta cuando aumenta la dimensión del espacio parámetrico.

De estos métodos el más utilizado es el muestreo de Gibbs, pues en este método se muestrea de distribuciones condicionales, las cuales son fáciles de generar. Pero si no se puede simular alguna distribución condicional, se utiliza el algoritmo de Metropolis-Hastings en el paso del muestreo de Gibbs. El muestreo de Gibbs tiene un amplio uso en problemas prácticos y puede ser usado tanto en la estadística clásica como en la estadística bayesiana.

En la estadística bayesiana, el muestreo de Gibbs se usa principalmente para la aproximación de distribuciones a posteriori, mientras para la estadística clásica el principal uso es para calcular la función de verosimilitud y algunas características de los estimadores.

La ventaja del muestreo de Gibbs se observa cuando se incrementa enormemente la dimensión del problema (es decir, aumenta la dimensión del parámetro $\boldsymbol{\theta}$).

El principal uso de los métodos de simulación es aproximar distribuciones a posteriori, y de ahí hacer inferencias. Sin duda, la mejor opción en un problema particular, es combinar varios de éstos métodos, con el fin de obtener una mejor precisión.

En el ejemplo de datos reales presentado en el apéndice A se decidió utilizar una distribución a priori no informativa de Jeffreys, lo cual es válido, debido a que las inferencias se realizan con la distribución a posteriori, la cual obtenemos al aplicar el teorema de Bayes. Si se hubiera elegido trabajar con una distribución a priori conjugada se esperarían resultados similares a los obtenidos aquí.

Cuando se realizó el muestreo directo se obtuvo una buena estimación a la media a posteriori. Cuando se implementó el muestreo por importancia, se tomó como distribución pivote la aproximación normal a la distribución a posteriori, lo cual arrojó excelentes resultados. Y cuando se realizó el muestreo de Gibbs, también se obtuvo una buena aproximación a la distribución a posteriori, al igual, la estimación obtenida a la media a posteriori fue buena, pero aquí el costo computacional fue mayor, pero la ventaja de implementar el muestreo de Gibbs se observa cuando la dimensión de parámetro que se desea estimar aumenta.

La predicción que se obtuvo del número de accidentes aéreos para el año 1997 es muy buena, pues se utilizó la distibución predictiva a posteriori y la predicción del número de accidentes aéreos para ese año es 26, cuando en realidad, según los datos proporcionados por la Secretaría de comunicaciones y transportes y el Instituto Mexicano del transporte fue de 27.

Apéndice A

Ejemplo con datos reales

La estadística tiene un sin fin de aplicaciones, tanto en áreas científicas como sociales. En este capítulo aplicaremos algunos de los métodos estudiados en esta tesis a un problema real de accidentes aéreos.

Se presentan datos estadísticos en relación a la aviación comercial de servicio regular, desde el año 1969 y hasta el año 1997, obtenida de la publicación técnica No. 152 de la Secretaría de Comunicaciones y Transportes e Instituto Mexicano del Transporte.

El convenio sobre aviación civil internacional y la seguridad aérea se creó en 1944, cuando los delegados de 52 naciones se reunieron en Chicago para firmar el Convenio sobre aviación civil internacional. La Organización de Aviación Civil Internacional (OACI) ha estado vinculada desde entonces a la historia de la aviación. A ella se deben todas las normas técnicas y reglamentos, así como la elaboración del marco jurídico, que ha permitido el desarrollo ordenado de la aviación civil internacional.

En la tabla A.1 se presenta la información estadística de los accidentes en líneas aéreas comerciales de servicio regular (doméstico e internacional) de los países miembros de la OACI.

Año	Número de accidentes	Número de pasajeros muertos
1969	33	965
1970	29	700
1971	32	884
1972	41	1209
1973	36	862
1974	29	1299
1975	21	467
1976	20	734
1977	24	516
1978	25	754
1979	31	877
1980	22	814
1981	21	362
1982	26	764
1983	20	809
1984	16	223
1985	22	1066
1986	17	331
1987	24	890
1988	25	699
1989	27	817
1990	22	440
1991	25	510
1992	29	1097
1993	34	936
1994	28	941
1995	26	710
1996	23	1135
1997	27	930

Tabla A.1: Accidentes aéreos internacionales de los países miembros de la OACI.

Se tomarán los datos de la tabla de los años 1969 a 1996 para hacer todos nuestros analisis y se hará una predicción para el número de accidentes aéreos para el año 1997.

En la figura A.1 se muestra un histograma del número de accidentes aéreos fatales (ver apéndice A6.1).

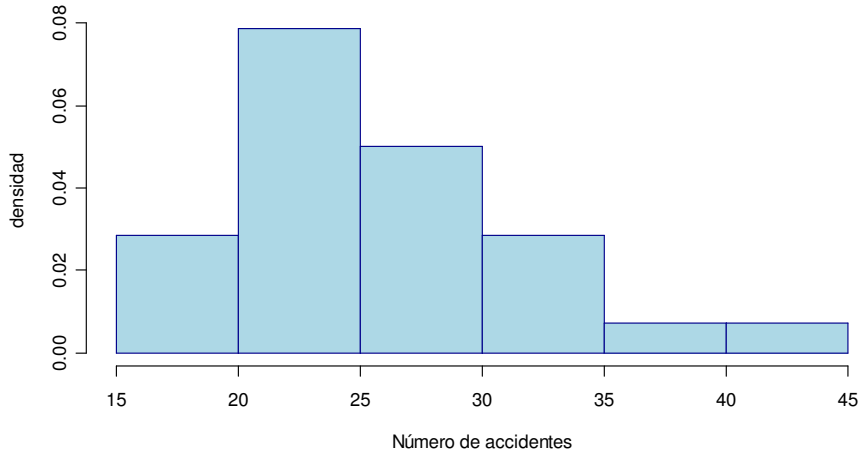

Figura A.1: Histograma del número de accidentes fatales.

Se supone que el número de accidentes fatales cada año es independiente al de otros años, con una distribución de Poisson.

$$X_i \sim Pn\left(x_i \mid \theta\right) = \exp\left(-\theta\right)\frac{\theta^{x_i}}{x_i!}, \quad \theta > 0.$$

se toma una distribución a priori no informativa de Jeffreys (2.2), con el fin de no introducir mucha subjetividad al problema, se le puede asignar ésta distribución, pues con la distribución a posteriori es con la que se hacen las inferencias. Entonces se tiene que

$$\begin{aligned}
f\left(x_i \mid \theta\right) &= \exp\left(-\theta\right)\frac{\theta^{x_i}}{x_i!}, \\
\log\left(f\left(x_i \mid \theta\right)\right) &= -\theta + x_i \log\left(\theta\right) - \log\left(x_i!\right), \\
\frac{\partial\left[\log\left(f\left(x_i \mid \theta\right)\right)\right]}{\partial\theta} &= -1 + \frac{x_i}{\theta}, \\
\frac{\partial^2\left[\log\left(f\left(x_i \mid \theta\right)\right)\right]}{\partial\theta^2} &= -\frac{x_i}{\theta^2}, \\
I\left(\theta\right) &\propto -E\left[-\frac{x_i}{\theta^2}\right] = \frac{E\left(x_i\right)}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}, \\
\pi\left(\theta\right) &\propto \theta^{-\frac{1}{2}}.
\end{aligned}$$

Entonces la distribución conjunta es

$$\begin{aligned} f(\mathbf{x}|\theta)\pi(\theta) &= \Pi_{i=1}^{n}\left[\frac{\theta^{x_i}\exp(-\theta)}{x_i!}\right]\theta^{-\frac{1}{2}} = \frac{\theta^{n\overline{\mathbf{x}}}\exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]}\theta^{-\frac{1}{2}} \\ &= \frac{\theta^{(n\overline{\mathbf{x}}+1)-1}\exp(-n\theta)\,\theta^{-\frac{1}{2}}}{\Pi_{i=1}^{n}[x_i!]} = \frac{\theta^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)-1}\exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]}, \end{aligned}$$

y la distribución marginal es

$$\begin{aligned} m(\mathbf{x}) &= \int f(\mathbf{x}\mid\theta)\,\pi(\theta)\,d\theta \\ &= \frac{\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}{\Pi_{i=1}^{n}[x_i!]\,n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}\int\frac{n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}{\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\theta^{(n\overline{\mathbf{x}}+\frac{1}{2})-1}\exp(-n\theta)\,d\theta \\ &= \frac{\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}{\Pi_{i=1}^{n}[x_i!]\,n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}. \end{aligned}$$

La distribución a posteriori es

$$\begin{aligned} \pi(\theta\mid\mathbf{x}) &= \frac{f(\mathbf{x}\mid\theta)\,\pi(\theta)}{m(x)} \\ &= \frac{n^{(n\overline{\mathbf{x}}+\frac{1}{2})}}{\Gamma\left(n\overline{x}+\frac{1}{2}\right)}\theta^{(n\overline{\mathbf{x}}+\frac{1}{2})-1}\exp(-n\theta) \end{aligned}$$

$$\theta\mid\mathbf{x}\sim Ga\left(\theta\mid\left(n\overline{\mathbf{x}}+\frac{1}{2}\right),n\right).$$

La figura A.2 nos muestra la distribución a posteriori $\pi(\theta\mid\mathbf{x})$ (ver apéndice A6.2).

Podemos obtener los momentos de la distribución a posteriori, y son

$$E(\theta\mid\mathbf{x}) = \frac{\left(n\overline{x}+\frac{1}{2}\right)}{n},\ \text{y}\ V(\theta\mid\mathbf{x}) = \frac{\left(n\overline{x}+\frac{1}{2}\right)}{n^2}.$$

El resultado numérico se muestra en la tabla A.2.

Esperanza	26.01786
Varianza	0.9292092
Intervalo creíble	(24.16259, 27.94077)

Tabla A.2. Momentos e intevalo creíble al 95 %
para la distribución a posteriori (apéndice A6.3).

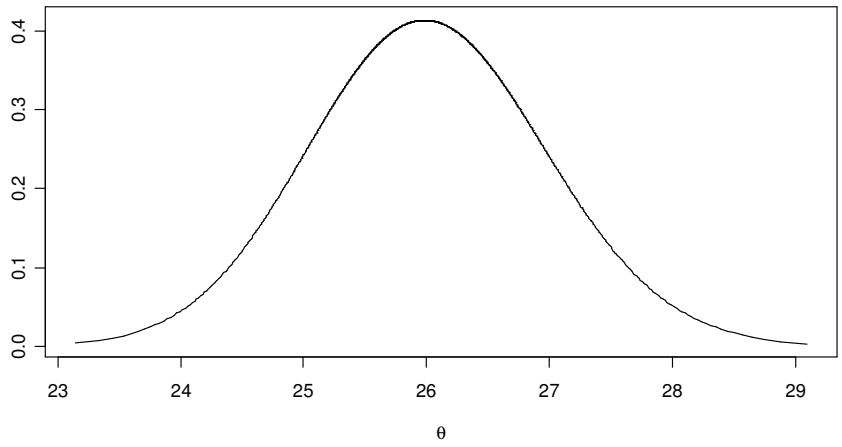

Figura A.2: Distribución a posteriori.

La distribución predictiva a posteriori es

$$\begin{aligned} m\left(x^{*}|\mathbf{x}\right) &= \int f\left(x^{*}|\theta\right)\pi\left(\theta|x\right)d\theta \\ &= \int \exp\left(-\theta\right)\frac{\theta^{x^{*}}}{x^{*}!}\left[\frac{(n)^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\theta^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)-1}\exp\left(-n\theta\right)\right]d\theta \\ &= \frac{(n)^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{x^{*}!\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\frac{\Gamma\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)}{(1+n)^{\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)}}\times \\ &\quad \int\frac{(1+n)^{\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{\Gamma\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)}\theta^{\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)-1}\exp\left(-\left(1+n\right)\right)d\theta \\ &= \frac{(n)^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}}{x^{*}!\Gamma\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)}\frac{\Gamma\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)}{(1+n)^{\left(x^{*}+n\overline{\mathbf{x}}+\frac{1}{2}\right)}}, \end{aligned}$$

$$x^{*}|\mathbf{x}\sim Pg\left(x^{*}\mid\left(n\overline{\mathbf{x}}+\frac{1}{2}\right),n,1\right).$$

La gráfica de la distribución predictiva a posteriori se observa en la figura A.3 (ver apéndice 6.4).

En la tabla A.3 se muestra el valor esperado, la varianza y un intervalo creíble al 95 % para la distribución predictiva a posteriori.

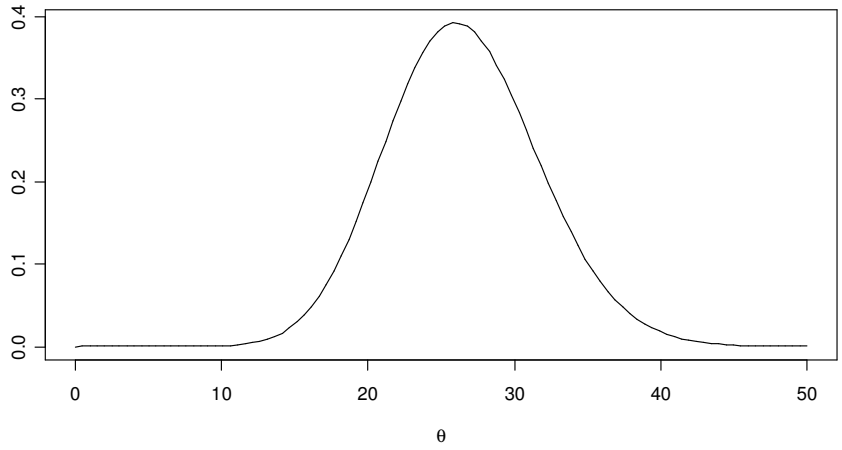

Figura A.3: Distribución predictiva a posteriori.

Esperanza	26.01786
Varianza	26.94707
Intervalo creíble	$(16, 37)$

Tabla A.3. Momentos de la distribución predictiva a posteriori.

El valor esperado de la predictiva a posteriori es una estimación para el número de accidentes aéreos para el año 1997, el cual es 26. Cuando en realidad, según los datos proporcionados en el reporte técnico de la Secretaría de Comunicaciones y Transportes y el Instituto Mexicano del Transporte, el número de accidentes aéreos para el año de 1997 es de 27. Por lo que se puede decir que esta es una muy buena predicción.

Si se aplica el muestreo directo para estimar $E(\theta \mid \mathbf{x})$, se genera

$$\theta^1, \theta^2, \ldots, \theta^m \sim \pi(\theta \mid \mathbf{x}) = Ga\left(\theta \mid \left(n\overline{\mathbf{x}} + \frac{1}{2}\right), n\right)$$

Como

$$E(\theta \mid \mathbf{x}) = \int \theta\pi(\theta \mid \mathbf{x})\, d\theta \approx \frac{1}{m}\sum_{i=1}^{m} \theta^i,$$

entonces, con $m = 10000$ se obtiene (ver apéndice A6.5):

	A posteriori	Muestreo directo
Media	26.01786	26.02888
Varianza	0.9292092	0.924308
intervalo creíble	(24.16259, 27.94077)	(24.17059, 27.93662)

Tabla A.4. Momentos e intervalo creíble para la estimación al 95 %.

En la tabla A.4 se observa que la aproximación es muy buena al igual que el intervalo creíble puesto que la diferencia de los valores obtenidos por el muestreo directo y los valores reales de la distribución a posteriori no es significativa.

Si ahora se aplica el muestreo importante para calcular la $E(\theta \mid \mathbf{x})$. Se tiene que conseguir una distribución pivote de la que sea fácil simular

$$\pi^*(\theta) \approx cf(\mathbf{x} \mid \theta)\pi(\theta)$$

y de ahi utilizar $\theta_j \sim \pi^*(\theta)$ para poder aproximar

$$\begin{aligned} E(\theta \mid \mathbf{x}) &= \int \theta\pi(\theta \mid \mathbf{x})\, d\theta = \frac{\int \theta f(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta}{\int f(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta} \\ &= \frac{\int \theta w(\theta)\pi^*(\theta)\, d\theta}{\int w(\theta)\pi^*(\theta)\, d\theta} \approx \frac{\frac{1}{M}\sum_{j=1}^{M} \theta_j w(\theta_j)}{\frac{1}{M}\sum_{j=1}^{M} w(\theta_j)}, \end{aligned}$$

donde

$$w(\theta) = \frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{\pi^*(\theta)}.$$

Si se utiliza como distribución pivote la aproximación normal (3.1),

$$\begin{aligned}
\pi(\theta \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta)\pi(\theta) = \frac{\theta^{\left(n\overline{\mathbf{x}}+\frac{1}{2}\right)-1}\exp(-n\theta)}{\Pi_{i=1}^{n}[x_i!]} \\
L(\theta) &= \log[f(\mathbf{x} \mid \theta)\pi(\theta)] \\
L(\theta) &= \left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)\log\theta - n\theta - \log(\Pi_{i=1}^{n}[x_i!]) \\
\frac{\partial L(\theta)}{\partial\theta} &= \frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{\theta} - n = 0 \\
\widehat{\theta} &= \frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{n} \\
\frac{\partial^2 L(\theta)}{\partial\theta^2} &= -\frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{\theta^2},
\end{aligned}$$

evaluando en el estimador

$$\begin{aligned}
\frac{\partial^2 L(\theta)}{\partial\theta^2} \quad \Big| \quad {}_{\theta=\widehat{\theta}} &= -\frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{\frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)^2}{n^2}} = -\frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)n^2}{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)^2} \\
&= -\frac{n^2}{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)},
\end{aligned}$$

se tiene que

$$\begin{aligned}
I^{\pi}(x) &= \frac{n^2}{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}, \\
[I^{\pi}(x)]^{-1} &= \frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{n^2},
\end{aligned}$$

entonces la distribución pivote por aproximación normal es

$$N\left(\frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{n}, \frac{\left(\sum_{i=1}^{n} x_i - \frac{1}{2}\right)}{n^2}\right).$$

Así, la estimación de $E(\theta \mid \mathbf{x})$ por muestreo importante se presenta en la tabla A.5 (ver apéndice 6.6).

	A posteriori	Muestreo por importancia
Media	26.01786	26.01279
Varianza	0.9292092	0.9275294
Intervalo creíble	(24.16259, 27.94077)	(24.12515, 27.90043)

Tabla A.5. Momentos e intervalo creíble para la estimación al 95 %.

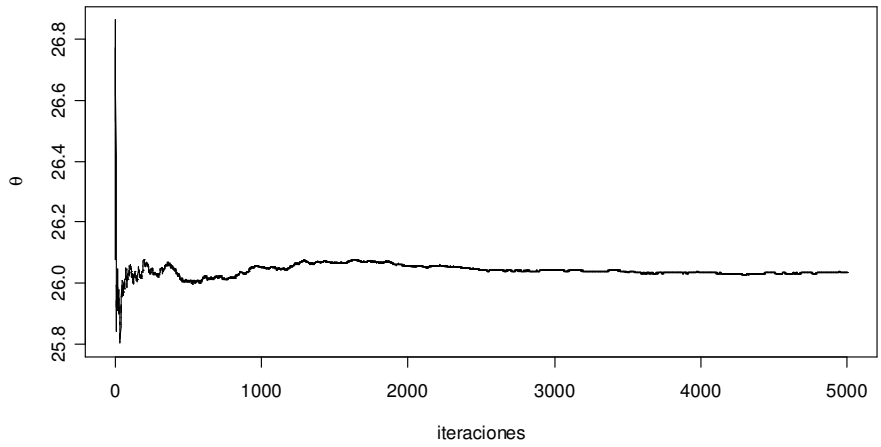

Figura A.4: Promedio ergódico para el parámetro theta por muestreo de Gibbs.

En la tabla A.5 se observa que la aproximación es muy buena pues la diferencia de los valores obtenidos en la estimación de la distribución a posteriori y los valores analíticos de la distribución a posteriori no es significativa.

Por otro lado, se aplica el muestreo de Gibbs para aproximar a la distribución a posteriori. Primero, se corre el algoritmo de Gibbs con una sola cadena de longitud 10000. Con el fin de observar en qué momento se alcanza la convergencia, se eliminan los primeros 5000 valores y se quedan 5000 valores. Con estos valores se grafica el promedio ergódico para el parámetro θ, que se puede ver en la figura A.4 (ver apéndice A6.7).

Mediante ésta gráfica se puede asegurar que la convergencia se alcanza a partir de la iteración 2500.

Ahora se toma una muestra de tamaño 50 para aproximar a la distribución a posteriori, con lo que se obteniene el histograma de la figura A.5 (ver apéndice A6.8).

En la figura A.5 se puede apreciar que el histograma obtenido con los valores simulados por el muestreo de Gibbs son una buena aproximación a la distribución a posteriori.

Con esta muestra se puede estimar el valor esperado, la varianza a posteriori y obtener un intervalo creíble para la estimación de $E(\theta \mid \mathbf{x})$ al 95 % (ver apéndice A6.9).

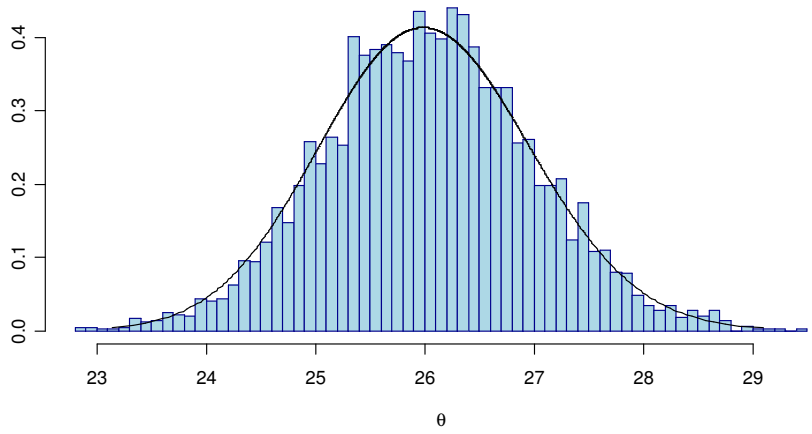

Figura A.5: Aproximación a la distribución a posteriori por muestreo de Gibbs.

	A posteriori	Muestreo de Gibbs
Media	26.01786	26.03613
Varianza	0.9292092	0.915703
Intervalo creíble	(24.16259, 27.94077)	(24.19954, 27.91502)

Tabla A.6. Momentos e intervalo creíble para la estimación al 95 %.

En la tabla A.6 se observa que la aproximación es muy buena dado que la diferencia de los valores obtenidos en la simulación y los valores de la distribución a posteriori no es significativa. La probabilidad de que el verdadero valor de la media se encuentre dentro del intervalo creíble es de 0.95, además también se observar que la longitud del intervalo creíble es relativamente pequeña.

Apéndice B

Notación

$Bi(x\|n,\theta)$	Distribución binomial con parámetros $0<\theta<1$, $n=1,\,2,\ldots$ y $x=0,\,1,\ldots,n$.
$Pn\,(x\|\lambda)$	Distribución Poisson con parámetro $\lambda>0$ y $x=0,\,1,\,2,\ldots$
$Be\,(x\|\alpha,\beta)$	Distribución beta con parámetros $\alpha>0$, $\beta>0$ y $0<x<1$.
$Ga\,(x\|\alpha,\beta)$	Distribución gamma con parámetros $\alpha>0$, $\beta>0$ y $x>0$.
$Exp\,(x\|\beta)$	Distribución exponencial con parámetro $\beta>0$ y $x\geq 0$.
$GI\,(x\|\alpha,\beta)$	Distribución gamma inversa con parámetros $\alpha>0$, $\beta>0$ y $x>0$.
$N\,(x\|\mu,\lambda)$	Distribución normal con parámetros $\mu\in\mathbb{R}$, $\lambda>0$.
$Pn\,(x\|\alpha,\beta)$	Distribución Poisson-gamma con parámetros $\alpha>0$, $\beta>0$ y $n=1,2,...$
i.i.d.	Independiente idénticamente distribuida.
$U\,(0,1)$	Distribución uniforme en el intervalo $(0,1)$.
$t_n\,(x\mid\lambda)$	Distribución t de Student con n grados de libertad con parámetro $\lambda>0$ y $x\in\mathbb{R}$.

Apéndice C

Códigos fuente en R

A2.1

```
# Momentos e intervalos creíbles de la distribución predictiva con distribución
a priori no informativa de Jeffreys
mon.aposteriori<-function(x,n)
{
    esp<-(sum(x)+(1/2))/( n)
    var<-((sum(x)+ (1/2))*(n+1))/(n)^2
    return(c(esp,var))
}
x<-c(24,25,31,31,22,21,26,20,16,22)
#1
mon.aposteriori(x,10)
# Intervalo predictivo
qnbinom(0.025,sum(x)+ (1/2),( 10)/(10+1))
qnbinom(0.975,sum(x)+ (1/2),( 10)/(10+1))
# Gráfica de la predictiva a posteriori
p.aposteriori<-function(xp,x,n)
{
    a<-((1/2)+sum(x))*log(n)+lgamma((1/2)+xp+sum(x))
    -lgamma(xp+(1/2))-lgamma((1/2)+sum(x))-( (1/2)+xp+sum(x))
    *(log(n+1))
    return(exp(a))
}
x.aposteriori<-seq(0,50, length=100)
x<-c(24,25,31,31,22,21,26,20,16,22)
plot(x.aposteriori, p.aposteriori(x.aposteriori, x,10), xlab=,ylab=, type="l")
```

A3.1

#Método de aproximación normal a la aposteriori
mu.sec_seq(0,1, length=100)
poste_dbeta(mu.sec,14,4)
plot(mu.sec,poste,type="l",ylim=c(0,5),xlab=expression(theta),
ylab="Densidad a posteriori")
var_((13/16)*(3/16))/16
aprox_dnorm(mu.sec,(13/16), sqrt(var))
lines(mu.sec,aprox,col=1,lty=2)
legend(0,5, c("Aproximación a posteriori","A posteriori exacta"), col = c(1,1),
lty = c(2, 1), pch = c(-1, -1))

A4.1

Método congruencial

Iniciar con una semilla arbitraria x_0

Iterar

$x_i = (69069x_{i-1}) \pmod{2^{32}}$

$u_i = 2^{-32}x_i$

Generar una distribución Normal estándar, $N(0,1)$

El método de Box-Muller (1958) prueba las salidas observadas normal independiente para dos variables aleatorias uniformes

1.- Generar U_1, U_2.

2.- Tomar

$x_1 = \sqrt{-2\log U_1}\cos(2\pi U_2)$,

$x_2 = \sqrt{-2\log U_1}\cos(2\pi U_2)$,

Generar una Distribución Exponencial, $\exp(\lambda)$

1.- Generar U.

2.- Tomar $x = -\log(U)/\lambda$

Este generador también se usa para la generación de una distribución geométrica $Geo(p)$. Si $x \sim Geo(p)$, $P(x = r) = P(r \leq E < r+1)$, con $E \sim \exp(-\log(1-p))$.

Generar una Distribución t-Student, $\Upsilon(\nu, 0, 1)$

Kinderman (1977) prueba una alternativa para la generación de una variable aleatoria normal y una variable aleatoria ji-cuadrada.

1.- Generar U_1, U_2.

2.- Si $U_1 < 0{,}5$, $x = 1/(4U_1 - 1)$ y $v = x^{-2}U_2$

en otro caso, $x = 4U_1 - 3$ y $v = U_2$.

3.- Si $v < 1 - (\mid x \mid /2)$ o $v < (1 + (x^2/v))^{-(v+1)/2}$, tomar x;

en otro caso, repetir.

Generar una Distribución Gamma, $\mathbf{G}(\alpha, 1)$

Los métodos simulados difieren acorde al valor de α (note que el factor escalar β se asume como 1). Cuando $\alpha > 1$, se sigue el algoritmo:

0.- Definimos $c_1 = \alpha - 1, c_2 = \left(\alpha - \left(\frac{1}{6\alpha}\right)\right)/c_1, c_3 = \frac{2}{c_1}, c_4 = 1 + c_3$ y $c_5 = \frac{1}{\sqrt{\alpha}}$

1.- Repetir

$\alpha > 2{,}5$

hasta $0 < U_1 < 1$.

2.- $W = \frac{c_2 U_2}{U_1}$

3.- Si $c_3 U_1 + W + W^{-1} \leq c_4$ o $c_3 \log U_1 - \log W + W \leq 1$, tomar $c_1 W$;

En otro caso, repetir.

Si α es muy grande $(\alpha > 50)$, es mejor usar una aproximación normal basada en el teorema del límite central.

Cuando $\alpha < 1$, un posible algoritmo es:

1.- Generar U y $y \sim G(\alpha + 1, 1)$.

2.- Tomar $yU^{1/\alpha}$

Las distribuciones Beta, Fisher y ji-cuadrada pueden ser simuladas usando este algoritmo dado que ellas pueden ser derivadas de la distribución Gamma.

Generar una distribución Binomial, $B(n, p)$

Cuando n es razonablemente pequeña $(n \leq 30)$, un algoritmo elemental para generar n variables aleatorias uniformes y contar éstos almenos p, Knuth (1981) prueba un algoritmo altermativo.

0.- Definimos $k = n$, $\theta = p$ y $x = 0$.

1.- Repetir

$i = [1 + k\theta]$

$u \sim Be(i, k + 1 - i)$

if $\theta > v$, $\theta = \frac{\theta}{v}$ y $k = i - 1$;

en otro caso,

$x = x + i$, $\theta = \frac{(\theta - v)}{(1 - v)}$ y $k = k - i$

hasta $k \leq K$.

2.- for $i = 1, 2, \ldots, k$,

generar U_i

if $U_i < p$, $x = x = x + 1$.

La constante K puede ser escogida como una función de orden n para incrementar la eficiencia del algoritmo.

Generación de una distribución Poisson, $\mathbf{P}(\lambda)$

Si λ es razonablemente pequeña $(\lambda < 30)$, un algoritmo simple para generar variables uniformes, en relación con el proceso de Poisson:

0.- $p = 1$, $N = 0$, $c = \exp(-\lambda)$

1.- Repetir

Generar U_i

$p = pU_i$, $N = N + 1$

hasta $p < c$.

3.-Tomar $x = N - 1$.

Para λ grande, Atkinson (1979) propone una alternativa mas eficiente.

0.- Definir $c = 0{,}767 - (3{,}36/\lambda)$, $\beta = \pi (3\lambda)^{-\frac{1}{2}}$, $\alpha = \beta\lambda$, $k = \log c - \lambda - \log \beta$.

1.- Repetir

Generar U_i

$x = [\alpha - \log((1 - U_1)/U_1)]/\beta$

hasta $x > -\frac{1}{2}$.

2.- Generar U_2.

3.- $N = [x + 0{,}5]$.

4.- Si $\alpha - \beta x + \log\left\{U_2/[1 + \exp(\alpha - \beta x)]^2\right\} \leq k + N \log \lambda - \log N!$

Tomar N;

en otro caso, repetir.

A4.2

Si no se conoce el valor de la media μ, pero se sabe que la desviación estándar es σ. Se determinará lo grande que debe ser el tamaño muestral para que la probabilidad de que $\mid \overline{X}_n - \mu \mid$ sea menor que $t > 0$ (t la discrepancia máxima) sea al menos $1 - \alpha$ (α la cota superior de la probabilidad de la discrepancia t).

De la desigualdad de Chebyshev aplicado a $\overline{X}_n$, puesto que $E\left(\overline{X}_n\right) = \mu$ y $Var\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$, para cualquier tamaño muestral n

$$P\left(\mid \overline{X}_n - \mu \mid \geq t\right) \leq \frac{\sigma^2}{nt^2}$$

puesto que n se debe elegir para que $P\left(\mid \overline{X}_n - \mu \mid < t\right) \geq 1 - \alpha$ resulta que se debe elegir n de forma que

$$\frac{\sigma^2}{nt^2} \leq \alpha.$$

Por tanto se necesita que

$$\frac{\sigma^2}{\alpha t^2} \leq n.$$

Si ahora tenemos proporciones. Sea p la probabilidad del evento A, y $f_A = \frac{n_A}{n}$ la frecuencia observada de A. Entonces, para cualquier número positivo ϵ se tiene

$$P(\mid f_A - f \mid \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2},$$

donde ϵ es la discrepancia máxima que vamos a permitir y δ la cota superior de la probabilidad de la discrepancia ϵ. Si no se conoce p (lo más común), el número de simulaciones n viene dado por

$$\frac{1}{4\delta\epsilon^2} \leq n.$$

A4.3

#código del ejemplo 4.1

```
N_100000
x2_runif(N,2,5)
y2_runif(N,0,0.1839)
f2_0.5*exp(-0.5*x2)
c2_0
for (i in 1:N)
{
     if (y2[i]<=f2[i])
     c2_c2+1
}
c2
A1_0.1839*3*(c2/N)
```

A4.4

#Muestreo directo

```
M=1000
y_rnorm(M)
mu.sec_qnorm(seq(0.001,0.999,len=1000),mean(y), sqrt(1/length(y)))
plot(mu.sec,dnorm(mu.sec,mean(y),sqrt(1/length(y))),type="l",
xlab=expression(theta), ylab=)
mu_rnorm(M,mean(y),sqrt(1/length(y)))
sum_0
for (i in 1:M)
{
     sum_sum+mu[i]
}
prom_sum/M
#media simulación
prom
#media analítica
mean(y)
#muestrear la varianza
summ_0
for (i in 1:M)
{
     summ_summ+(mu[i])^2
}
promm_summ/M
#varianza simulación
```

```
vvar_promm-(prom) ^2
vvar
#varianza analítica
v_1/M
v
#error estandar estimado
see_sqrt((1/(M-1))*vvar)
see
```

A4.5

```
# Muestreo por importancia
N_1000
theta_rnorm(N)
xbarra_mean(theta)
varianza_1/N
thetaimpor_rt(N,N)
t1_0
t2_0
w1_0
for (i in 1:N)
{
    w1[i]_ dnorm(xbarra, thetaimpor[i], sqrt(varianza))/dt(thetaimpor[i],N)
    t2_w1[i]+t2
    t1_thetaimpor[i]*w1[i]+t1
}
pomedio_t1/t2
#simulación
pomedio
#analítica
xbarra
```

A4.6

```
#método de aceptación-rechazo
var_((13/16)*(3/16))/16
k_0.0004433101
i_1
thetar_0
while (i<100)
{
    theta_rnorm(1,(13/16),sqrt(var))
    uu_runif(1,0,1)
    if (k*uu*dnorm(theta,(13/16),sqrt(var)) < (theta^13)*((1-theta)^3) )
```

```
    {
        thetar[i]_theta
        i=i+1
    }
}
mean(thetar)
var(thetar)
hist(thetar,xlab=expression(theta),ylab="",ylim=c(0,6),col="light blue",
border="dark blue",probability=T,main="")
mu.sec_seq(0,1, length=100)
poste_dbeta(mu.sec,14,4)
lines(mu.sec,poste,type="l")
```

A5.1

```
#Metrópolis
theta_1
alfa_4
beta_5
x_rexp(10)
for (i in 2:500)
{
    can_runif(1, 0, 100)
    vaunif_runif(1, 0, 1)
    f_((can^(10+alfa-1))*exp(-can*(sum(x)+beta)))
    /((theta[i-1]^(10+alfa-1))*
    exp(-theta[i-1]*(sum(x)+beta)))
    alpha_min(f, 1)
    if (vaunif<=alpha)
    {
        theta[i]_can
    }
    else
    {
        theta[i]_theta[i-1]
    }
}
#gráfica
plot(theta, main=, xlab="n", ylab=expression(theta), type="l")
```

A5.2

```
#Metropolis-Hastings
theta_1
```

```
alfa_4
beta_5
x_rexp(10)
for (i in 2:1000)
{
    can_rchisq(1, 2)
    vaunif_runif(1, 0, 1)
    f_((can^(10+alfa-1))*exp(-can*(sum(x)+beta))
        *(theta[i-1]^((can/2)-1))*exp(-theta[i-1]/2))
        /((theta[i-1]^(10+alfa-1))*exp(-theta[i-1]
        *(sum(x)+beta))*(can^((theta[i-1]/2)-1))
        *exp(-can/2))
    alpha_min(f, 1)
    if (vaunif<=alpha)
    {
        theta[i]_can
    }
    else
    {
        theta[i]_theta[i-1]
    }
}
plot(theta, xlab="Iteraciones", ylab=expression(theta), type="l")
```

A5.3

```
#Muestreo de Gibbs
re_0
for (j in 1:500)
{
    n_16
    alfa_2
    beta_4
    y_runif(1,0,1)
    x_0
    for (i in 1:10)
    {
        x[i+1]_rbinom(1, 16, y[i])
        y[i+1]_rbeta(1, x[i+1]+2,16-x[i+1]+4)
    }
    re[j]_x[10]
}
```

```
plot(re, main="Muestreo de Gibbs", xlab="Muestra", ylab=expression(theta),
type="l")
hist(re,50, col="light blue", border="dark blue", ylab="Frecuencia",
xlab=expression(theta), main="Histograma de datos generados")
```

A5.4

```
#Muestreo de Gibbs con múltiples cadenas
resul_0
for (j in 1:500)
{
    theta1_runif(1,0,1)
    theta2_runif(1,0,1)
    i_1
    while (i<16)
    {
        the1_rexp(1, theta2[i])
        if (the1<5)
        {
            the2_rexp(1, the1)
            if(the2<5)
            {
                i_i+1
                theta1[i]_the1
                theta2[i]_the2
            }
        }
    }
    resul[j]_theta1[15]
}
plot(resul, main="Muestreo de Gibbs", xlab="muestra",
ylab=expression(theta), type="l")
hist(resul,50, probability=T, col="light blue", border="dark blue",
ylab="Densidad", xlab=expression(theta), main=)
```

A5.5

```
#promedio ergódico para el ejemplo de Metropolis-Hastings
vprom_0
for (i in 1:1000)
{
    vprom[i]_mean(theta[1:i])
}
plot(vprom, xlab="iteraciones", ylab=expression(theta), type="l")
```

A5.6

```
#Graficar el promedio ergódico de la simulación para theta.del ejemplo de muestreo de Gibbs
vprom_0
for (i in 1:500)
{
    vprom[i]_mean(resul[1:i])
}
plot(vprom, xlab="iteraciones", ylab=expression(theta), type="l")
```

A6.1

```
#Histograma de datos
datos_0
datos<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24, 25, 27, 22, 25, 29, 34, 28, 26, 23)
hist(datos, 6, probability=T, col="light blue", border="dark blue",
xlab="Número de accidentes",
ylab="densidad", main=" ")
```

A6.2

```
#Gráfica de la distribución a posteriori
x_0
x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24, 25, 27, 22, 25, 29, 34, 28, 26, 23)
theta.seq<-qgamma(seq(0.001,.999,len=1000),(sum(x)+(1/2)),28)
y_dgamma(theta.seq,(sum(x)+(1/2)),28)
plot(theta.seq,y,xlab=expression(theta),ylab=,main=,type="l")
```

A6.3

```
x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24, 25, 27, 22, 25, 29, 34, 28, 26, 23)
#media analítica
mediaa_(sum(x)+(1/2))/28
mediaa
#varianza analítica
varianza_(sum(x)+(1/2))/(28^2)
varianza
qgamma(0.025,sum(x)+(1/2),28)
qgamma(0.975,sum(x)+(1/2),28)
```

A6.4

```
# Gráfica de la predictiva a posteriori
p.aposteriori<-function(xp,x,n)
{
```

```
        a<-((1/2)+sum(x))*log(n)+lgamma((1/2)+xp+sum(x))
             -lgamma(xp+(1/2))-lgamma((1/2)+sum(x))-( (1/2)
             +xp+sum(x))*(log(n+1))
        return(exp(a))
    }
    x.aposteriori<-seq(0,50, length=100)
    x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24,
25, 27, 22, 25, 29, 34, 28, 26, 23)
    plot(x.aposteriori, p.aposteriori(x.aposteriori, x,28), xlab=expression(theta),
    ylab=, type="l")
    #valor esperado y varianza de la distribución predictiva a posteriori
```

$$E\left(x^{*} \mid \theta\right)=\frac{1 \cdot\left(n \overline{x}+\frac{1}{2}\right)}{n}, \quad Var\left(x^{*} \mid \theta\right)=\frac{1 \cdot\left(n \overline{x}+\frac{1}{2}\right)(\beta+1)}{n^{2}}$$

```
    espe_(sum(x)+0.5)/28
    espe
    vvar_((sum(x)+0.5)*29)/(28^2)
    vvar
    #Intervalo predictivo
    qnbinom(0.025,sum(x)+0.5,28/29)
    qnbinom(0.975,sum(x)+0.5,28/29)
```

A6.5

```
    #Cálculo de la media a por muestreo directo
    m_10000
    x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24,
25, 27, 22, 25, 29, 34, 28, 26, 23)
    mu_rgamma(m,(sum(x)+(1/2)),28)
    mean(mu)
    var(mu)
    quantile(mu,0.025)
    quantile(mu,0.975)
```

A6.6

```
    #Cálculo de la media a posteriori mediante muestreo importante
    x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24,
25, 27, 22, 25, 29, 34, 28, 26, 23)
    n_28
    mmed_(sum(x)-0.5)/n
    desvva_sqrt((sum(x)-0.5)/(n^2))
    theta_rnorm(10000,mmed,desvva)
    p1_dpois(sum(x),28*theta)
```

```
p2_1/sqrt(theta)
p3_dnorm(theta,mmed,desvva)
p4_(p1*p2)/p3
esperanza_sum(theta*p4)/sum(p4)
esperanza
espcuad_sum((theta^2)*p4)/sum(p4)
vvar_espcuad-(esperanza^2)
vvar
#intervalo creíble
a_esperanza-1.96*sqrt(vvar)
a
b_esperanza+1.96*sqrt(vvar)
b
```

A6.7

```
#Muestreo de Gibbs, periodo de calentamiento
x_0
x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24,
25, 27, 22, 25, 29, 34, 28, 26, 23)
for (i in 1:10000)
{
    theta[i]_rgamma(1,(sum(x)+(1/2)),28)
}
thetar_theta[5000:10000]
#Graficar el promedio ergódico de la simulación para theta.
vprom_0
for (i in 1:5000)
{
    vprom[i]_mean(thetar[1:i])
}
plot(vprom, xlab="iteraciones", ylab=expression(theta), type="l")
```

A6.8

```
#múltiples cadenas
x_0
x<-c(33, 29, 32, 41, 36, 29, 21, 20, 24, 25, 31, 22, 21, 26, 20, 16, 22, 17, 24,
25, 27, 22, 25, 29, 34, 28, 26, 23)
for (j in 1:50)
{
    for (i in 1:2500)
    {
        theta[i]_rgamma(1,(sum(x)+(1/2)),28)
```

```
    }
    thetar[j]_theta[2500]
}
hist(thetar,50,probability=T,col="light blue", border="dark blue",
xlab=expression(theta),ylab="",main=" ")
theta.seq<-qgamma(seq(0.001,.999,len=1000),(sum(x)+(1/2)),28)
y_dgamma(theta.seq,(sum(x)+(1/2)),28)
lines(theta.seq,y)
```

A6.9

```
#media y varianza por Muestreo de Gibbs
mean(thetar)
var(thetar)
quantile(thetar,0.025)
quantile(thetar,0.975)
```

Bibliografía

[1] Alamilla López Norma Edith. (2004). Selección bayesiana de modelos: Una aplicación a los modelos de regresión lineal heteroscedásticos usando factores de Bayes intrínsecos. Tesis de maestria. IIMAS-UNAM.

[2] Berger, James O. (1985). Statistical Decision Theory and Bayesian Analysis. Segunda edición, Springer-Verlag.

[3] Bernardo, José M. y Smith, Adrian F. M. (1994), Bayesian Theory. John Wiley & Sons Ltd.

[4] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. J. Roy. Statist. Soc. B, **41**, 113-147.

[5] Carlin, B. P. and Louis, T. A. (2000). Bayes and Empirical Methods for Data Analysis, Second edition. Chapman & Hall/CRC.

[6] Casella, George y Berger, Roger I. (2002). Statistical Inference. Duxbury/Tomson Learning.

[7] Casella, George y George, Edward I. (1992). Explaining the Gibbs Sampler. American Statistical Association, 167-174.

[8] Chen, Ming-Hui., Shao, Qi-Man., Ibrahim, Joseph G. (2000). Monte Carlo Methods in Bayesian Computation. Springer-Verlag New York, Inc.

[9] Devroye, L. 1986. Grid Methods in Simulation and Random Variate Generation. j-COMPUTING. **37**, 71-84.

[10] Gelman, Andrew., Carlin, John B., Stern, Hal S., Rubin, Donald B. (1995). Bayesian Data Analysis. Chapman and Hall.

[11] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6): 721-741.

[12] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85 398–409.

[13] Grimmett, Geoffrey and Stirzaker, David R. (2001). Probability and Random Processes. Oxford University Press.

[14] Gutiérrez Peña, Eduardo. Métodos Computacionales en la inferencia Bayesiana. IIMAS, UNAM.

[15] Hastings W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97-109.

[16] Pérez Pérez, Olivia. (1996). Introducción a la teoría de la decisión Estadística. Tesis de licenciatura. Facultad de Ciencias, UNAM.

[17] Robert, Christian P. (2001). The Bayesian Choise. Springer-Verlag.

[18] R. L. Wolpert (1991) Monte Carlo Integration in Bayesian Statistical Analysis. Contemporary Mathematics 115,101-116.

[19] Robert, C. P., y G. Casella (1999). Monte Carlo Statistical Methods. Springer Verlag.

[20] Ripley, Brian D. (1987). Stochastic Simulation. Jonh Wiley & Sons.

[21] Secretaría de Comunicaciones y Transportes, Instituto Mexicano del Transporte. Elementos para el análisis de la seguridad en el transporte aéreo comercial en México. Publicación Técnica No. 152. Sanfandila, Qro. 2000.

[22] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society B 55, 3-23.

[23] Simith, A. F. M. and Gelfand, A. E. (1992). Bayesian Statistics Without Tears: a Sampling Resampling Perspective. American Statistician 46, 84-88.

[24] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). Annals of Satistics 22, 1701-1762.

[25] Tierney, L and Kadene, J. B. (1986). Accurate approximations for posterior moments and marginal densities. J. Amer. Statist Assoc, **81**, 82-86.

[26] Walsh 2004. Markov Chain Monte Carlo and Gibbs Sampling. Lecture Notes for EEB 581, version 26 April 2004.