Para cada uno de los problemas que se proponen, calcular la estimación requerida mediante el algoritmo de Montecarlo por cadenas de Markov especificado. Aplicar el método de las medias por lotes para construir intervalos de confianza con probabilidad de cobertura $95\,\%$ de las estimaciones realizadas. Documentar en *RMarkdown* todo el proceso llevado a cabo para resolver cada problema.

Puede ser de utilidad considerar esta página de *Wikipedia*, en la que puede encontrarse un listado de distintas distribuciones de probabilidad, y este paquete de *R*, que permite manejar distribuciones de probabilidad adicionales a las estándar.

Los problemas propuestos son los siguientes:

1. Consideremos una distribución Beta$(2.7, 6.3)$ truncada al intervalo (c, d), con $c = 0.1$ y $d = 0.9$. Es decir, su función de densidad es

$$f(x) \propto \begin{cases} x^{2.7-1}(1-x)^{6.3-1} & \text{si } c < x < d, \\ 0 & \text{en otro caso.} \end{cases}$$

   Se pide comparar con esa función de densidad un histograma de los valores generados

   - mediante el algoritmo de Metropolis-Hastings usando Beta$(2, 6)$ como muestreador independiente;
   - mediante el algoritmo de Metropolis-Hastings usando Unif(c, d) como muestreador independiente;
   - mediante un paseo aleatorio de Metropolis, haciendo uso del paquete `mcmc`.

   Para una correcta construcción del gráfico comparativo, téngase en cuenta que la constante de normalización de f es $1/B(2.7, 6.3)(F(d) - F(c))$, donde F es la función de distribución de Beta$(2.7, 6.3)$ (la función B está implementada mediante la función `beta` de *R*).

   En cada uno de los tres apartados anteriores, estimar el valor de $\mathbb{E}_f[\boldsymbol{X}]$.

   Repetir el ejercicio tomando $c = 0.25$ y $d = 0.75$.

2. Consideremos el vector aleatorio $\boldsymbol{X} = (X_1, X_2)$ con función de densidad

$$f(x_1, x_2) \propto \mathrm{e}^{-\frac{x_1^2 x_2^2 + x_1^2 + x_2^2 - 8x_1 - 8x_2}{2}}$$

   para todo $x_1, x_2 \in \mathbb{R}$.

   - Haciendo uso del paquete `mcmc`, estimar el valor de $\mathbb{E}_f[\boldsymbol{X}]$ mediante un paseo aleatorio de Metropolis.
   - Demostrar que las densidades condicionales completas, $f_1(x_1 \mid x_2)$ y $f_2(x_2 \mid x_1)$, corresponden a distribuciones normales, determinando su esperanza y su varianza. Aplicar el muestreador de Gibbs para estimar el valor de $\mathbb{E}_f[\boldsymbol{X}]$.

3. En un estudio médico sobre infecciones causadas por partos por cesárea se han considerado tres factores de influencia: un indicador acerca de si la cesárea estaba o no planificada; un indicador acerca de si se presentaban factores de riesgo adicionales en el momento del parto; y un indicador acerca de si se habían proporcionado antibióticos como profilaxis (es decir, como medida preventiva).

   Los datos obtenidos se facilitan en el cuadro 1, donde la respuesta y_i es el número de infecciones que se observaron entre las n_i pacientes sujetas a los mismos factores de influencia.

Número de partos		Cesárea planificada	Factores de riesgo	Antibióticos
Con infección	Total			
y_i	n_i	z_{i1}	z_{i2}	z_{i3}
11	98	1	1	1
1	18	0	1	1
0	2	0	0	1
23	26	1	1	0
28	58	0	1	0
0	9	1	0	0
8	40	0	0	0

Cuadro 1: Datos acerca de infecciones causadas por cesáreas

   El conjunto de datos anterior se puede modelizar usando un *modelo probit*, en el que, dados unos determinados factores de influencia z_{i1}, z_{i2} y z_{i3}, la probabilidad de que una cesárea cause una infección viene dada por

$$p_i = \Phi(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3})$$

   donde Φ es la función de distribución de la distribución normal estándar $\mathrm{N}(0, 1)$ (es decir, $\Phi(x) = \mathbb{P}(Z \leq x)$ y se puede calcular con la función `pnorm` de *R*, estableciendo el argumento `log.p` a `TRUE` si se desea calcular $\log \Phi$).

   Por tanto, tomando $\boldsymbol{z}_i = (1, z_{i1}, z_{i2}, z_{i3})^{\mathrm{T}}$ y $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, se tiene que

$$Y_i \sim \mathrm{Binom}(n_i, p_i), \qquad \text{con } p_i = \Phi\big(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\beta}\big)$$

   para cada combinación de factores de influencia.

   Una distribución a priori adecuada para los parámetros de interés es $\beta_i \overset{\text{i. i. d.}}{\sim} \mathrm{N}\Big(0, \frac{1}{\lambda}\Big)$, para un cierto hiperparámetro λ. La densidad a posteriori de $\boldsymbol{\beta}$ es entonces

$$f_{\mathrm{pos}}(\boldsymbol{\beta} \mid y_1, \dots, y_7) \propto \left(\prod_{i=1}^{7} p_i^{y_i} (1 - p_i)^{n_i - y_i} \right) \mathrm{e}^{-\frac{\lambda}{2} \sum_{j=0}^{3} \beta_j^2}$$

   Se pide estimar mediante un paseo aleatorio de Metropolis los valores esperados a posteriori de los parámetros β_i para los datos proporcionados. Para ello, tomar $\lambda = 10$ como valor del hiperparámetro.

4. El cuadro 2 recoge, para diez bombas refrigerantes de una planta nuclear, el número de fallos observados así como el tiempo a lo largo del cual se han observado esos fallos.

Bomba	1	2	3	4	5	6	7	8	9	10
Fallos (F_i)	5	1	5	14	3	19	1	1	4	22
Tiempo (T_i)	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

Cuadro 2: Datos acerca de fallos de diez bombas de una planta nuclear

Si consideramos que el número de fallos por unidad de tiempo de la bomba i-ésima sigue una distribución de Poisson con parámetro λ_i, el conjunto de datos anterior se puede modelizar a priori usando el siguiente *modelo jerárquico*:

$$\begin{aligned} F_i &\sim \text{Pois}(\lambda_i T_i) \\ \lambda_i &\sim \text{Gamma}(\alpha, \beta) \\ \beta &\sim \text{Gamma}(\gamma, \delta) \end{aligned}$$

donde α, γ y δ son los hiperparámetros del modelo.

La densidad a posteriori de $\boldsymbol{\lambda}$ y β es, entonces

$$\begin{aligned} f_{\text{pos}}(\boldsymbol{\lambda}, \beta \mid F_1, \dots, F_{10}, T_1, \dots, T_{10}) &\propto \left(\prod_{i=1}^{10} (\lambda_i T_i)^{F_i} \mathrm{e}^{-\lambda_i T_i}\right)\left(\prod_{i=1}^{10} \beta^{\alpha} \lambda_i^{\alpha-1} \mathrm{e}^{-\beta\lambda_i}\right)\beta^{\gamma-1}\mathrm{e}^{-\delta\beta} \\ &= \left(\prod_{i=1}^{10} \lambda_i^{F_i+\alpha-1} \mathrm{e}^{-(T_i+\beta)\lambda_i}\right)\beta^{10\alpha+\gamma-1}\mathrm{e}^{-\delta\beta} \end{aligned}$$

Se pide estimar, mediante un muestreador de Gibbs, los valores esperados a posteriori de los parámetros λ_i para los datos proporcionados. Para ello tomar $\alpha = 1.8, \gamma = 0.01$ y $\delta = 1$ como valores de los hiperparámetros.

5. El conjunto de datos `cars` relaciona la distancia de frenado y con la velocidad x de una muestra de coches. Este conjunto de datos se puede modelizar con un modelo lineal cuadrático

$$y_i = a + bx_i + cx_i^2 + \varepsilon_i$$

donde asumimos que $\varepsilon_i \overset{\text{i. i. d.}}{\sim} \mathrm{N}(0, \sigma^2)$.

Asumiendo una distribución a priori uniforme para los parámetros, su densidad a posteriori es

$$f_{\text{pos}}\left(a, b, c, \frac{1}{\sigma^2} \,\middle|\, \boldsymbol{x}, \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{N/2} \mathrm{e}^{-\frac{\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2}{2\sigma^2}}$$

donde N es el tamaño de la muestra de coches.

Se pide estimar los valores esperados a posteriori de los parámetros a, b y c

- mediante un paseo aleatorio de Metropolis, haciendo uso del paquete `mcmc`;
- mediante un muestreador de Gibbs.

2. Consideremos el vector aleatorio $\boldsymbol{X} = (X_1, X_2)$ con función de densidad

$$f(x_1, x_2) \propto \mathrm{e}^{-\frac{x_1^2 x_2^2 + x_1^2 + x_2^2 - 8x_1 - 8x_2}{2}}$$

para todo $x_1, x_2 \in \mathbb{R}$.

- Haciendo uso del paquete `mcmc`, estimar el valor de $\mathbb{E}_f[\boldsymbol{X}]$ mediante un paseo aleatorio de Metropolis.
- Demostrar que las densidades condicionales completas, $f_1(x_1 \mid x_2)$ y $f_2(x_2 \mid x_1)$, corresponden a distribuciones normales, determinando su esperanza y su varianza. Aplicar el muestreador de Gibbs para estimar el valor de $\mathbb{E}_f[\boldsymbol{X}]$.

$$f_1(x_1 \mid x_2) \propto \exp\left\{-\frac{1}{2}\left(x_1^2(x_2^2+1) - 8x_1\right)\right\}$$

$$= \exp\left\{-\frac{(x_2^2+1)}{2}\left(x_1^2 - \frac{8}{(x_2^2+1)}x_1\right)\right\}$$

$$\propto \exp\left\{-\frac{(x_2^2+1)}{2}\left(x_1^2 - \frac{4}{(x_2^2+1)}\right)^2\right\}$$

$$x_1 \mid x_2 \sim N\left(\frac{4}{x_2^2+1}, \frac{1}{x_2^2+1}\right)$$

$$x_2 \mid x_1 \sim N\left(\frac{4}{x_1^2+1}, \frac{1}{x_1^2+1}\right)$$

3. En un estudio médico sobre infecciones causadas por partos por cesárea se han considerado tres factores de influencia: un indicador acerca de si la cesárea estaba o no planificada; un indicador acerca de si se presentaban factores de riesgo adicionales en el momento del parto; y un indicador acerca de si se habían proporcionado antibióticos como profilaxis (es decir, como medida preventiva).

   Los datos obtenidos se facilitan en el cuadro 1, donde la respuesta y_i es el número de infecciones que se observaron entre las n_i pacientes sujetas a los mismos factores de influencia.

Número de partos		Cesárea planificada	Factores de riesgo	Antibióticos
Con infección	Total			
y_i	n_i	z_{i1}	z_{i2}	z_{i3}
11	98	1	1	1
1	18	0	1	1
0	2	0	0	1
23	26	1	1	0
28	58	0	1	0
0	9	1	0	0
8	40	0	0	0

Cuadro 1: Datos acerca de infecciones causadas por cesáreas

El conjunto de datos anterior se puede modelizar usando un *modelo probit*, en el que, dados unos determinados factores de influencia z_{i1}, z_{i2} y z_{i3}, la probabilidad de que una cesárea cause una infección viene dada por

$$p_i = \Phi(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3})$$

donde Φ es la función de distribución de la distribución normal estándar $\mathrm{N}(0,1)$ (es decir, $\Phi(x) = \mathbb{P}(Z \leq x)$ y se puede calcular con la función `pnorm` de *R*, estableciendo el argumento `log.p` a TRUE si se desea calcular $\log \Phi$).

Por tanto, tomando $\boldsymbol{z}_i = (1, z_{i1}, z_{i2}, z_{i3})^{\mathrm{T}}$ y $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, se tiene que

$$Y_i \sim \mathrm{Binom}(n_i, p_i), \qquad \text{con } p_i = \Phi(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\beta})$$

para cada combinación de factores de influencia.

Una distribución a priori adecuada para los parámetros de interés es $\beta_i \overset{\text{i. i. d.}}{\sim} \mathrm{N}\left(0, \frac{1}{\lambda}\right)$, para un cierto hiperparámetro λ. La densidad a posteriori de $\boldsymbol{\beta}$ es entonces

$$f_{\text{pos}}(\boldsymbol{\beta} \mid y_1, \dots, y_7) \propto \left(\prod_{i=1}^{7} p_i^{y_i} (1-p_i)^{n_i - y_i} \right) \mathrm{e}^{-\frac{\lambda}{2} \sum_{j=0}^{3} \beta_j^2}$$

Se pide estimar mediante un paseo aleatorio de Metropolis los valores esperados a posteriori de los parámetros β_i para los datos proporcionados. Para ello, tomar $\lambda = 10$ como valor del hiperparámetro.

$$\log\left(f_{pos}(\beta \mid y)\right) \propto \log\left(\prod_{i=1}^{7} p^{y_i}(1-p_i)^{n_i - y_i}\right) - \frac{\lambda}{2}\sum_{j=0}^{3}\beta_j^2$$

$$= \sum_{i=1}^{7}\left[y_i \log p_i + (n_i - y_i)\log(1-p_i)\right] - \frac{\lambda}{2}\sum_{j=0}^{3}\beta_j^2$$

$$\hookrightarrow \log \Phi(z_i^T \beta) \qquad \hookrightarrow \log(1-\Phi(z_i^T\beta)) \overset{!}{=} \log(\Phi(-z_i^T\beta))$$

4. El cuadro 2 recoge, para diez bombas refrigerantes de una planta nuclear, el número de fallos observados así como el tiempo a lo largo del cual se han observado esos fallos.

Bomba	1	2	3	4	5	6	7	8	9	10
Fallos (F_i)	5	1	5	14	3	19	1	1	4	22
Tiempo (T_i)	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

Cuadro 2: Datos acerca de fallos de diez bombas de una planta nuclear

Si consideramos que el número de fallos por unidad de tiempo de la bomba i-ésima sigue una distribución de Poisson con parámetro λ_i, el conjunto de datos anterior se puede modelizar a priori usando el siguiente *modelo jerárquico*:

$$\begin{aligned} F_i &\sim \text{Pois}(\lambda_i T_i) \\ \lambda_\text{i} &\sim \text{Gamma}(\alpha, \beta) \\ \beta &\sim \text{Gamma}(\gamma, \delta) \end{aligned}$$

donde α, γ y δ son los hiperparámetros del modelo.

La densidad a posteriori de $\boldsymbol{\lambda}$ y β es, entonces

$$\begin{aligned} f_{\text{pos}}(\boldsymbol{\lambda}, \beta \mid F_1, \dots, F_{10}, T_1, \dots, T_{10}) &\propto \left(\prod_{i=1}^{10} (\lambda_i T_i)^{F_i} e^{-\lambda_i T_i}\right)\left(\prod_{i=1}^{10} \beta^\alpha \lambda_i^{\alpha-1} e^{-\beta\lambda_i}\right)\beta^{\gamma-1}e^{-\delta\beta} \\ &= \left(\prod_{i=1}^{10} \lambda_i^{F_i+\alpha-1} e^{-(T_i+\beta)\lambda_i}\right)\beta^{10\alpha+\gamma-1}e^{-\delta\beta} \end{aligned}$$

Se pide estimar, mediante un muestreador de Gibbs, los valores esperados a posteriori de los parámetros λ_i para los datos proporcionados. Para ello tomar $\alpha = 1.8, \gamma = 0.01$ y $\delta = 1$ como valores de los hiperparámetros.

$$f(\lambda_i \mid \beta) \propto \lambda_i^{F_i+\alpha-1} \cdot \exp\{-(T_i+\beta)\lambda_i\}$$

$$\hookrightarrow \lambda_i \mid \beta \sim \text{Gamma}(F_i+\alpha, T_i+\beta)$$

$$f(\beta \mid \underline{\lambda}) \propto \beta^{10\alpha+\gamma-1} e^{-\delta\beta} \cdot \prod_{i=1}^{10} e^{-\lambda_i\beta}$$

$$= \beta^{10\alpha+\gamma-1} e^{-\beta\left(\delta+\sum_i \lambda_i\right)}$$

$$\hookrightarrow \beta \mid \underline{\lambda} \sim \text{Gamma}\left(10\alpha+\gamma,\ \delta+\sum_i \lambda_i\right)$$

5. El conjunto de datos `cars` relaciona la distancia de frenado y con la velocidad x de una muestra de coches. Este conjunto de datos se puede modelizar con un modelo lineal cuadrático

$$y_i = a + bx_i + cx_i^2 + \varepsilon_i$$

donde asumimos que $\varepsilon_i \overset{\text{i. i. d.}}{\sim} \mathrm{N}(0, \sigma^2)$.

Asumiendo una distribución a priori uniforme para los parámetros, su densidad a posteriori es

$$f_{\text{pos}}\left(a, b, c, \frac{1}{\sigma^2} \,\middle|\, \boldsymbol{x}, \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{N/2} \mathrm{e}^{-\frac{\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2}{2\sigma^2}}$$

donde N es el tamaño de la muestra de coches.

Se pide estimar los valores esperados a posteriori de los parámetros a, b y c

- mediante un paseo aleatorio de Metropolis, haciendo uso del paquete `mcmc`;
- mediante un muestreador de Gibbs.

$$f\left(a \mid b, c, \frac{1}{\sigma^2}, \underline{x}, \underline{y}\right) \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - a - bx_i - cx_i^2\right)^2 \right\}$$

$$\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(a^2 - 2a\left(y_i - bx_i - cx_i^2\right)\right) \right\}$$

$$= \exp\left\{ -\frac{N}{2\sigma^2} \left(a^2 - 2a\left(\frac{\sum_{i=1}^{N}\left(y_i - bx_i - cx_i^2\right)}{N}\right)\right) \right\}$$

$$\hookrightarrow a \mid b, c, \frac{1}{\sigma^2}, \underline{x}, \underline{y} \sim \mathrm{N}\left(\frac{\sum_{i=1}^{N}\left(y_i - bx_i - cx_i^2\right)}{N}, \frac{\sigma^2}{N}\right)$$

$$f\left(b \mid a, c, \frac{1}{\sigma^2}, \underline{x}, \underline{y}\right)$$

$$\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - a - b x_i - c x_i^2 \right)^2 \right\}$$

$$\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(b^2 x_i^2 - 2 b x_i \left(y_i - a - c x_i^2 \right) \right) \right\}$$

$$\propto \exp\left\{ -\frac{1}{2\sigma^2} \left(b^2 \sum_{i=1}^{N} x_i^2 - 2b \sum_{i=1}^{N} x_i \left(y_i - a - c x_i^2 \right) \right) \right\}$$

$$\propto \exp\left\{ -\frac{\sum_{i=1}^{N} x_i^2}{2\sigma^2} \left(b^2 - 2b \cdot \frac{\sum_{i=1}^{N} x_i \left(y_i - a - c x_i^2 \right)}{\sum_{i}^{N} x_i^2} \right) \right\}$$

$$\hookrightarrow b \mid a, c, \frac{1}{\sigma^2}, \underline{x}, \underline{y} \sim N\left(\frac{\sum_{i=1}^{N} x_i \left(y_i - a - c x_i^2 \right)}{\sum_{i}^{N} x_i^2}, \frac{\sigma^2}{\sum_{i=1}^{N} x_i} \right)$$

$$f(c \mid a, b, \sigma^2, \underline{x}, \underline{y})$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(c_i^2 x_i^4 - 2cx_i^2\left(y_i - a - bx_i\right)\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(c_i^2\sum_{i=1}^{N}x_i^4 - 2c\sum_{i=1}^{N}x_i^2\left(y_i - a - bx_i\right)\right)\right\}$$

$$\hookrightarrow c \mid \ldots \sim N\left(\frac{\sum_{i=1}^{N}\left(y_i - a - bx_i\right)x_i^2}{\sum_{i=1}^{N}x_i^4}, \frac{\sigma^2}{\sum_{i=1}^{N}x_i^4}\right)$$

$$f_{\text{pos}}\left(a,b,c,\frac{1}{\sigma^2}\,\middle|\,\boldsymbol{x},\boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{N/2} \mathrm{e}^{-\frac{\sum_{i=1}^{N}\left(y_i-a-bx_i-cx_i^2\right)^2}{2\sigma^2}}$$

$$\left(\frac{1}{\sigma^2}\right)^{N/2} \exp\left\{-\left(\frac{\sum_{i=1}^{N}\left(y_i-a-bx_i-cx_i^2\right)^2}{2}\right)\cdot\frac{1}{\sigma^2}\right\}$$

↳ $$\frac{1}{\sigma^2} \sim G\left(N/2+1,\ \frac{1}{2}\sum_{i=1}^{N}\left(y_i-a-bx_i-cx_i^2\right)^2\right)$$

$$f_{\text{pos}}\left(a, b, c, \frac{1}{\sigma^2} \,\middle|\, \boldsymbol{x}, \boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{N/2} \mathrm{e}^{-\frac{\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2}{2\sigma^2}}$$

$$\log f\left(a, b, c, \frac{1}{\sigma^2} \,\middle|\, \underline{x}, \underline{y}\right)$$

$$\propto -\frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - a - bx_i - cx_i^2\right)^2$$

$$\sum_{N=0}^{+\infty}\int_0^1 \frac{1}{(N-n^+)!} p^{n^c}(1-p)^{2N-n^c}\lambda^N \, dp$$

$$= \sum_{N=0}^{+\infty}\frac{\lambda^N}{(N-n^+)!}\underbrace{\int_0^1 p^{n_c}(1-p)^{2N-n_c}\, dp}_{=}$$