Contenido

Capítulo 1

Tópicos básicos

1.1 Teoría de la decisión

El problema estadístico de estimar un parámetro se puede ver dentro del contexto de la teoría de decisión: la estimación que proveemos, sea en el ámbito de la estadística clásica o la estadística bayesiana, depende de los datos muestrales, $\mathbf{X}$, de tal forma que si éstos cambian, nuestra estimación también cambia. De esta manera, el proceso de estimación puede ser representado como una función que toma un conjunto de datos muestrales y los convierte en una estimación de nuestro parámetro de interés, $A(\mathbf{X})$ o simplemente A. En la teoría de decisión, la anterior función se conoce como una regla de decisión.

Así como en la vida cotidiana, por la incertidumbre del futuro(en el ámbito estadístico, por la incertidumbre acerca del parámetro), toda acción que uno toma (toda estimación que uno provea) puede traer consigo un grado de falla o riesgo. Y es necesario tomar la acción óptima que de alguna forma minimice ese riesgo. Formalizando esta idea intuitiva, tenemos la función de pérdida L que asocia cada dupla de la acción tomada y el parámetro de interés θ, $(A,\ \theta)$ con un número no negativo que cuantifica la pérdida que ocasiona la acción (o la estimación) A con respecto al parámetro θ.

Es claro que se desea escoger aquella acción que minimice de alguna forma la pérdida que ésta ocasiona, pero la función L no se puede minimizar directamente, puesto que:

- En el ámbito de la estadística clásica, el parámetro θ se considera fijo, y los datos muestrales $\mathbf{X}$ aleatorios, así como la función de pérdida L depende de $\mathbf{X}$, entonces ésta también será una variable aleatoria, y no se puede minimizar directamente. Por lo tanto se define el riesgo o la pérdida promedio como la esperanza matemática de L; denotando el riesgo como R, éste está definido como $R = E(L)$ (la esperanza se toma con respecto a la distribución probabilística de $\mathbf{X}$).
- En el ámbito de la estadística bayesiana, θ es una cantidad aleatoria, y la herramienta fundamental para conocer características de θ es su función de densidad posterior $p(\theta|\mathbf{X})$. En este caso, el riesgo R se define como
$$R = E(L) = \int L(A,\theta)p(\theta|\mathbf{X})d\theta$$

En cualquier de los dos casos anteriores, buscaremos la estimación que minimice el riesgo R. Ilustramos los anteriores conceptos en los siguientes ejemplos tanto en la estadística clásica como en la estadística bayesiana.

Ejemplo 1.1.1. Sea X_i con $i = 1, \cdots, n$ una muestra aleatoria con media θ y varianza σ^2, ambas fijas, y suponga que se desea encontrar el mejor estimador de θ bajo la función de pérdida cuadrática

dada por

$$L(A,\theta) = (A-\theta)^2$$

cuyo riesgo asociado está dado por $R = E(A-\theta)^2$. En primer lugar buscaremos dicho estimador dentro de todas las formas lineales de X_i, es decir, los estimadores de la forma $A = \sum_{i=1}^n c_i X_i$, de esta forma, el riesgo se puede expresar como

$$\begin{aligned} R = E(A-\theta)^2 &= Var(A) + (E(A)-\theta)^2 \\ &= \sum_{i=1}^n c_i^2\sigma^2 + \theta^2(\sum_{i=1}^n c_i - 1)^2 \end{aligned}$$

Y al buscar los coeficientes c_i que minimizan la anterior expresión, encontramos que $c_i = \theta^2/(\sigma^2 + n\theta^2)$ para todo i. Como estos coeficientes conducen a un estimador que depende del parámetro desconocido, concluimos que no hay ningún estimador que minimiza el riesgo.

Para encontrar una solución, es necesario restringir aún más el rango de estimadores, para eso, se restringe que $\sum_{i=1}^n c_i = 1$, de esta forma el riesgo está dado por $R = \sum c_i^2\sigma^2$, y al minimizar $\sum c_i^2$ sujeto a la restricción de $\sum c_i = 1$. La solución está dada por $c_i = 1/n$ para todo i, y así encontramos que el mejor estimador (en el sentido de minimizar el riesgo de la función de pérdida cuadrática) dentro de todas formas lineales con $\sum c_i = 1$ es la media muestral $\bar{X}$.

Ejemplo 1.1.2. Suponga que se desea estimar un parámetro de interés θ en el contexto de la estadística bayesiana y denotamos la función de densidad posterior de θ como $p(\theta|\mathbf{X})$, entonces si utilizamos la función de pérdida cuadrática, entonces el riesgo asociado será

$$R = E(L(A,\theta)) = E(A-\theta)^2 = Var(\theta) + (E(\theta) - A)^2$$

que es minimizado si $A = E(\theta)$. Es decir la mejor acción para estimar θ es utilizar la esperanza de θ tomada con respecto a la distribución posterior $p(\theta|\mathbf{X})$.

Ejemplo 1.1.3. En el mismo contexto del ejemplo anterior, si cambiamos la función de pérdida a la siguiente

$$L(A,\theta) = |A-\theta| = (A-\theta)I_{(A\geq\theta)} + (\theta - A)I_{(\theta > A)}$$

Y el riesgo está dado por

$$\begin{aligned} R &= E(L(A,\theta)) \\ &= \int L(A,\theta)p(\theta|\mathbf{X})d\theta \\ &= \int_{(A\geq\theta)} (A-\theta)p(\theta|\mathbf{X})d\theta + \int_{(\theta>A)} (\theta - A)p(\theta|\mathbf{X})d\theta \end{aligned}$$

Derivando el riesgo con respecto a la acción A, se tiene que

$$\frac{\partial R}{\partial A} = \int_{(A\geq\theta)} p(\theta|\mathbf{X})d\theta - \int_{(\theta>A)} p(\theta|\mathbf{X})d\theta$$

Igualando a cero, tenemos que

$$\int_{(A\geq\theta)} p(\theta|\mathbf{X})d\theta = \int_{(\theta>A)} p(\theta|\mathbf{X})d\theta = 0.5$$

Y concluimos que la acción A que induce menor riesgo corresponde al percentil 50 % o la mediana de la distribución posterior de θ.

De los anteriores ejemplos vemos que bajo un mismo contexto, cuando se utilizan diferentes funciones de pérdidas, también obtenemos distintas estimaciones.

1.2 Algunos resultados de probabilidad

A continuación se presentan definiciones y resultados de probabilidad en términos de notación se utilizará indistintamente la expresión de integral, $\int$, que implicará la integral, en el caso de las variables aleatorias continuas, o la sumatoria, en el caso de las variables aleatorias discretas.

Definición 1.2.1. *Sean* $\mathbf{X} = (X_1, \ldots, X_p)'$, $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ *dos vectores aleatorios definidos sobre los espacios de muestreo* $\mathcal{X}$, $\mathcal{Y}$, *respectivamente. Suponga que la distribución conjunta de estos vectores aleatorios está dada por* $p(\mathbf{X}, \mathbf{Y})$. *La distribución marginal de* $\mathbf{X}$ *está dada por*

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{Y}) \, d\mathbf{Y} \tag{1.2.1}$$

y la distribución condicional de $\mathbf{X}$ *dado* $\mathbf{Y}$ *como*

$$p(\mathbf{X} \mid \mathbf{Y}) = \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{Y})} \tag{1.2.2}$$

Resultado 1.2.1. *Suponga los vectores* $\mathbf{X}$, $\mathbf{Y}$ *y un tercer vector* $\mathbf{Z} = (Z_1, \ldots, Z_r)'$ *definido sobre el espacio de muestreo* $\mathcal{Z}$. *Entonces se tiene que*

$$p(\mathbf{X} \mid \mathbf{Z}) = \int p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \, d\mathbf{Y} \tag{1.2.3}$$

y

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} \tag{1.2.4}$$

Prueba. En primer lugar, nótese que

$$\begin{aligned} \int p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \, d\mathbf{Y} &= \int \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})} \, d\mathbf{Y} \\ &= \frac{1}{p(\mathbf{Z})} \int p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \, d\mathbf{Y} \\ &= \frac{1}{p(\mathbf{Z})} p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z}) \end{aligned}$$

Por otro lado,

$$\frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} = \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})} \Big/ \frac{p(\mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})} = \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Y}, \mathbf{Z})} = p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z})$$

■

Definición 1.2.2. *Sean* $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ *vectores aleatorios, se dice que* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *con respecto a* $\mathbf{Z}$ *si satisfacen la siguiente expresión*

$$p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z}) p(\mathbf{Y} \mid \mathbf{Z}) \tag{1.2.5}$$

Resultado 1.2.2. *Si* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *con respecto a* $\mathbf{Z}$, *entonces se tiene que*

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z}) \tag{1.2.6}$$

Prueba. Como $p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = \dfrac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Z})}$, entonces

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{p(\mathbf{Y}, \mathbf{Z})} = \frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})p(\mathbf{Z})}{p(\mathbf{Y}, \mathbf{Z})} = \frac{p(\mathbf{X} \mid \mathbf{Z})p(\mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} = p(\mathbf{X} \mid \mathbf{Z})$$

■

Resultado 1.2.3. *Si* $\mathbf{X}$ *es independiente de* $\mathbf{Y}$*, entonces* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *dada cualquier otro vector, digamos* $\mathbf{Z}$*.*

Prueba. Nótese que

$$p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z})p(\mathbf{Y} \mid \mathbf{Z}) = p(\mathbf{X} \mid \mathbf{Z})p(\mathbf{Y} \mid \mathbf{Z})$$

puesto que, utilizando la hipótesis de independencia, se tiene que

$$p(\mathbf{X} \mid \mathbf{Y}) = p(\mathbf{X})$$

■

1.3 Teorema de Bayes

Desde la revolución estadística de Pearson y Fisher, la inferencia estadística busca encontrar los valores que parametrizan a la distribución desconocida de los datos. El primer enfoque, propuesto por Pearson, afirmaba que si era posible observar a la variable de interés en todos y cada uno de los individuos de una población, entonces era posible calcular los parámetros de la distribución de la variable de interés; por otro lado, si sólo se tenía acceso a una muestra representativa, entonces era posible calcular una estimación de tales parámetros. Sin embargo, Fisher discrepó de tales argumentos, asumiendo que las observaciones están sujetas a un error de medición y por lo tanto, así se tuviese acceso a toda la población, es imposible calcular los parámetros de la distribución de la variable de interés.

Del planteamiento de Fisher resultaron una multitud de métodos estadísticos para la estimación de los parámetros poblacionales. Es decir, si la distribución de $\mathbf{Y}$ está parametrizada por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, $\boldsymbol{\theta} \in \Theta$ con Θ el espacio paramétrico inducido por el comportamiento de la variable de interés, el objetivo de la teoría estadística inferencial es calcular una estimación $\hat{\boldsymbol{\theta}}$ del parámetro $\boldsymbol{\theta}$ por medio de los datos observados. En este enfoque, los parámetros se consideran cantidades fijas y constantes. Sin embargo, en la última mitad del siglo XX, algunos investigadores estadísticos comenzaron a reflexionar acerca de la naturaleza de $\boldsymbol{\theta}$ y enfocaron la inferencia estadística de una manera distinta: asumiendo que la distribución de la variable de interés está condicionada a valores específicos de los parámetros. Es decir, en términos de notación, si la variable de interés es $\mathbf{Y}$, su distribución condicionada a los parámetros toma la siguiente forma $p(\mathbf{Y} \mid \boldsymbol{\theta})$. Esto implica claramente que en este nuevo enfoque la naturaleza de los parámetros no es constante sino estocástica.

En términos de inferencia para $\boldsymbol{\theta}$, es necesario encontrar la distribución de los parámetros condicionada a la observación de los datos. Para este fin, es necesario definir la distribución conjunta de la variable de interés con el vector de parámetros.

$$p(\boldsymbol{\theta}, \mathbf{Y}) = p(\boldsymbol{\theta})p(\mathbf{Y} \mid \boldsymbol{\theta})$$

A la distribución $p(\boldsymbol{\theta})$ se le conoce con el nombre de distribución *previa* y en ella se enmarcan todas y cada una de las creencias que se tienen acerca del comportamiento estocástico del vector de parámetros antes de que ocurra la recolección de los datos y $p(\mathbf{Y} \mid \boldsymbol{\theta})$ es la distribución de muestreo o verosimilitud

o distribución de los datos. Por otro lado, la distribución del vector de parámetros condicionada a los datos observados está dada por

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \mathbf{Y})}{p(\mathbf{Y})} = \frac{p(\boldsymbol{\theta})p(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})} \tag{1.3.1}$$

A la distribución $p(\boldsymbol{\theta} \mid \mathbf{Y})$ se le conoce con el nombre de distribución *posterior* y en ella se enmarcan las creencias actualizadas acerca del comportamiento estocástico del vector de parámetros teniendo en cuenta los datos observados $\mathbf{Y}$. Nótese que la expresión (1.3.1) se compone de una fracción cuyo denominador no depende del vector de parámetros y considerando a los datos observados como fijos, corresponde a una constante y puede ser obviada. Por lo tanto, otra representación de la regla de Bayes está dada por

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.3.2}$$

Gelman, Carlin, Stern & Rubin (2003) menciona que esta expresión se conoce como la distribución *a posterior no-normalizada* y encierra el núcleo técnico de la inferencia bayesiana. La constante $p(\mathbf{Y})$ faltante en la expresión 1.3.2 se da a continuación:

Resultado 1.3.1. *La expresión $p(\mathbf{Y})$ corresponde a una constante k tal que*

$$k = p(\mathbf{Y}) = E_{\boldsymbol{\theta}}[p(Y \mid \boldsymbol{\theta})]$$

Prueba. Nótese que

$$k = p(\mathbf{Y}) = \int p(\mathbf{Y}, \boldsymbol{\theta})\ d\boldsymbol{\theta} = \int p(\boldsymbol{\theta})p(\mathbf{Y} \mid \boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

entonces

$$\begin{aligned} k &= \int p(\mathbf{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &= E_{\boldsymbol{\theta}}[p(Y \mid \boldsymbol{\theta})] \end{aligned}$$

■

Curiosamente, el reverendo Thomas Bayes nunca publicó este resultado, sino que después de su fallecimiento, su amigo, el filósofo Richard Price, encontró los escritos dentro de sus pertenencias, y éstos fueron publicados en el 1764 en *Philosophical Transactions of the Royal Society of London.* Aunque el teorema de Bayes fue nombredo a honor de Thomas Bayes, estamos casi seguros que de que él mismo no sospechaba del gran impacto de este hermoso resultado. De hecho, aproximadamente una década más tarde el gran Pierre-Simon Laplace también descrubrió el mismo principio, y dedicó gran parte de su vida extendiéndolo y formalizándolo. Más aún, él analizó grandes volumenes de datos relacionados a los nacimientos en diferentes paises para confirmar esta teoría, y sentó las bases de ésta. A continuación se presenta un ejemplo simple de este sencillo pero poderoso teorema.

Ejemplo 1.3.1. Uno de los primeros acercamientos de cualquier profesional a la estadística bayesiana se da en un curso básico de probabilidades en donde el docente presenta con cierta rigurosidad los conceptos básicos e introductorios de la teoría de probabilidad. En un sobrevuelo de tales conceptos es posible recordar términos como experimento, espacio muestral, función de probabilidad y sigma álgebra. Justo después del repaso de rigor acerca de los axiomas de probabilidades y sus teoremas más significativos, el curso da una curva cerrada y el alumno es introducido en conceptos más profundos como la probabilidad condicional.

En estos tópicos, tanto el maestro como el alumno asumen que los temas básicos ya están entendidos y que no existe necesidad de volver atrás. A manera de introducción, los autores desean hacer notar a los

lectores que requieren de herramientas de modelamiento más sofisticadas, que es necesario volver atrás - al menos en esta primera página - para sentar las bases de la autopista de alta velocidad como lo es el análisis bayesiano. No tiene sentido que el investigador utilice las poderosas herramientas bayesianas si no entiende que sus bases probabilísticas están bien sustentadas.

Para entrar en detalle, vamos a utilizar un ejemplo en donde el lector se sentirá identificado con aquellas épocas universitarias de un curso de probabilidades: suponga que una fábrica del sector industrial produce bolígrafos y que la producción está a cargo de tres máquinas. La primera máquina produce el 50 % del total de bolígrafos en el año, la segunda máquina produce el 30 % y la última maquina produce el restante 20 %. Por supuesto, esta producción esta sujeta al error y por tanto, basados en la experiencia, es posible reconocer que, de los artículos producidos por la primera máquina, el 5 % resultan defectuosos; de los artículos producidos por la segunda máquina, el 2 % resultan defectuosos y , de los artículos producidos por la última máquina, el 6 % resultan defectuosos.

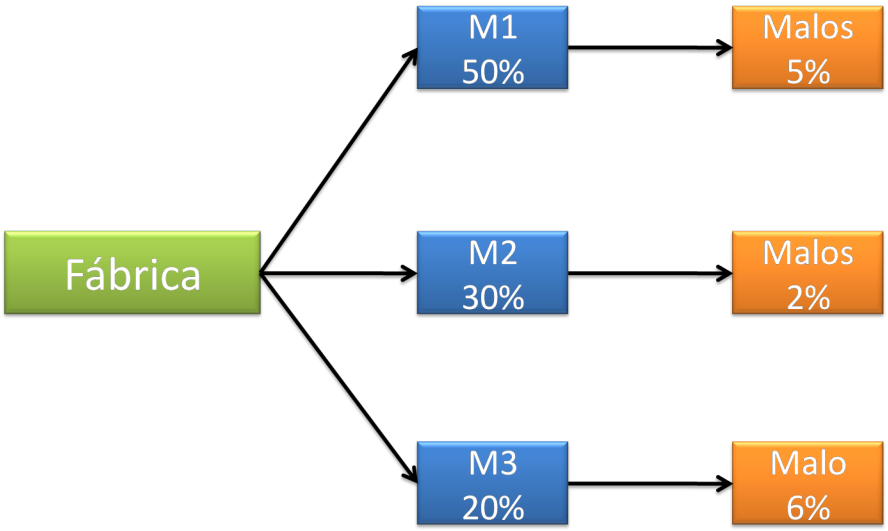


Figura 1.1: *Plano del proceso industrial en la fábrica de bolígrafos*

Una pregunta natural que surge es acerca de la probabilidad de selección de un artículo defectuoso y para responder a esta pregunta con «rigurosidad de probabilísta» es necesario enfocar nuestra atención en los tópicos básicos que dejamos atrás. En primer lugar el experimento en cuestión es la selección de un bolígrafo. Para este experimento, una terna $(\Omega, \mathfrak{F}, P)$ [1], llamada comúnmente espacio de medida o espacio de probabilidad, está dada por

1. El espacio muestral: $\Omega = \{\text{defectuoso}, \text{No defectouso}\}$
2. La σ-álgebra: $\mathfrak{F} = \{\Omega, \phi, \{\text{Defectuoso}\}, \{\text{No Defectuoso}\}\}$

[1] Ω denota el conjunto de todos lo posibles resultados del experimento, $\mathfrak{F}$ denota una σ-álgebra y P hace referencia ana medida de probabilidad propiamente definida.

3. La función de probabilidad:

$$\begin{aligned} p : \mathfrak{F} &\longrightarrow [0,1] \\ \Omega &\longrightarrow 1 \\ \phi &\longrightarrow 0 \\ \{Defectuoso\} &\longrightarrow P(D) \\ \{Defectuoso\} &\longrightarrow 1 - P(D) \end{aligned}$$

en donde, acudiendo al teorema de probabilidad total, se define

$$p(D) = p(D \mid M1)P(M1) + p(D \mid M2)P(M2) + p(D \mid M3)P(M3)$$

Sin embargo, también es posible plantearse otro tipo de preguntas que sirven para calibrar el proceso de producción de artículos defectuosos. Por ejemplo, cabe preguntarse acerca de la probabilidad de que habiendo seleccionado un artículo defectuoso, éste provenga de la primera máquina[2]. En esta ocasión, el experimento ha cambiado y ahora se trata de seleccionar un artículo defectuoso y para responder a tal cuestionamiento, se debe establecer rigurosamente el espacio de probabilidad que puede estar dado por

1. El espacio muestral: $\Omega = \{M1, M2, M3\}$
2. La σ-álgebra: $\mathfrak{F}^+ = \{\Omega, \phi, \{M1\}, \{M2, M3\}\}$
3. La función de probabilidad:

$$\begin{aligned} p : \mathfrak{F}^+ &\longrightarrow [0,1] \\ \Omega &\longrightarrow 1 \\ \phi &\longrightarrow 0 \\ \{M1\} &\longrightarrow p(M1 \mid D) \\ \{M2, M3\} &\longrightarrow 1 - p(M1 \mid D) \end{aligned}$$

en donde, acudiendo a la definición de probabilidad condicional, se define

$$p(M1 \mid D) = \frac{p(D \mid M1)P(M1)}{p(D \mid M1)P(M1) + p(D \mid M2)P(M2) + p(D \mid M3)P(M3)}$$

La anterior función de probabilidad se conoce con el nombre de regla de probabilidad de Bayes y, aparte de ser el baluarte de la mayoría de investigaciones estadísticas que se plantean hoy en día, ha sido la piedra de tropiezo de muchos investigadores radicales que trataron de estigmatizar este enfoque tildando a sus seguidores de mediocres matemáticos y pobres probabilistas afirmando que la regla de probabilidad de Bayes es sólo un artilugio diseñado para divertirse en el tablero.

Pues bien, la interpretación de la regla de bayes se puede realizar en el sentido de actualización de la estructura probabilística que gobierna el experimento. Y esta actualización tiene mucho sentido práctico cuando se cae en la cuenta de que la vida real está llena de calibradores y que las situaciones generadas son consecuencia de algún cambio estructural. De esta forma, el conocimiento de la probabilidad de que el artículo sea producido por la primera máquina se actualiza al conocer que este artículo particular es defectuoso y de esta manera calibra la estructura aleatoria que existe detrás del contexto de la fábrica de bolígrafos. Aparte de servir para resolver problemas como el anteriormente mencionado, la regla de bayes ha marcado el comienzo de un nuevo enfoque de análisis de datos, no solamente porque hace explícitas las relaciones causales entre los procesos aleatorios, sino también porque facilita la inferencia estadística y la interpretación de los resultados.

[2]Por supuesto que la pregunta también es válida al indagar por la probabilidad de que habiendo seleccionado un artículo defectuoso, éste provenga de la segunda o tercera máquina.

En el campo de la medicina, también se ha visto un gran número de la aplicación del teorema de Bayes. A continuación se enuncia uno de ellos:

Ejemplo 1.3.2. El Grupo de Trabajo de Servicios Preventivos de los Estados Unidos (USPSTF por sus siglas en inglés) hizo unas nuevas y controversiales recomendaciones sobre la detección del cáncer de mama (ver página $http://www.uspreventiveservicestaskforce.org/uspstf/uspsbrca.htm$), dentro de los cuales, no recomienda el examen de la mamografía en mujeres entre 40 y 49 años de edad, afirmando que la práctica bienal de este examen debe ser una decisión individual según el contexto particular de la paciente, mientras que por muchos años, se han dicho a las mujeres que se debe realizar la mamografía una vez cumplidos los 40 años. Por otro lado, USPSTF sí recomienda tal práctica de forma bienal en grupos de mujeres de entre 50 y 74 años de edad, puesto que USPSTF no encontró suficiente evidencia de beneficio o daño adicional en realizar este examen en mujeres mayores que los 74 años. Otra recomendación que hizo USPSTF es no realizar auto exámanes de senos, contrario a las recomendaciones y consejos que da la mayoría de los profesionales y organizaciones de la salud, incluyendo la *Amerian Cancer Society* (ver $http://www.cancer.org/acs/groups/cid/documents/webcontent/003164-pdf.pdf$).

El autor del blog, después de algunas averiguaciones, encontró que

- Los expertos estiman que un 12.3 % de las mujeres desarrollan formas invasivas del cáncer de mama durante la vida.
- La probabilidad de que una mujer desarrolle el cáncer de mama entre los 40 y los 49 años de edad es 1 en 69, y esta probabilidad aumenta a medida que envejezca, de tal forma que llega a ser de 1 en 38 en mujeres de entre 50 y 59 años.
- El cáncer de mama es más difícil de detectar en mujeres jóvenes puesto que el tejido mamario es más denso y fibroso. Los expertos estiman que la tasa de un falso positivo es de 97.8 por cada 1000 mujeres de 40 y 49 años, y esta tasa disminuye a 86.6 por cada 1000 mujeres entre 50 y 59 años.
- La tasa de un falso negativo es de 1 por cada 1000 mujeres de 40 y 49 años, y es de 1.1 por cada 1000 mujeres entre 50 y 59 años.

Resumiendo las anteriores afirmaciones, tenemos las siguientes probabilidades

Probabilidad	Edad	
	40 - 49 años	50 - 59 años
p(Cáncer)	1/69=0.01449	1/38=0.02632
p(No cáncer)	68/69=0.9855	37/38=0.97368
p(Positivo| No cáncer)	0.0978	0.0866
p(Negativo| No cáncer)	0.9022	0.9134
p(Negativo| Cáncer)	0.001	0.0011
p(Positivo| Cáncer)	0.999	0.9989

Utilizando la regla de Bayes, se puede calcular las siguientes probabilidades para mujeres de 40 y 49 años:

$$\begin{aligned} P(\text{Cáncer}|\text{Positivo}) &= \frac{P(\text{Positivo}|\text{Cáncer})P(\text{Cáncer})}{P(\text{Positivo}|\text{Cáncer})P(\text{Cáncer}) + P(\text{Positivo}|\text{No cáncer})P(\text{No cáncer})} \\ &= \frac{0.999 * 0.01449}{0.999 * 0.01449 + 0.0978 * 0.9855} \\ &= 0.1305 \end{aligned}$$

$$
\begin{aligned}
P(\text{Cáncer}|\text{Negativo}) &= \frac{P(\text{Negativo}|\text{Cáncer})P(\text{Cáncer})}{P(\text{Negativo}|\text{Cáncer})P(\text{Cáncer}) + P(\text{Negativo}|\text{No cáncer})P(\text{No cáncer})} \\
&= \frac{0.001 * 0.01449}{0.001 * 0.01449 + 0.9022 * 0.9855} \\
&= 0.0000163
\end{aligned}
$$

Similarmente, se puede calcular estas dos probabilidades para las mujeres de 50 y 59 años. Los resulta-

Probabilidad	Edad	
	40 - 49 años	50 - 59 años
P(Cáncer|Positivo)	0.1305985	0.23769
P(Cáncer|Negativo)	0.0000163	0.0000326
P(No cáncer|Positivo)	0.8694223	0.7623123
P(No cáncer|Negativo)	0.9999837	0.9999674

dos de la anterior tabla muestran cómo se cambia la probabilidad de tener cancer condicionado en los resultados de la pruebe. Entre estos valores se puede ver que, con un resultado positivo en el examen, la probabilidad de tener efectivamente el cáncer es aproximadamente diez puntos porcentuales más bajo en mujeres de edad de 40 y 49 años, de donde se puede sustentar la recomendación de no efectuar este examen en mujeres de este rango de edad.

1.4 Inferencia bayesiana

El enfoque bayesiano, además de especificar un modelo para los datos observados $\mathbf{Y} = (y_1, \ldots, y_n)$ dado un vector de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, usualmente en forma de densidad condicional $p(\mathbf{Y} \mid \boldsymbol{\theta})$, supone que $\boldsymbol{\theta}$ es aleatorio y que tiene un densidad *previa* $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$, donde $\boldsymbol{\eta}$ es un vector de hiper-parámetros. De esta forma, la inferencia concerniente a $\boldsymbol{\theta}$ se basa en una densidad *posterior* $p(\boldsymbol{\theta} \mid \mathbf{Y})$.

En términos de estimación, inferencia y predicción, el enfoque Bayesiano supone dos momentos o etapas:

1. Antes de la recolección de las datos, en donde el investigador propone, basado en su conocimiento, experiencia o fuentes externas, una distribución de probabilidad *previa* para el parámetro de interés. Con esta distribución es posible calcular estimaciones puntuales y por intervalo con el fin de confirmar que la distribución propuesta se ajusta al problema de estudio. En esta etapa, basados en la distribución *previa*, también es posible hacer predicciones de cantidades observables.

2. Después de la recolección de los datos. Siguiendo el teorema de Bayes, el investigador actualiza su conocimiento acerca del comportamiento probabilístico del parámetro de interés mediante la distribución *posterior* de este. Con esta distribución es posible calcular estimaciones puntuales y por intervalo justo como en el enfoque frecuentista. En esta etapa, basados en la distribución *posterior*, también es posible hacer predicciones de cantidades observables y pruebas de hipótesis acerca de la adecuación del mejor modelo a los datos observados.

1.4.1 Inferencia *previa*

Con las anteriores expresiones es posible calcular la probabilidad *previa* de que $\boldsymbol{\theta}$ esté en una determinada región G como

$$Pr(\boldsymbol{\theta} \in G) = \int_G p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.1}$$

En esta primera etapa también es posible calcular, con fines confirmatorios (Carlin & Louis 1996), la estimación puntual para el vector $\boldsymbol{\theta}$ dada por alguna medida de tendencia central para la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. En particular, si se escoge la media, entonces

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.2}$$

También es posible calcular una región C de $100 \times (1-\alpha)\,\%$ de credibilidad[3] para $\boldsymbol{\theta}$ que en esta primera etapa es tal que

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C) = \int_C p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.3}$$

1.4.2 Inferencia *posterior*

Una vez recolectados los datos, se actualizan las cálculos descritos en la sección anterior. Podemos calcular la probabilidad *posterior* de que $\boldsymbol{\theta}$ esté en la región G dados los datos observados como

$$Pr(\boldsymbol{\theta} \in G \mid \mathbf{Y}) = \int_G p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{1.4.4}$$

También es posible calcular la estimación puntual para el vector $\boldsymbol{\theta}$ dados los datos observados. Ésta está dada por alguna medida de tendencia central para la distribución $p(\boldsymbol{\theta} \mid \mathbf{Y})$. En particular, si se escoge la media, entonces

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \mathbf{Y}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{1.4.5}$$

La región C de $100 \times (1-\alpha)\,\%$ de credibilidad es tal que

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C \mid \mathbf{Y}) = \int_C p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{1.4.6}$$

También la distribución posterior del parámetro $\boldsymbol{\theta}$ es útil para el procedimiento de juzgamiento de hipótesis en el ámbito del análisis bayesiano. Esto se lleva a cabo por medio del factor de Bayes que se presentará más adelante.

1.4.3 Inferencia predictiva

En términos de inferencia predictiva existen dos etapas que cubren las «actuales» suposiciones acerca del vector de parámetros $\boldsymbol{\theta}$. En una primera etapa - antes de la observación de los datos - la suposición «actual» de $\boldsymbol{\theta}$ está dada por la densidad *previa* $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. En estos términos, utilizando el Resultado 1.3.1, la distribución predictiva *previa* de $\mathbf{Y}$ está dada por

$$p(\mathbf{y}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{1.4.7}$$

[3] La interpretación de las regiones de credibilidad bayesianas difiere de la interpretación de las regiones de confianza frecuentistas. La primera se refiere a la probabilidad de que el verdadero valor de $\boldsymbol{\theta}$ esté en la región. La segunda se refiere a la región de la distribución muestral para $\boldsymbol{\theta}$ tal que, dados los datos observados, se podría esperar que el $100 \times \alpha\,\%$ de las futuras estimaciones de $\boldsymbol{\theta}$ no pertenecieran a dicha región.

La segunda etapa - después de la recolección de los datos - actualiza las suposiciones acerca de $\boldsymbol{\theta}$ puesto que ahora éste sigue una distribución *posterior* dada por (1.3.1). Por lo tanto, la distribución predictiva *posterior* de $\mathbf{Y}$ está dada por

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \mathbf{Y}) &= \int p(\tilde{\mathbf{y}}, \boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \end{aligned} \tag{1.4.8}$$

donde $p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta})$ es la distribución de los datos evaluada en los nuevos valores $\tilde{\mathbf{y}}$. La segunda línea de la anterior igualdad se obtiene utilizando el resultado 1.2.1 y la última línea se obtiene del resultado 1.2.2 de la independencia condicional.

1.5 Información *previa*

La escogencia de una distribución previa es muy importante en el análisis bayesiano, puesto que ésta afecta directamente en la distribución posterior, tal como lo ilustra el teorema de Bayes. En primer lugar, la distribución previa debe describir adecuadamente los conocimientos previos sobre los parámetros objetivos de estimación. Por ejemplo, si se cree que un parámetro toma valores cercanos a 10, entonces la distribución escogida para representarla también debe tomar valores cercanos a 10, por ejemplo, una distribución normal centrada en ese valor. Por otro lado, dado que en la literatura existe un gran número de distribuciones, algunas muy similares entre ellas, a la hora de escoger una distribución previa también debe tener en cuenta las implicaciones a la hora de efectuar cálculos de la estimación puntual o de intervalo de crediblidad, procurando en la mayoría de casos, obtener una distribución posterior fácil de manejar. A continuación exponemos algunos aspectos generales relacionados con esta distribución previa.

1.5.1 Distribuciones conjugadas

Como se verá en los capítulos siguientes, muchos problemas de inferencia bayesiana comparten la agradable cualidad de que la forma funcional de la distribución *previa* para el parámetro de interés resulta ser la misma de la distribución *posterior*. Por ejemplo:

- Cuando se tiene una muestra aleatoria de variables con distribución Bernoulli de parámetro θ, es factible pensar que una distribución *previa* apropiada para este parámetro es la distribución Beta; bajo este escenario, la distribución *posterior* también resulta ser Beta.
- En el caso en que se quiera modelar el parámetro θ concerniente a una variable aleatoria con distribución Poisson, es posible asignar como candidata para distribución *previa* a la distribución Gamma; en este caso la distribución *posterior* también resulta ser Gamma.

Las distribuciones conjugadas son deseadas en el análisis bayesiano pues en primer lugar, la distribución posterior del parámetro θ es considerada como la actualización del conocimiento acerca de este después de la recolección de los datos, entonces al tener la misma forma funcional que la distribución previa, puede ser comparada a ésta y así ver claramente cómo es la influencia de los datos observados sobre la creencia acerca de θ; en segundo lugar, el hecho de que la distribución posterior sea de la misma forma funcional que la previa permite que la actualización de información se pueda llevar a cabo sistemáticamente, pues cada vez que se observan nuevos datos, la anterior distribución posterior puede ser tomada como la distribución previa y así producir una nueva distribución posterior.

A continuación exponemos la definición rigurosa de las distribuciones conjungadas y algunos tópicos relacionados.

Definición 1.5.1. *Sea $\mathcal{F} = \{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$ una familia de distribuciones de probabilidad. Una familia de distribuciones $\mathcal{P}$ se dice conjugada con respecto a $\mathcal{F}$ si para toda distribución* previa *$p(\boldsymbol{\theta}) \in \mathcal{P}$ y para toda distribución de muestreo o verosimilitud de las observaciones $p(\mathbf{Y} \mid \boldsymbol{\theta})$, $p(\boldsymbol{\theta} \mid \mathbf{Y})$ también pertenece a la familia $\mathcal{P}$.*

Esta definición es en la mayoría de los casos prácticos muy útil. Sin embargo, Migon & Gamerman (1999) describe los siguientes dos casos en donde esta definición es completamente inútil:

1. (Caso amplio) Sea $\mathcal{P} = \{$*Todas las distribuciones de probabilidad*$\}$ y $\mathcal{F}$ cualquier familia de distribuciones de probabilidad. Entonces $\mathcal{P}$ es conjugada con respecto a $\mathcal{F}$ puesto que toda posible distribución *posterior* será un miembro de $\mathcal{P}$.

2. (Caso restringido) Sea $\mathcal{P} = \{p \mid p(\theta = \theta_0) = 1\}$, esto es, $\mathcal{P}$ corresponde a todas las distribuciones concentradas en un punto. Sea $\mathcal{F}$ cualquier familia de distribuciones de probabilidad. De esta manera, la distribución *posterior* de θ estará dada por

$$p(\theta \mid Y) \propto p(Y \mid \theta)p(\theta) = \begin{cases} p(Y \mid \theta) \times 1 & \text{si } \theta = \theta_0 \\ p(Y \mid \theta) \times 0 & \text{si } \theta \neq \theta_0 \end{cases}$$
$$= \begin{cases} p(Y \mid \theta) & \text{si } \theta = \theta_0 \\ 0 & \text{si } \theta \neq \theta_0 \end{cases}$$

   De lo anterior y dado que $\int p(\theta \mid Y)\ d\theta = 1$, entonces $p(Y \mid \theta) = 1$ si y sólo si $\theta = \theta_0$. Con el anterior razonamiento, se concluye que $\mathcal{P}$ es conjugada con respecto a $\mathcal{F}$.

Por lo tanto, se deben buscar distribuciones *previa* que sean conjugadas de una forma tan amplia que permita proponer una distribución *previa* adecuada, pero al mismo tiempo tan restringida para que la definición de conjugada tenga sentido práctico. Ahora introducimos una familia de distribuciones muy importante para el desarrollo de la teoría estadística, tanto en el ámbito bayesiano como en el clásico.

Familia exponencial

Dependiendo de la naturaleza del parámetro θ, la familia exponencial puede ser uniparamétrica o multiparamétrica. En el primer caso, una distribución de probabilidad pertenece a la familia exponencial uniparamétrica si se puede escribir de la forma

$$p(Y \mid \theta) = \exp\{d(\theta)T(y) - c(\theta)\}h(y) \tag{1.5.1}$$

donde $T(y)$ y $h(y)$ son funciones que dependen de y únicamente, y $d(\theta)$ y $c(\theta)$ son funciones que depende de θ únicamente. Análogamente, una distribución de probabilidad pertenece a la familia exponencial multi-paramétrica si se puede escribir de la forma

$$p(Y \mid \boldsymbol{\theta}) = \exp\{\mathbf{d}(\boldsymbol{\theta})'\mathbf{T}(y) - c(\boldsymbol{\theta})\}h(y) \tag{1.5.2}$$

donde $\mathbf{T}(y)$ y $\mathbf{d}(\boldsymbol{\theta})$ son funciones vectoriales, $h(y)$ y $c(\boldsymbol{\theta})$ son funciones reales.

La ventaja de la familia exponencial radica en que es una familia relativamente restringuida de distribuciones y a la vez conserva la propiedad de ser distribuciones conjugadas, tal como muestra el siguiente resultado:

Resultado 1.5.1. *Sea Y una variable aleatoria con función de densidad perteneciente a la familia exponencial uniparamétrica, entonces la familia exponencial uniparamétrica es conjugada con respecto a sí misma.*

Prueba. Observando la expresión (1.5.1), se debe encontrar una distribución *previa* en la familia exponencial uniparamétrica, tal que la distribución *posterior*, resultante del producto de la distribución *previa* con la verosimilitud, sea también miembro de la familia exponencial uniparamétrica. Con base en lo anterior, la distribución *previa*, parametrizada por el hiperparámetro α, debe ser una función exponencial de los términos $d(\theta)$ y $c(\theta)$ como lo afirma Jordan (2004). Esto es,

$$p(\theta \mid \alpha) \propto \exp\{w(\alpha)d(\theta) - \delta c(\theta)\}, \tag{1.5.3}$$

donde δ es una constante real (posiblemente dependiente de α). Por otro lado, para garantizar que $p(\theta \mid \alpha)$ sea una auténtica función de densidad se normaliza de la siguiente manera

$$p(\theta \mid \alpha) = \frac{1}{k(\alpha, \delta)} \exp\{w(\alpha)d(\theta) - \delta c(\theta)\}, \tag{1.5.4}$$

con

$$k(\alpha, \delta) = \int \exp\{w(\alpha)d(\theta) - \delta c(\theta)\}\ d\theta.$$

De esta manera, no es difícil comprobar que la definición de distribución *previa*, parametrizada por el hiper-parámetro α, pertenece a la familia exponencial, puesto que

$$p(\theta \mid \alpha) = \exp\{\underbrace{w(\alpha)}_{d(\alpha)}\underbrace{d(\theta)}_{T(\theta)} - \underbrace{\ln k(\alpha, \delta)}_{c(\alpha)}\}\underbrace{\exp\{-\delta c(\theta)\}}_{h(\theta)}. \tag{1.5.5}$$

Por otro lado, del teorema de Bayes se tiene que

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta)p(\theta \mid \alpha) \\ &= \exp\{w(\alpha)d(\theta) + d(\theta)T(y) - c(\theta) - \ln k(\alpha, \delta)\}\exp\{-\delta c(\theta)\}h(y) \\ &= \exp\{\underbrace{[\alpha + T(y)]}_{d(y)}\underbrace{d(\theta)}_{T(\theta)} - \underbrace{[\ln k(\alpha, \delta) - \ln h(y)]}_{c(y)}\}\underbrace{\exp\{-(\delta + 1)c(\theta)\}}_{h(\theta)} \\ &\propto \exp\{[w(\alpha) + T(y)]d(\theta)\}\exp\{-(\delta + 1)c(\theta)\}. \end{aligned}$$

Por lo tanto, la distribución *posterior* resultante también pertenece a la familia exponencial uniparamétrica. ■

La extensión del anterior resultado para el caso cuando tenemos una muestra aleatoria de observaciones es sencilla, tal como se expone a continuación:

Resultado 1.5.2. *Sean $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ una muestra aleatoria de variables distribuidas con función de densidad común perteneciente a la familia exponencial uniparamétrica, cuya función de densidad conjunta $p(\mathbf{Y} \mid \theta)$ también pertenece a la familia exponencial uniparamétrica. Bajo las anteriores condiciones la familia exponencial uniparamétrica es conjugada con respecto a sí misma.*

Prueba. La demostración es inmediata utilizando el resultado anterior y notando que la forma funcional de la densidad conjunta para $\mathbf{Y}$ es

$$p(\mathbf{Y} \mid \theta) = \exp\left\{d(\theta)\sum_{i=1}^{n} T(y_i) - nc(\theta)\right\}\prod_{i=1}^{n} h(y_i) \tag{1.5.6}$$

la cual hace parte de la familia exponencial. ■

Otra extensión del resultado 1.5.1 corresponde al caso cuando la distribución de la observación está reparametrizado por un vector de parámetros $\boldsymbol{\theta}$. A continuación se expone el resultado y la prueba correspondiente.

Resultado 1.5.3. *Sean Y una variable aleatoria con función de densidad perteneciente a la familia exponencial multiparamétrica. Sea $\boldsymbol{\theta}$ el parámetro de interés con distribución* previa *parametrizada por un vector de hiperparámetros $\boldsymbol{\eta}$ y perteneciente a la familia exponencial multiparamétrica. Entonces la familia exponencial multiparamétrica es conjugada con respecto a sí misma.*

Prueba. En primer lugar, la distribución de probabilidad de Y perteneciente a la familia exponencial multiparamétrica está dada por (1.5.2). Siguiendo el mismo razonamiento de la demostración del Resultado 1.5.1, la distribución *previa* del parámetro de interés debe estar definida de la siguiente manera

$$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \exp\left\{ \underbrace{w(\boldsymbol{\eta})'}_{\mathbf{d}(\boldsymbol{\eta})} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\boldsymbol{\theta})} - \underbrace{\ln k(\boldsymbol{\eta}, \delta)}_{c(\boldsymbol{\eta})} \right\} \underbrace{\exp\{-\delta c(\boldsymbol{\theta})\}}_{h(\boldsymbol{\theta})}, \tag{1.5.7}$$

con

$$k(\boldsymbol{\eta}, \delta) = \int \exp\{w(\boldsymbol{\eta})'\mathbf{d}(\boldsymbol{\theta}) - \delta c(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}.$$

Utilizando el teorema de Bayes, se tiene que, la distribución *posterior* del parámetro θ es

$$\begin{aligned} p(\boldsymbol{\theta} \mid Y) &\propto p(Y \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \\ &= \exp\{\mathbf{T}(y)'\mathbf{d}(\boldsymbol{\theta}) - c(\boldsymbol{\theta}) + w(\boldsymbol{\eta})'\mathbf{d}(\boldsymbol{\theta}) - \delta c(\boldsymbol{\theta}) - \ln k(\boldsymbol{\eta}, \delta) + \ln h(y)\} \\ &= \exp\left\{ \underbrace{(w(\boldsymbol{\eta}) + \mathbf{T}(y))'}_{\mathbf{d}(y)} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\theta)} - \underbrace{[\ln k(\boldsymbol{\eta}, \delta) - \ln h(y)]}_{c(y)} \right\} \underbrace{\exp\{-(\delta + 1)c(\boldsymbol{\theta})\}}_{h(\boldsymbol{\theta})} \end{aligned}$$

La anterior expresión también hace parte de la familia exponencial biparamétrica y con esto se concluye la demostración ■

Nótese que el anterior resultado también cobija situaciones donde la verosimilitud sea perteneciente a la familia exponencial uniparamétrica. Más aún, a cualquier familia exponencial multiparamétrica de orden menor o igual al orden de la distribución *previa*.

Resultado 1.5.4. *Sean $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ una muestra aleatoria con función de densidad conjunta o verosimilitud dada (1.4.4). Bajo este escenario la familia exponencial multi-paramétrica es conjugada con respecto a sí misma.*

Prueba. La demostración sigue los mismos lineamentos que la demostración del resultado anterior

concluyendo que la distribución *posterior* de $\boldsymbol{\theta}$ está dada por

$$\begin{aligned}
&p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \\
&= \exp\left\{ \sum_{i=1}^{n} \mathbf{T}(y_i)' \mathbf{d}(\boldsymbol{\theta}) - n c(\boldsymbol{\theta}) + \boldsymbol{\eta}' \mathbf{d}(\boldsymbol{\theta}) - \delta c(\boldsymbol{\theta}) - \ln k(\boldsymbol{\eta}, \delta) + \sum_{i=1}^{n} \ln h(y_i) \right\} \\
&= \exp\left\{ \underbrace{\left(\boldsymbol{\eta} + \sum_{i=1}^{n} \mathbf{T}(y_i) \right)'}_{\mathbf{d}(\mathbf{y})} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\theta)} - \underbrace{\left[\ln k(\boldsymbol{\eta}, \delta) - \sum_{i=1}^{n} \ln h(y_i) \right]}_{c(\mathbf{y})} \right\} \\
&\times \underbrace{\exp\left\{ -(\delta + n) c(\boldsymbol{\theta}) \right\}}_{h(\boldsymbol{\theta})}
\end{aligned}$$

La anterior expresión también hace parte de la familia exponencial. ■

Ahora, estudiamos las expresiones relacionadas con la distribución predictiva de nuevas observaciones dentro del contexto de la familia exponencial:

Resultado 1.5.5. *Sea Y una variable aleatoria con función de densidad perteneciente a la familia exponencial, dada por (1.5.1). Sea θ el parámetro de interés con distribución* previa *en la familia exponencial biparamétrica. La distribución predictiva* previa *de Y está dada por*

$$p(Y) = \frac{k(\alpha + T(y), \delta + 1)}{k(\alpha, \delta)} h(y) \tag{1.5.8}$$

donde

$$k(a, b) = \int \exp\{ w(a) d(\theta) - b c(\theta) \} \, d\theta$$

Prueba.

$$\begin{aligned}
p(Y) &= \int p(\theta) p(Y \mid \theta) \, d\theta \\
&= \int \exp\{ w(\alpha) d(\theta) - \ln k(\alpha, \delta) - \delta c(\theta) \} \exp\{ d(\theta) T(y) - c(\theta) \} h(y) d\theta \\
&= \frac{h(y)}{k(\alpha, \delta)} \int \exp\{ [w(\alpha) + T(y)] d(\theta) - (\delta + 1) c(\theta) \} d\theta \\
&= \frac{k(\alpha + T(y), \delta + 1) h(y)}{k(\alpha, \delta)}
\end{aligned}$$

donde

$$k(\alpha, \delta) = \int \exp\{ w(\alpha) d(\theta) - \delta c(\theta) \} \, d\theta$$

y

$$k(\alpha + T(y), \delta + 1) = \int \exp\{ [w(\alpha) + T(y)] d(\theta) - (\delta + 1) c(\theta) \} \, d\theta.$$

■

La extensión al caso de contar con una muestra aleatoria de observaciones se encuentra a continuación:

Resultado 1.5.6. *Sea* $\mathbf{Y} = \{Y_1 \ldots, Y_n\}$ *una muestra aleatoria con función de densidad conjunta perteneciente a la familia exponencial, dada por (1.4.4). Sea* θ *el parámetro de interés con distribución* previa *dada por (1.4.5). La distribución predictiva* previa *de* $\mathbf{Y}$ *está dada por*

$$p(\mathbf{Y}) = \frac{k(\alpha + T(\mathbf{y}), \delta + n)}{k(\alpha, \beta)} h(\mathbf{y}) \tag{1.5.9}$$

donde k *se define tal como en el resultado anterior.*

Prueba. La prueba se tiene de inmediato siguiendo los lineamentos de la demostración del anterior resultado. ■

Resultado 1.5.7. *En términos de la distribución predictiva* posterior*, se tiene que para una sola observación* $\tilde{y}$*, ésta está dada por*

$$p(\tilde{y} \mid Y) = \frac{k(\alpha + T(y) + T(\tilde{y}), \delta + 2)}{k(\alpha + T(y), \delta + 1)} h(\tilde{y}) \tag{1.5.10}$$

y en el caso en donde se tiene una muestra aleatoria, entonces la distribución predictiva posterior *para una nueva muestra* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$ *de tamaño* n^* *está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{k(\alpha + T(\mathbf{y}) + T(\tilde{\mathbf{y}}), \delta + n + n^*)}{k(\alpha + T(\mathbf{y}), \delta + n)} h(\tilde{\mathbf{y}}) \tag{1.5.11}$$

Prueba. De la definición de distribución predictiva *posterior* dada por la expresión (1.4.8) se tiene que

$$\begin{aligned} p(\tilde{y} \mid Y) &= \int p(\tilde{y} \mid \theta) p(\theta \mid y) \, d\theta \\ &= \int \exp\{d(\theta)T(\tilde{y}) - c(\theta)\} h(\tilde{y}) \frac{\exp\{[w(\alpha) + T(y)]d(\theta) - (\delta + 1)c(\theta)\}}{k(\alpha + T(y), \delta + 1)} \, d\theta \\ &= \frac{h(\tilde{y})}{k(w(\alpha) + T(y), \delta + 1)} \int \exp\{[\alpha + T(y) + T(\tilde{y})]d(\theta) - (\delta + 2)c(\theta)\} \, d\theta \\ &= \frac{k(\alpha + T(y) + T(\tilde{y}), \delta + 2)}{k(\alpha + T(y), \delta + 1)} h(\tilde{y}), \end{aligned}$$

con

$$k(\alpha + T(y) + T(\tilde{y}), \delta + 2) = \int \exp\{[w(\alpha) + T(y) + T(\tilde{y})]d(\theta) - (\delta + 2)c(\theta)\} \, d\theta.$$

La demostración para la nueva muestra se lleva a cabo de manera análoga. ■

1.5.2 Distribuciones *previa* no informativas

Cuando no existe una base poblacional sobre el parámetro de interés o cuando existe total ignorancia de parte del investigador acerca del comportamiento de probabilístico del parámetro, es necesario definir distribuciones *previa* que sean no informativas. Es decir, definir distribuciones *previa* que jueguen un papel mínimo en términos de influencia en la distribución *posterior*. Una característica de estas distribuciones es que su forma es vaga, plana o difusa, cumpliendo así el objetivo de no influenciar a la distribución *posterior*. Por tanto la pregunta de interés que surge en este instante es: ¿cómo seleccionar distribuciones *previa* no informativas[4] sobre el parámetro de interés?

[4]Existen muchas denominaciones para las distribuciones uniformes que no son informativas. Por ejemplo, Box Tiao proponen el nombre de distribuciones localmente uniformes para asegurar que cumplan con las condiciones de función de densidad de probabilidad en un rango particular del espacio paramétrico. Sin embargo, en este texto vamos a utilizar la expresión ≪no informativa≫ al referirse a este tipo de distribuciones a previa.

En los anteriores términos, la distribución uniforme define una distribución *previa* que cumple con las características de no información en la mayoría de escenarios. Específicamente en aquellos problemas en donde el parámetro de interés está limitado a un espacio de muestreo acotado. Por ejemplo, en la distribución Binomial, el parámetro de interés está limitado al espacio de muestreo $[0, 1]$. Sin embargo, no en todos los problemas encaja la distribución uniforme. Nótese, por ejemplo, que en el caso en que la distribución exponencial se acomode a los datos como candidata a verosimilitud, entonces el espacio de muestreo del parámetro de interés estaría dado por $(0, \infty)$ en cuyo caso la distribución uniforme no sería conveniente puesto que sería una distribución impropia en el espacio de muestreo del parámetro de interés. Es decir

$$\text{si } p(\theta) \propto kI_{\Theta}(\theta), \text{ entonces } \int_{\Theta} p(\theta)\ d(\theta) \longrightarrow \infty.$$

donde Θ denota espacio de muestreo del parámetro θ y I denota la función indicadora. Por otro lado, una característica importante que debe tener una distribución *previa* no informativa es que sea invariante en términos de transformaciones matemáticas. Es decir, si el parámetro de interés es θ con distribución *previa* no informativa dada por $p(\theta)$, y sea $\phi = h(\theta)$ una transformaición de θ por medio de la función h, entonces la distribución *previa* de ϕ también debería ser no informativa. Sin embargo, la teoría de probabilidad afirma que la distribución de probabilidad de una transformación está dada por

$$p(\phi) = p(\theta) \mid \frac{d\theta}{d\phi} \mid = p(\theta) \mid h'(\theta) \mid^{-1} \tag{1.5.12}$$

y claramente si la función h no es una función lineal, entonces los resultados encontrados por medio de este enfoque indicarían que la distribución *previa* $p(\phi)$ sería informativa contradiciendo los supuestos de $p(\theta)$. El siguiente ejemplo ilustra este planteamiento:

Ejemplo 1.5.1. Suponga que el parámetro de interés es θ y que está restringido a un espacio de muestreo dado por el intervalo $[0, 1]$. Si se supone completa ignorancia acerca del comportamiento del parámetro, entonces una buena opción, con respecto a la distribución *previa*, sería la distribución uniforme en el intervalo $[0, 1]$. Es decir, la distribución *previa* no informativa estaría dada por

$$p(\theta) = I_{[0,1]}(\theta)$$

Suponga ahora que existe una transformación del parámetro de interés dada por $\phi = h(\theta) = \ln(\theta)$. Por tanto, siguiendo (1.5.12) se tiene que la distribución de ϕ está dada por

$$p(\phi) = I_{(-\infty,0)}(\phi)e^{\phi}$$

la cual es informativa con respecto al parámetro ϕ. Sin embargo, es el mismo problema y existe una contradicción en términos de que para θ se desconoce todo, pero para una función ϕ existe evidencia de que el parámetro se comporta de cierta manera.

Para palear las anteriores diferencias, es necesario encontrar una distribución *previa* no informativa que sea invariante a transformaciones matemáticas. La distribución *previa* no informativa de Jeffreys, definida a continuación, cuenta con esta agradable propiedad.

Definición 1.5.2. *Si la verosimilitud de los datos está determinada por un único parámetro θ, la distribución* previa *no informativa de Jeffreys tiene distribución de probabilidad dada por*

$$p(\theta) \propto (I(\theta))^{1/2} \tag{1.5.13}$$

con $I(\theta)$ la información de Fisher definida como

$$\begin{aligned} I(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\log p(\mathbf{Y} \mid \theta)\right]^2\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{Y} \mid \theta)\right\} \end{aligned}$$

Si la verosimilitud de los datos está determinada por un vector de parámetros $\boldsymbol{\theta}$*, la distribución* previa *no informativa de Jeffreys tiene distribución de probabilidad dada por*

$$p(\theta) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2} \tag{1.5.14}$$

donde $\mathbf{I}$ *es la matriz de información de Fisher, cuyo elemento en la fila* i *y columna* j *está definida como*

$$\begin{aligned}\mathbf{I}_{[ij]}(\boldsymbol{\theta}) &= E\left\{\left[\frac{\partial}{\partial\theta_i}\log p(\mathbf{Y}\mid\theta)\right]\left[\frac{\partial}{\partial\theta_j}\log p(\mathbf{Y}\mid\boldsymbol{\theta})\right]\right\}\\ &= -E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\mathbf{Y}\mid\boldsymbol{\theta})\right\}\end{aligned}$$

donde θ_i *y* θ_j *son los elementos* i *y* j *del vector* $\boldsymbol{\theta}$*.*

Nótese que si la verosimilitud de las observaciones pertenecen a la familia de distribuciones exponencial, entonces la distribución previa de Jeffreys no es difícil de calcular. Por otro lado nótese que la distribución previa no informativa de Jeffreys depende, de cierta manera, del mecanismo probabilístico que rige a los datos. Lo anterior hace que ciertos críticos de la estadística bayesiana critiquen este enfoque puesto que se supone que la formulación de la distribución a previa es independiente de los datos observados.

A continuación se evidencia la propiedad de esta distribución previa de seguir siendo no informativa con diferentes parametrizaciones.

Resultado 1.5.8. *La distribución* previa *no informativa de Jeffreys es invariante a transformaciones uno a uno. Es decir, si* $\phi = h(\theta)$*, entonces* $p(\phi) \propto (I(\phi))^{1/2}$*.*

Prueba. En primer lugar nótese que

$$I(\theta) = \mathbf{J}(\phi) \mid \frac{\partial\phi}{\partial\theta} \mid^2$$

puesto que al utilizar la regla de la cadena del cálculo matemático se tiene que

$$\begin{aligned}\mathbf{J}(\phi) = -E\left[\frac{\partial^2\log p(\mathbf{Y}\mid\phi)}{\partial\phi^2}\right] &= -E\left[\frac{\partial}{\partial\phi}\left(\frac{\partial\log p(\mathbf{Y}\mid\phi)}{\partial\phi}\right)\right]\\ &= -E\left[\frac{\partial}{\partial\theta}\left(\frac{\partial\log p(\mathbf{Y}\mid\phi)}{\partial\phi}\right)\mid\frac{\partial\theta}{\partial\phi}\mid\right]\\ &= -E\left[\frac{\partial^2\log p(\mathbf{Y}\mid\phi)}{d\theta^2}\mid\frac{\partial\theta}{\partial\phi}\mid^2\right]\\ &= -E\left[\frac{\partial^2\log p(\mathbf{Y}\mid\theta=h^{-1}(\phi))}{d\theta^2}\mid\frac{\partial\theta}{\partial\phi}\mid^2\right]\\ &= I(\theta)\mid\frac{\partial\theta}{\partial\phi}\mid^2\end{aligned}$$

Ahora, de la definición de función de distribución para una función y utilizando (1.4.11), se tiene que

$$p(\phi) = p(\theta)\mid\frac{\partial\theta}{\partial\phi}\mid\propto(I(\theta))^{1/2}\mid\frac{\partial\theta}{\partial\phi}\mid\propto I(\phi)^{1/2}\mid\frac{\partial\phi}{\partial\theta}\mid\mid\frac{d\theta}{\partial\phi}\mid= I(\phi)^{1/2}$$

■

En Box & Tiao (1992, p. 59) citan una Tabla de resumen en donde se encuentran distribuciones a previa no informativas para las distribuciones probabilísticas más comunes. A continuación se exponen algunos ejemplos que utilizan este enfoque.

Ejemplo 1.5.2. Si Y es una variable aleatoria con distribución Binomial, entonces el espacio de muestreo del parámetro de interés será el intervalo $[0, 1]$; sería conveniente utilizar la función de distribución uniforme sobre este intervalo como distribución *previa* no informativa. Con el enfoque de Jeffreys se llega a este mismo resultado puesto que: la información de Fisher para la distribución binomial es $J(\theta) = n/\theta(1-\theta)$ dado que

$$\log p(Y \mid \theta) = \log \binom{n}{y} + y \log(\theta) + (n-y)\log(1-\theta)$$

y

$$\frac{\partial^2 \log p(Y \mid \theta)}{\partial \theta^2} = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}$$

Por lo tanto al calcular la esperanza, y por consiguiente la información de Fisher, se tiene que

$$I(\theta) = -E\left[\frac{d^2 \log p(Y \mid \theta)}{d\theta^2}\right] = \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Es decir, la distribución *previa* no informativa para el parámetro de interés θ es proporcional a $\theta^{-1/2}(1-\theta)^{-1/2}$, la cual comparte la misma forma estructural de una distribución $Beta(1/2, 1/2)$ que a su vez es idéntica a la distribución uniforme. En términos de la distribución *posterior* para el parámetro de interés, se tiene que

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta)p(\theta) \\ &\propto \theta^y (1-\theta)^{n-y} \theta^{-1/2} (1-\theta)^{-1/2} \\ &= \theta^{y+1/2-1} (1-\theta)^{n-y+1/2-1} \end{aligned}$$

Por tanto, la distribución de $\theta \mid Y$ es $Beta(y + 1/2, n - y + 1/2)$. Por construcción, esta distribución no está alterada ni influenciada por la distribución *previa* pues la misma es no informativa.

Ejemplo 1.5.3. Si $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es una muestra aleatoria de variables con distribución de Poisson, entonces el espacio de muestreo del parámetro de interés será el intervalo $(0, \infty)$; por tanto utilizar la distribución uniforme como distribución *previa* no informativa no es conveniente. Ahora, la información de Fisher para la distribución conjunta es $I(\theta) = n/\theta$ puesto que

$$\log p(\mathbf{Y} \mid \theta) = -n\theta + \log(\theta) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!)$$

y

$$\frac{\partial^2 \log p(\mathbf{Y} \mid \theta)}{\partial \theta^2} = -\frac{\sum_{i=1}^{n} y_i}{\theta^2}$$

Por lo tanto al calcular la esperanza, y por consiguiente la información de Fisher, se tiene que

$$I(\theta) = -E\left[\frac{\partial^2 \log p(\mathbf{Y} \mid \theta)}{\partial \theta^2}\right] = \frac{\sum_{i=1}^{n} E(y_i)}{\theta^2} = \frac{n}{\theta}$$

Es decir, la distribución *previa* no informativa para el parámetro de interés es proporcional a $\theta^{-1/2}$. En términos de la distribución *posterior* para el parámetro de interés, se tiene que

$$p(\theta \mid Y) \propto p(Y \mid \theta)p(\theta) \propto e^{-n\theta} \theta^{\sum_{i=1}^{n} y_i} \theta^{-1/2} = e^{-n\theta} \theta^{\sum_{i=1}^{n} y_i - 1/2}$$

Por tanto, la distribución de $\theta \mid \mathbf{Y}$ es $Gamma(\sum_{i=1}^{n} y_i + 1/2, n)$. Por construcción, esta distribución no está alterada ni influenciada por la distribución *previa* pues la misma es no informativa.

Ejemplo 1.5.4. Suponga que $\mathbf{Y} = \{Y_1 \ldots, Y_n\}$ es una muestra aleatoria con distribución normal de parámetros $(\theta, \sigma^2)'$. Se puede verificar que la matriz de información de Fisher para el vector de parámetros está dada por

$$\begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{1.5.15}$$

cuyo determinante está dado por $\frac{n^2}{2\sigma^6}$. Por lo tanto, la distribución a previa no informativa de Jeffreys está dada por

$$p(\theta, \sigma^2) \propto 1/\sigma^3 \tag{1.5.16}$$

1.6 Pruebas de hipótesis

A excepción del juzgamiento de hipótesis, las inferencias que hacen los estadísticos bayesianos, acerca de poblaciones normales, son muy similares a las que los estadísticos de la tradición frecuentista, de Neyman y Pearson, hacen. Consideremos la siguiente situación. Un instrumento mide la posición de un objeto con un determinado error. Éste error está distribuido de manera uniforme en el intervalo (-1cm, 1cm). Supongamos que el instrumento midió la posición de un objeto en +0.9999cm del origen. Planteamos la siguiente hipótesis nula, H: La posición real del objeto es exactamente el origen. Imagine que planteamos este problema de inferencia estadística a los profesores López (frecuentista clásico) y Cepeda (acérrimo bayesiano). Razonamiento del frecuentista: Si la hipótesis nula es verdadera, ha ocurrido un evento con una probabilidad (a dos colas) de ocurrencia de 0.0001 o menos. Mediante un criterio razonable (nivel de significación), este es un evento muy raro y por lo tanto rechaza H. Razonamiento del bayesiano: El bayesiano ve las cosas desde un punto de vista diferente. Dada una observación, la verosimilitud asociada con la posición del objeto en el intervalo -0.0001 y +1.9999 es la misma, 0.5. Fuera de esos límites la verosimilitud es nula. Ahora, el origen está dentro de la región en donde la verosimilitud es máxima; por lo tanto sea cual sea la distribución a previa asociada al parámetro de posición, la distribución a posterior tomara el valor cero en cualquier lugar fuera del intervalo -0.0001 y +1.9999. Así, con la observación disponible, no hay evidencia para el rechazo de H. Bajo esta paradoja, Brewer (2002) sugiere que ambos estadísticos tienen razón, pero a la vez están equivocados. El frecuentista tiene razón en afirmar que, con la evidencia disponible, ha ocurrido un evento extraordinariamente extraño o que la hipótesis nula es falsa. El bayesiano tiene razón en argumentar que, en términos de la situación, no hay evidencia en contra de la hipótesis nula. Esta paradoja se presenta porque los bayesianos tienden a trabajar dentro de la situación que ellos creen que existe (o al menos creen que ellos creen que existe) y la lógica bayesiana se mueve en ese marco de referencia. Los bayesianos hacen las inferencias en términos de la verosimilitud de los eventos observados, mientras que los frecuentistas hacen inferencias en términos de eventos que ni siquiera han ocurrido. .

1.6.1 Factor de Bayes

El juzgamiento de hipótesis del enfoque frecuentista se puede efectuar en el ámbito Bayesiano por medio del *Factor de Bayes*. Suponiendo que existen dos modelos $M1$ y $M2$ candidatos para $\mathbf{Y}$, se define el *Factor de Bayes* en favor del modelo $M1$ como la razón de las densidades marginales de los datos para los dos modelos y es posible demostrar que es equivale a la siguiente expresión

$$FB = \frac{p(\mathbf{Y} \mid M1)}{p(\mathbf{Y} \mid M2)} = \frac{Pr(M1 \mid \mathbf{Y})/Pr(M2 \mid \mathbf{Y})}{Pr(M1)/Pr(M2)} \tag{1.6.1}$$

Para evaluar esta última expresión es necesario recurrir a la densidad previa y posterior del parámetro de interés, asumiendo que los modelos están parametrizados por éstos. Se puede ver que cuando los

modelos $M1$ y $M2$ tienen la misma distribución previa, entonces el factor de Bayes se reduce a la razón de densidad posterior de los dos modelos. Adicionalmente este factor sólo está definido cuando la integral de la densidad marginal de $\mathbf{Y}$ bajo cada modelo converge. En la expresión (1.6.1) se claro que valores grandes del factor muestra evidencias a favor del modelo $M1$, valores menores de 1 a favor del modelo $M2$, mientras que valores cercanos a 1 no muestra evidencias claras hacia ninguno de los dos modelos.

En Gelman, Carlin, Stern & Rubin (1995) presenta el siguiente ejemplo sencillo sobre la presencia o ausencia de la enfermedad hemofilia, una enfermedad genética especialmente grave las mujeres. Para una mujer quien tiene un hermano portador del gen, el parámetro θ describe la presencia o ausencia del gen en ella, y toma valores de 1 (presencia del gen) y 0 (ausencia del gen). La distribución previa del parámetro es $P(\theta = 1) = P(\theta = 0) = 0.5$. El objetivo es evaluar el sistema $M_1 : \ \theta = 1$ y $M_2 : \ \theta = 0$ con base en el hecho de que ella tiene dos hijos ambos no portadores del gen. De esta forma

$$FB = \frac{p(y_1 = 0,\ y_2 = 0|\theta = 1)}{p(y_1 = 0,\ y_2 = 0|\theta = 0)} = \frac{0.25}{1} = 0.25$$

De donde se evidencia mayor apoyo a la hipótesis $\theta = 0$.

1.6.2 Valor p Bayesiano

En la inferencia clásica, se define el valor p como la probabilidad de que la estadística de prueba tome valores más extremos a los observados, y se compara con el nivel de significancia, previamente establecidad, para tomar decisión acerca de una hipótesis nula. En el ámbito Bayesiano, el valor p se define como la probabilidad de que la estadística de prueba T calculado sobre los datos replicados y^{rep} sean más extremos al observado, y la probabilidad se toma sobre la distribución posterior del parámetro θ y la distribución predictiva posterior de y^{rep}. Específicamente, queda determinado por

$$p_B = \int \int_{T(y^{rep}) \geq T(y)} p(y^{rep}|\theta) p(\theta|y) dy^{rep} d\theta$$

A diferencia del valor p clásico donde solo valores pequeños muestran evidencia en contra de la hipótesis nula, un valor p Bayesiano extremo (menor a 0.01 o mayor a 0.99) sugiere que los valores observados difícilmente pueden ser replicadso si el modelo fuera verdadero.

1.7 Criterios de información

Los criterios de información constituyen una herramienta muy importante en el modelamiento estadístico, pues contribuye a la selección de modelos de manera simple. Existen una variedad de estos criterios, a continuación se describen los dos criterios más comunes en el análisis bayesiano.

Criterio DIC

El criterio de información de devianza (denotada por DIC por los iniciales en inglés) es una generalización del popular criterio AIC para los modelos jerárquicos, y se basa en el concepto de la devianza que se define como

$$D(y, \boldsymbol{\theta}) = -2 * \log(p(y|\boldsymbol{\theta})) \tag{1.7.1}$$

cuya media posterior es una medida usual del ajuste del modelo. Dempster (1974) sugirió graficar la distribución posteriori de la devianza para observar el ajuste del modelo a los datos. Una estimación

de esta media posterior se basa en simulación de M valores $\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^M$ de la distribución posterior de $\boldsymbol{\theta}$, y está dada por

$$\hat{E}_D = \frac{1}{M} \sum_{m=1}^{M} D(y, \boldsymbol{\theta}^m)$$

El DIC se define como

$$DIC = \hat{E}_D + p_D$$

Donde p_D es el número efectivo de parámetros. Nótese que en la anterior formulación, el DIC se puede descomponer en dos partes: la parte de la bondad de ajuste del modelo, medido a través de E_D, y la parte que mide la complejidad del modelo p_D. Otra formulación equivalente del DIC se obtiene teniendo en cuenta que

$$p_D = \hat{E}_D - \hat{D}$$

Donde $\hat{D} = -2 * \log(p(y|\hat{\boldsymbol{\theta}}))$ con $\hat{\boldsymbol{\theta}}$ denotando la media posterior de $\boldsymbol{\theta}$; es decir, $\hat{D}$ es la estimación de la devianza usando $\hat{\boldsymbol{\theta}}$, y p_D se puede ver como la media posterior de la devianza menos la devianza de las medias posterior (Spiegelhalter, Best, Carlin & VanderLinde 2002). De esta forma, el DIC también se puede escribir como

$$DIC = \hat{D} + 2p_D$$

Interpretación de DIC: El modelo con el menor DIC es considerado como el modelo que mejor predice un conjunto de datos con la misma estructura que los datos observados. Al respecto se deben tener en cuenta las siguientes consideraciones:

- El DIC puede ser negativo puesto que $p(y|\theta)$ puede tomar valores mayores a 1 asociado a una devianza pequeña.
- p_D, y por consiguiente DIC, no es invariante a parametrizaciones del modelo. Se sugiere en la práctica usar parametrizaciones que conducen a la normalidad en la distribución posterior.

Criterio AIC y BIC

El criterio de información de Akaike (AIC) fue formalmente presentado en Akaike (1974). Este criterio mide la pérdida de información al ajustar un modelo a un conjunto de datos; por esto, se buscan modelos que arrojen valores pequeños de AIC. Posteriormente Cavanaugh (1997) introdujo el factor de corrección para evitar que el AIC escoja modelos con demasiados parámetros en situaciones de tamaño de muestra pequeño. Por otro lado, el criterio de información bayesiano BIC, también conocido como el criterio de Schwarz (Schwarz 1978), también está formulado en términos de la función de verosimilitudel modelo y del número de parámetros. La expresión de estos criterios es como sigue:

$$\begin{aligned} AIC &= -2\log(p(y|\hat{\boldsymbol{\theta}})) + 2p \\ AIC_c &= AIC + \frac{2p^2 + 2p}{n - p - 1} \\ BIC &= -2\log(p(y|\hat{\boldsymbol{\theta}})) + p\log(n) \end{aligned}$$

Donde p es el número de parámetros en el modelo y n el número de datos observados. Cabe resaltar que en el criterio BIC hay una mayor penalización por el número excesivo de parámetros que en el criterio AIC, y en la práctica se prefieren los modelos con un BIC menor.

Nota: Se debe recalcar que los dos criterios tienen diferentes enfoques, el criterio BIC se enfoca en identificar el modelo verdadero, mientras que el criterio DIC enfoca en encontrar el modelo con mejor capacidad de predicción.

1.8 Acerca de la notación

Antes de empezar las próximas secciones, es necesario revisar la notación que se seguirá de ahora en adelante. Del teorema de Bayes resultan tres grandes definiciones que constituyen la base de la estadística Bayesiana y que a lo largo de este texto se mencionarán diferenciándolas por medio de la notación. El símbolo más importante de la estadística matemática es p, el cual indica que existe una distribución de probabilidad para los datos, para el vector de parámetros, condicional o no. De hecho todos las definiciones y resultados anteriores han estado supeditadas al uso de esta monótona notación. En el ámbito de la notación de investigación internacional es común diferenciar las distribuciones con el fin de hacer más ameno el estudio del enfoque Bayesiano. En este texto se seguirá esta distinción. Un ejemplo claro en donde p representa cuatro funciones distintas en una sola ecuación es el siguiente:

$$p(\theta \mid y) = p(y \mid \theta)\frac{p(\theta)}{p(y)}$$

Gelman, Carlin, Stern & Rubin (1995) explica por qué la notación simple, con el uso (a veces abuso) de la letra p es más rigurosa de lo que, a simple vista, pueda parecer y comenta que,

> En realidad no me gusta la notación que la mayoría de los estadísticos usan: f, para distribuciones de muestreo; π, para distribuciones a previa y L, para verosimilitudes. Este estilo de notación se desvía de lo que realmente es importante. La notación no debería depender del orden en que las distribuciones son especificadas. Todas ellas son distribuciones de probabilidad, eso es lo realmente importante.

Esto tiene sentido, aún más cuando se estudian las propiedades estadísticas de los estimadores desde el punto de vista de la teoría de la medida. Siendo así, el símbolo p se refiere a una notación para una medida de probabilidad, quizás inducida por un elemento aleatorio. De hecho, en la ecuación que determina la regla de Bayes, cada una de las p son medidas de probabilidad que no comparten el mismo espacio de medida (ni la misma σ-álgebra, ni el mimo espacio muestral).

De hecho, todo queda claro al realizar un diagrama que permita ver el espacio de salida y el espacio de llegada de los elementos aleatorios que inducen (si es el caso), cada una de las distribuciones de probabilidad. Por otra parte, Bob Carpenter, concluye que

> [Una vez resuelto el problema de identificación de los espacios] la notación estadística depende en gran manera del contexto y aunque la regla de Bayes no necesite de mucha explicación, es necesario conocerlo todo acerca del contexto para poder interpretar las funciones que la conforman... El problema se hace mucho más agudo para los estadísticos novatos, pero eso se resuelve con la práctica. Una vez que uno sabe lo que está haciendo, se vuelve obvia la referencia de la distribución p.

Por lo anterior, es natural que algunos de los textos clásicos de estadística matemática, parezcan olvidar el contexto de las diferentes medidas de probabilidad. En realidad no es que lo olviden, lo que pasa es que los autores no son novatos y asumen que el lector sigue la idea de la referencia de la p en cuestión. Sin embargo, y lo digo por mi y sólo por mí, sería mejor que no asumieran esa idea. De esta manera, el estudio de estos textos sería un poco menos denso.

Capítulo 2

Modelos uniparamétricos

Los modelos que están definidos en términos de un solo parámetro que pertenece al conjunto de los números reales se definen como modelos uniparamétricos. Este capítulo estudia modelos, discretos y continuos, que son comunes de implementar en la práctica. Dado que todos ellos son inducidos por familias de probabilidad conjugadas, entonces las estimaciones posteriores para los parámetros pueden hallarse sin necesidad de sofisticaciones computacionales. Es decir, con el uso de una simple calculadora de bolsillo, es posible realizar inferencia bayesiana propiamente dicha. Por lo tanto, en este capítulo, será menor el uso de software estadístico. Sin embargo, para cada modelo se incluye la sintaxis de `JAGS`, para un ejemplo práctico que permite la familiarización e interiorización del ambiente computacional de este software que será indispensable en el desarrollo de capítulos posteriores.

2.1 Modelo Bernoulli

Suponga que Y es una variable aleatoria con distribución Bernoulli, su distribución está dada por

$$p(Y \mid \theta) = \theta^y (1-\theta)^{1-y} I_{\{0,1\}}(y), \tag{2.1.1}$$

Como el parámetro θ está restringido al espacio $\Theta = [0,1]$, entonces es posible formular varias opciones para la distribución previa del parámetro. En particular, la distribución uniforme restringida al intervalo $[0,1]$ o la distribución Beta parecen ser buenas opciones. Dado que la distribución uniforme es un caso particular de la distribución Beta, entonces vamos a trabajar con ésta. Por lo tanto la distribución previa del parámetro θ está dada por

$$p(\theta \mid \alpha, \beta) = \frac{1}{Beta(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} I_{[0,1]}(\theta). \tag{2.1.2}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.1.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid Y \sim Beta(y+\alpha, \beta - y + 1)$$

Prueba.

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \frac{I_{\{0,1\}}(y)}{Beta(\alpha,\beta)} \theta^y \theta^{\alpha-1}(1-\theta)^{\beta-1}(1-\theta)^{1-y} I_{[0,1]}(\theta) \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{\beta-y+1-1} I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Beta(y+\alpha, \beta - y + 1)$. ■

Del anterior resultado, podemos ver que la familia de distribución Beta es conjugada con respecto a la familia de distribución Bernoulli. Ahora consideramos cuál sería la distribución previa no informativa de Jeffreys para el parámetro θ. De acuerdo a la Definición 1.5.2, tenemos que

$$p(\theta) \propto I(\theta)^{1/2}$$

donde $I(\theta)$ es la información de Fisher acerca del parámetro θ, que en este caso está dada por

$$\begin{aligned} I(\theta) &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{Y} \mid \theta)\right\} \\ &= -E\left\{\frac{\partial^2}{\partial\theta^2}\{Y\log\theta + (1-Y)\log(1-\theta)\}\right\} \\ &= E\left\{\frac{Y}{\theta^2} + \frac{1-Y}{(1-\theta)^2}\right\} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

De esta forma, tenemos que la distribución previa no informativa de Jeffreys debe ser proporcional a $\theta^{-1/2}(1-\theta)^{-1/2}$, el cual corresponde a la distribución $Beta(1/2, 1/2)$ cuya función de densidad se muestra en la figura 2.1 la cual asigna iguales pesos a los valores extremos del parámetro de interés y la no informatividad se representa en la simetría de la función alrededor del valor 0.5.

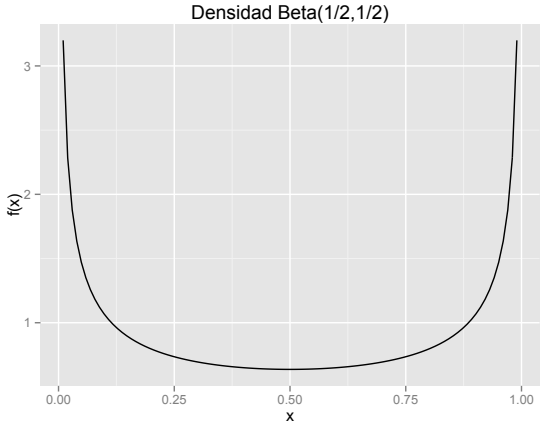


Figura 2.1: *Distribución previa no informativa de Jeffreys para el parámetro de una distribución Bernoulli*

Resultado 2.1.2. *La distribución predictiva previa para una observación y está dada por*

$$p(Y) = \frac{Beta(y+\alpha, \beta - y + 1)}{Beta(\alpha, \beta)} I_{\{0,1\}}(y), \tag{2.1.3}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(Y) &= \int p(Y \mid \theta)p(\theta \mid \alpha, \beta)\ d\theta \\ &= \int_0^1 \theta^y(1-\theta)^{1-y} I_{\{0,1\}}(y) \frac{1}{Beta(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\ d\theta \\ &= \frac{Beta(y+\alpha, \beta-y+1)}{Beta(\alpha,\beta)} I_{\{0,1\}}(y) \int_0^1 \frac{\theta^{y+\alpha-1}(1-\theta)^{\beta-y+1-1}}{Beta(y+\alpha,\beta-y+1)}\ d\theta \\ &= \frac{Beta(y+\alpha, \beta-y+1)}{Beta(\alpha,\beta)} I_{\{0,1\}}(y) \end{aligned}$$

Nótese que en la anterior demostración, la integral al lado derecho de la tercera igualdad es igual a la unidad, puesto que la expresión matemática dentro de la integral corresponde a la función de densidad de una variable aleatoria con distribucion $Beta$, que tiene rango en el intervalo $(0, 1)$. Por otro lado se deben verificar las dos condiciones de función de densidad. Es decir

1. $p(Y) > 0$ $(\forall y \in Y)$. Esta condición se tiene trivialmente puesto que la función matemática Beta siempre toma valores positivos.
2. $\int p(y)\ dx = 1$. En este caso, esta función es discreta definida en el conjunto $\{0, 1\}$. Por lo tanto esta condición es equivalente a

   $$\sum_{y\in\{0,1\}} P(Y = y) = \sum_{y\in\{0,1\}} \frac{Beta(y+\alpha, \beta-y+1)}{Beta(\alpha,\beta)} = 1$$

   Lo cual se verifica fácilmente teniendo en cuenta las propiedades de la función matemática Beta y de la función matemática Gamma.

■

La distribución predictiva dada en 2.1.3 está basada únicamente en la distribución previa del parámetro θ, una vez observada la variable Y se puede pensar en actualizar la distribución predictiva basando en la distribución posterior del parámetro, esta distribución se da en el siguiente resultado.

Resultado 2.1.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva observación $\tilde{y}$ está dada por*

$$p(\tilde{y} \mid Y) = \frac{Beta(\tilde{y}+y+\alpha, \beta-\tilde{y}-y+2)}{Beta(y+\alpha,\beta-y+1)} I_{\{0,1\}}(\tilde{y}), \tag{2.1.4}$$

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(\tilde{y} \mid Y) &= \int p(\tilde{y} \mid \theta)p(\theta \mid Y)\ d\theta \\ &= \int_0^1 \theta^{\tilde{y}}(1-\theta)^{1-\tilde{y}} I_{\{0,1\}}(\tilde{y}) \frac{\theta^{y+\alpha-1}(1-\theta)^{\beta-y+1-1}}{Beta(y+\alpha,\beta-y+1)}\ d\theta \\ &= \frac{Beta(\tilde{y}+y+\alpha, \beta-\tilde{y}-y+2)}{Beta(y+\alpha,\beta-y+1)} I_{\{0,1\}}(\tilde{y}) \\ &\qquad \times \int_0^1 \frac{\theta^{\tilde{y}+y+\alpha-1}(1-\theta)^{\beta-\tilde{y}-y+2-1}}{Beta(\tilde{y}+y+\alpha,\beta-\tilde{y}-y+2)}\ d\theta \\ &= \frac{Beta(\tilde{y}+y+\alpha, \beta-\tilde{y}-y+2)}{Beta(y+\alpha,\beta-y+1)} I_{\{0,1\}}(\tilde{y}) \end{aligned}$$

■

Ahora, en la práctica rara vez se observa la realización de una única variable aleatoria Bernoulli Y, sino una muestra de variables aleatorias $Y_1, \cdots, Y_n$. En este caso, la distribución posterior del parámetro θ está dada en el siguiente resultado.

Resultado 2.1.4. *Cuando se tiene una muestra aleatoria $Y_1, \ldots, Y_n$ de variables con distribución Bernoulli de parámetro θ, entonces la distribución posterior del parámetro de interés es*

$$\theta \mid Y_1, \ldots, Y_n \sim Beta\left(\sum_{i=1}^{n} y_i + \alpha, \beta - \sum_{i=1}^{n} y_i + n\right)$$

La demostración se deja como ejercicio.

Ejemplo 2.1.1. Es común en muchos países del mundo que se presenten encuestas de opinión electoral unas semanas antes de las elecciones presidenciales. Dentro de este tipo de encuestas se acostumbra a indagar acerca del favoritismo de los candidatos involucrados en la contienda electoral. Suponga que un candidato presidencial llamado José Pérez está interesado en conocer su intención de voto previa a las elecciones. Para esto, él contrata a una firma encuestadora para la realización de un muestreo probabilístico entre la población votante. El resultado de este estudio puede hacer cambiar o afirmar las estrategias publicitarias y la redefinición de la campaña electoral. La firma encuestadora decide implementar una estrategia de muestreo con un tamaño de muestra de doce mil personas. A cada respondiente se le realiza la siguiente pregunta: **Si las elecciones presidenciales fueran mañana. ?Usted votaría por el candidato José Pérez?**

Las respuestas a esta pregunta son realizaciones de una muestra aleatoria de doce mil variables con densidad Bernoulli. Los resultados del estudio arrojan que 6360 personas de las personas entrevistadas, es decir un 53 por ciento, votarían por el suscrito candidato. Técnicamente se debe analizar esta cifra puesto que las implicaciones de ganar en una primera vuelta son grandes en el sentido económico, logístico y administrativo. Claramente, el dato 53 por ciento asegura una ventaja dentro de la muestra de doce mil personas. Sin embargo, es necesario realizar un estudio más profundo acerca de la caracterización estructural de la intención de voto del candidato en la población de todos los votantes.

Con base en lo anteriormente expuesto, se decide utilizar la inferencia bayesiana puesto que existe información previa de un estudio anterior, contratado por el mismo candidato unos meses atrás en donde se entrevistaron a mil personas, con un favoritismo que estaba alrededor del 35 por ciento. Esta situación conlleva a la utilización de la metodología bayesiana que incorpora la información pasada acerca del mismo fenómeno.

El estadístico de la firma encuestadora decide utilizar una distribución previa[1] $Beta(\alpha = 350, \beta = 650)$. Utilizando el resultado 2.1.4, se contempla que la distribución posterior del parámetro de interés, que representa la probabilidad de éxito en las elecciones presidenciales, es $Beta(6360 + 350, 650 - 6360 + 12000) = Beta(6710, 6290)$. Por lo tanto, utilizando la distribución posterior, se estima que la intención de voto por el candidato es de $\frac{6710}{6710+6290} = \frac{6710}{13000} = 0.516$ y este valor equivale a la media de la distribución posterior. Este mismo análisis puede ejecutarse en `JAGS`, mediante el uso del siguiente código computacional

Sin embargo, si no se tuviese información previa como la suministrada por el estudio de meses anteriores, el análisis bayesiano sugeriría trabajar con una distribución previa no informativa, que en este caso, correspondería a una $Beta(\alpha = 0.5, \beta = 0.5)$. siguiendo el mismo análisis, se tiene que la distribución posterior es $Beta(6360.5, 5640.5)$. Finalmente, se estimaría que la intención de voto por el candidato es de $\frac{6350.5}{12001} = 0.529$. Las figuras 2.2 y 2.3 muestran el comportamiento de las distribuciones previas y posteriores en ambos escenarios. Nótese que la distribución no informativa influye muy poco en el comportamiento de la distribución posterior.

[1] Como se verá más adelante, es conveniente definir los parámetros de la distribución previa como α igual al número de votantes a favor y β igual al número de votantes en contra.

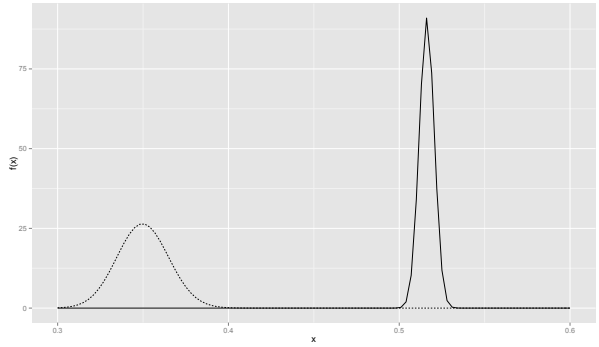


Figura 2.2: *Distribución previa informativa (línea punteada) y distribución posterior (línea sólida) para el ejemplo de las encuestas electorales.*

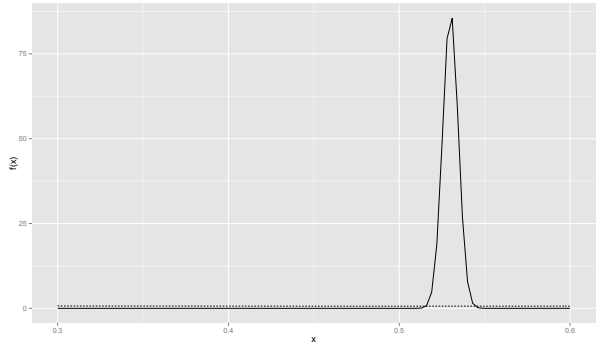


Figura 2.3: *Distribución previa no informativa (línea punteada) y distribución posterior (línea sólida) para el ejemplo de las encuestas electorales.*

Utilizando el siguiente código en R, es posible conocer los intervalos de credibilidad para las dos distribuciones posteriores. Es posible concluir que en ambos escenarios, el candidato aventaja significativamente a sus contrincantes y, salvo algún cambio drástico en el comportamiento del electorado, ganará las elecciones. Lo anterior se deduce puesto que el intervalo de credibilidad al 95 % no contiene ningún valor menor a 0.5

```
qbeta(c(0.025, 0.975), 6710, 6290)

## [1] 0.5076 0.5247

qbeta(c(0.025,0.975), 6350.5, 5640.5)

## [1] 0.5207 0.5385
```

Por otro lado, el siguiente código en JAGS permite obtener el mismo tipo de inferencia creando tres cadenas de Markov cuya distribución de probabilidad coincide con la distribución posterior del ejemplo.

```
# Datos
y <- c(1, 0, 1,..., 0)
n <-length(y)
```

```
# Modelo Bernoulli
Bern.model <-function() {
for(i in 1:n)
{
y[i]~dbern(theta)
}
theta~dbeta(350, 650)
}

bern.data <- list("y","n")
bern.param <- c("theta")
bern.inits <- function(){list("theta"=c(0.5))}
set.seed(123)

bern.fit <- jags(data=bern.data, inits=bern.inits, bern.param,
n.chains=3, n.iter=10000, n.burnin=1000, n.thin=10, model.file=bern.model)

print(bern.fit)
```

2.2 Modelo Binomial

Cuando se dispone de una muestra aleatoria de variables con distribución Bernoulli $Y_1, \ldots, Y_n$, la inferencia bayesiana se puede llevar a cabo usando la distribución Binomial, puesto que es bien sabido que la suma de variables aleatorias Bernoulli

$$S = \sum_{i=1}^{n} Y_i$$

sigue una distribución Binomial. Es decir:

$$p(S \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s} I_{\{0,1,\ldots,n\}}(s), \tag{2.2.1}$$

Nótese que la distribución binomial es un caso general para la distribución Bernoulli, cuando $n = 1$. Entonces, así como en la distribución Bernoulli, el parámetro θ está restringido al espacio $\Theta = [0, 1]$. Luego, es admisible proponer que θ siga una distribución Beta. Por tanto la distribución previa del parámetro θ está dada por la expresión (2.1.2). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.2.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid S \sim Beta(s + \alpha, \beta - s + n)$$

Prueba.

$$\begin{aligned} p(\theta \mid S) &\propto p(S \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \frac{\binom{n}{s} I_{\{0,1,\ldots,n\}}(s)}{Beta(\alpha, \beta)} \theta^s \theta^{\alpha-1} (1-\theta)^{\beta-1} (1-\theta)^{n-s} I_{[0,1]}(\theta) \\ &\propto \theta^{s+\alpha-1} (1-\theta)^{\beta-s+n-1} I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se llega a una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Beta(s+\alpha, \beta - s + n)$.

■

Del resultado anterior podemos ver que el estimador bayesiano de θ está dada por la media de la distribución posterior, dada por

$$\hat{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta} \tag{2.2.2}$$

En la práctica, se acostumbra a escoger los hiperparámetros α y β de tal forma que correspondan al número de éxitos y fracasos obtenidos en los datos previa, respectivamente. De esta forma, $\hat{\theta}_P = \alpha/(\alpha+\beta)$ corresponde a la estimación previa del parámetro θ. Por otro lado, el estimador clásico de θ está dado por $\hat{\theta}_C = s/n$. Entonces es posible notar que el estimador bayesiano de θ en (2.2.2) de alguna forma combina el estimador clásico y el estimador previa. Más aún, se puede ver que $\hat{\theta}_B$ se puede escribir como un promedio ponderado entre la estimación clásica y la estimación previa. Puesto que

$$\begin{aligned}
\hat{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta} &= \frac{s}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta} \\
&= \frac{n}{n+\alpha+\beta}\frac{s}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta} \\
&= \frac{n}{n+\alpha+\beta}\hat{\theta}_C + \frac{\alpha+\beta}{n+\alpha+\beta}\hat{\theta}_P
\end{aligned}$$

De esta forma, queda en evidencia que la estimación bayesiana de θ siempre será un valor intermedio entre la estimación clásica y la estimación previa. La gráfica 2.4 da una ilustración acerca de la anterior afirmación, en donde se puede observar que para una distribución previa concentrada en $2/7$ y una función de verosimilitud[2] con máximo en $8/10$, se tiene una distribución posterior centrada en $10/17$; es decir, la estimación bayesiana se encuentra situada entre la estimación previa y la estimación clásica.

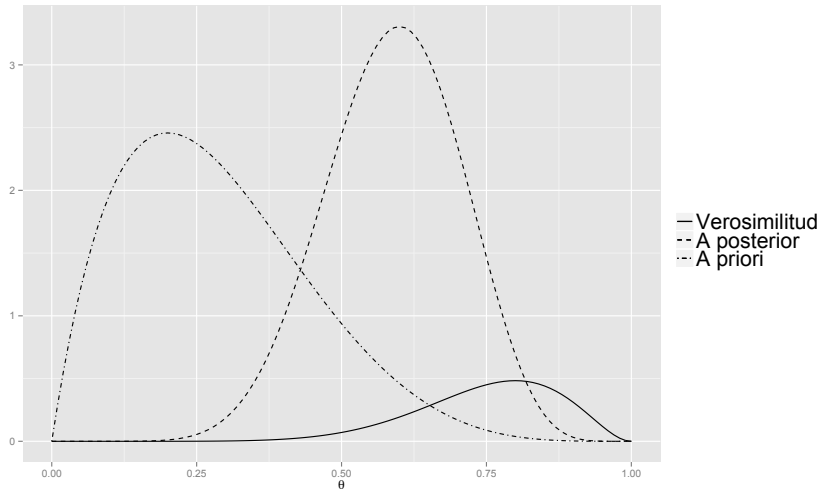


Figura 2.4: *Función de verosimilitud, función de densidad previa y posterior para* $\alpha = 2$, $\beta = 5$, $s = 8$ *y* $n = 10$.

[2]La función de verosimilitud es una función del parámetro y sólo se puede graficar una vez se hayan observado las realizaciones de la variable aleatoria.

Por otro lado, entre más grande sea el tamaño muestral n, más cercano estará $\hat{\theta}_B$ de $\hat{\theta}_C$ o equivalentemente la función de densidad posterior de θ estará más concentrada en s/n; mientras que entre mayor número de datos tenga la muestra de la distribución previa ($\alpha + \beta$=número de datos), más cercano estará $\hat{\theta}_B$ de $\hat{\theta}_P$ y la densidad posterior de θ estará más concentrada en $\alpha/(\alpha + \beta)$.

Para ilustrar lo anterior, suponga que la distribución previa de θ está dada con $\alpha = \beta = 5$, es decir la estimacion previa es 0.5, y suponga además que la estimación clásica es 0.33, pero el tamaño muestral n incrementa manteniendo constante la estimacion clásica. En la figura 2.5 se muestra la estimación posterior de θ, es evidente que a medida que el tamaño muestral n aumenta, la estimación posterior se acerca más a la estimación clásica.

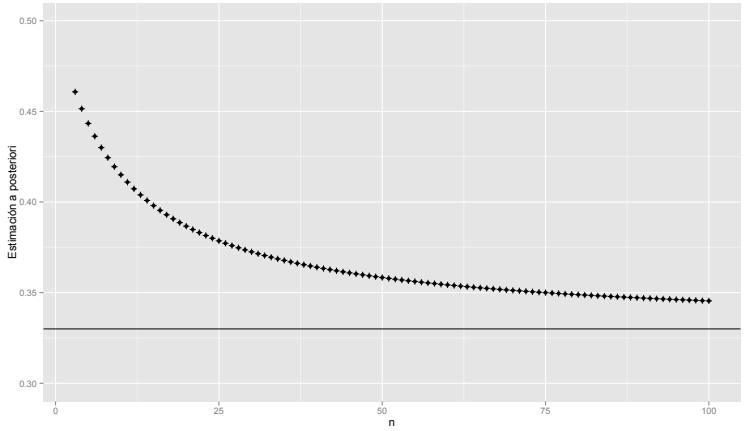


Figura 2.5: *Estimación posterior de θ para diferentes valores de n y s con $\alpha = \beta = 5$.*

Anteriormente, se comentó que se acostumbra a escoger los parámetros α y β que correspondan al número de éxitos y fracasos en la información previa, sin embargo, la información previa puede no presentarse de esta forma. Por ejemplo, en algunas situaciones, la información previa puede proveer el valor de θ, es decir, el valor de $\hat{\theta}_P$, y el valor de la desviación estándar de la estimación (comúnmente conocido como el error estándar). Por ejemplo, suponga que $\hat{\theta}_P = 0.5$ con un error estándar de 0.1, entonces podemos encontrar los valores de α y β de las expresiones $\frac{\alpha}{\alpha+\beta} = 0.5$ y $\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = 0.1$, de donde se tiene que $\alpha = 12$ y $\beta = 12$, y la distribución a priori correspondiente $Beta(12, 12)$ tiene una esperanza de 0.05 y una desviación estándar de 0.1. Se puede ver que entre mayor sea la desviación estándar, menores resultan los valores de α y β, que conducen a una distribución previa menos informativa.

Ahora, se vio anteriomente que la distribución previa no informativa de Jeffreys corresponde a la distribución $Beta(1/2, 1/2)$, la cual conduce a la distribución posterior $Beta(s + 1/2, n - s + 1/2)$, que a su vez nos lleva al estimador

$$\hat{\theta}_B = \frac{s + 1/2}{n + 1} \tag{2.2.3}$$

La anterior expresión es comparable con el estimador clásico $\hat{\theta_C} = \frac{s}{n}$, en el sentido de que los dos son aplicables cuando no se dispone de ninguna información previa. Podemos observar que aparte del alto grado de similitud que tienen los dos estimadores, es preferible usar el estimador (2.2.3) en situaciones donde el valor teórico de θ es muy pequeño, y como consecuencia en la muestra $s = 0$, por ejemplo, cuando θ representa el porcentaje de personas que esten infectados con algún virus poco común. En estos casos, el estimador clásico $\hat{\theta}_C = 0$ sugiriendo que ningún porcentaje de la población está infectado, conclusión que puede ser errónea; por otro lado, el estimador bayesiano $\hat{\theta}_B = \frac{0.5}{n+1}$, el cual tiende a ser un porcentaje muy pequeño a medida que aumente el tamaño muestral n, pero nunca

llega a dar el valor 0 como la estimación de θ.

En el siguiente resultado, se encuentra la distribución predictiva previa para una variable binomial S.

Resultado 2.2.2. *La distribución predictiva previa para la observación particular de la suma de variables aleatorias Bernoulli, s, está dada por una distribución Beta-Binomial dada por*

$$p(S) = \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s). \tag{2.2.4}$$

Prueba. De la definición de función de distribución predictiva previa se tiene que

$$\begin{aligned}
p(S) &= \int p(S \mid \theta) p(\theta \mid \alpha, \beta)\ d\theta \\
&= \int_0^1 \binom{n}{s} \theta^s (1-\theta)^{n-s} I_{\{0,1,\ldots,n\}}(s) \frac{1}{Beta(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\ d\theta \\
&= \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s) \\
&\qquad \times \int_0^1 \frac{\theta^{s+\alpha-1}(1-\theta)^{\beta-s+n-1}}{Beta(s+\alpha, \beta - s + n)}\ d\theta \\
&= \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s)
\end{aligned}$$

■

Una vez observados los valores muestrales, podemos encontrar la distribución predictiva posterior para una nueva variable binomial $\tilde{S}$ en una muestra de tamaño $\tilde{n}$. Esta distribución se encuentra en el siguiente resultado.

Resultado 2.2.3. *Después de la recolección de los datos $y_1, \cdots, y_n$, la distribución predictiva posterior para una nueva variable $\tilde{S}$ en una muestra del tamaño $\tilde{n}$ está dada por*

$$p(\tilde{s} \mid S) = \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}{Beta(s+\alpha, \beta - s + n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}), \tag{2.2.5}$$

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}
p(\tilde{s} \mid S) &= \int p(\tilde{s} \mid \theta) p(\theta \mid S)\ d\theta \\
&= \int_0^1 \binom{\tilde{n}}{\tilde{s}} \theta^{\tilde{s}} (1-\theta)^{\tilde{n}-\tilde{s}} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}) \frac{\theta^{s+\alpha-1}(1-\theta)^{\beta-s+n-1}}{Beta(s+\alpha, \beta - s + n)}\ d\theta \\
&= \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}{Beta(s+\alpha, \beta - s + n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}) \\
&\qquad \times \int_0^1 \frac{\theta^{\tilde{s}+s+\alpha-1}(1-\theta)^{\beta-\tilde{s}-s+n+\tilde{n}-1}}{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}\ d\theta \\
&= \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta - \tilde{s} - s + n + \tilde{n})}{Beta(s+\alpha, \beta - s + n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s})
\end{aligned}$$

■

En la anterior distribución predictiva, se necesita calcular funciones Beta. Cuando los tamaños muestrales n, $\tilde{n}$ y/o los parámetros de la distribución previa α y β son muy grandes, puede presentar

problemas numéricos al momento de calcular directamente estas funciones Beta. Por ejemplo, supongamos que $n = 1000$, $s = 650$, $\alpha = 200$, $\beta = 300$ y $\tilde{n} = 800$, de esta forma, los posibles valores para $\tilde{s}$ son $0, 1, \cdots, 800$, y se tiene que la probabilidad de que $\tilde{s}$ tome el valor 500 está dada por

$$Pr(\tilde{s} = 500|S) = \binom{800}{500} \frac{Beta(1350, 950)}{Beta(850, 650)} \tag{2.2.6}$$

y desafortunadamente, al evaluar `beta(1350,950)` y `beta(850,650)` en R da el valor de 0 para ambas expresiones. Planteamos la siguiente solución numérica cuando se quiere calcular la función predictiva (2.2.5) en muestras grandes. El problema central es el cómputo de $\frac{Beta(a,b)}{Beta(c,d)}$ con $a \geq c$ y $b \geq d$, valores enteros. Podemos ver que

$$\begin{aligned}
&\frac{Beta(a,b)}{Beta(c,d)} \\
&= \frac{(a-1)!(b-1)!(c+d-1)!}{(c-1)!(d-1)!(a+b-1)!} \\
&= \frac{(a-1)(a-2)\cdots(a-(a-c))(b-1)(b-2)\cdots(b-(b-d))}{(a+b-1)(a+b-2)\cdots(a+b-(a+b-c-d))} \\
&= \frac{a^{a-c}(1-\frac{1}{a})(1-\frac{2}{a})\cdots(1-\frac{a-c}{a})b^{b-d}(1-\frac{1}{b})(1-\frac{2}{b})\cdots(1-\frac{b-d}{b})}{(a+b)^{a+b-c-d}(1-\frac{1}{a+b})(1-\frac{2}{a+b})\cdots(1-\frac{a+b-c-d}{a+b})} \\
&= \underbrace{\left(\frac{a}{a+b}\right)^{a-c}}_{t1}\underbrace{\left(\frac{b}{a+b}\right)^{b-d}}_{t2}\underbrace{(1-\frac{1}{a})(1-\frac{2}{a})\cdots(1-\frac{a-c}{a})}_{t3} \\
&\quad \underbrace{(1-\frac{1}{b})(1-\frac{2}{b})\cdots(1-\frac{b-d}{b})}_{t4}\underbrace{(1-\frac{1}{a+b})(1-\frac{2}{a+b})\cdots(1-\frac{a+b-c-d}{a+b})}_{t5}
\end{aligned}$$

Calculando separadamente los términos $t1$, $t2$, $t3$, $t4$ y $t5$ podemos calcular $\frac{Beta(a,b)}{Beta(c,d)}$ para valores grandes de a, b, c y d. La siguiente función `prob` calcula la densidad (2.2.5) para un valor particular de $\tilde{s}$ usando la anterior técnica.

```
prob<-function(s.mono,n.mono,s,n,alfa,beta){
  a<-s.mono+s+alfa; b<-n.mono-s.mono+n-s+beta
  c<-s+alfa; d<-n-s+beta
  t1<-(a/(a+b))^(a-c); t2<-(b/(a+b))^(b-d)
  t3<-prod(1-c(1:(a-c))/a); t4<-prod(1-c(1:(b-d))/b)
  t5<-prod(1-c(1:(a+b-c-d))/(a+b))
  if(a==c){resul<- t2*t4/t5}
  if(b==d){resul<-t1*t3/t5}
  if(a>c&b>d){resul<-choose(n.mono,s.mono)*t1*t2*t3*t4/t5}
  resul
}
```

Si queremos examinar la distribución predictiva para todos valores de la variable $\tilde{S}$, podemos usar los siguientes códigos

```
n<-1000; s<-650
alfa<-200; beta<-300
n.mono<-800
res<-rep(NA,(1+n.mono))
for(i in 1:length(res)){
  res[i]<-prob(i-1,n.mono,s,n,alfa,beta)
}
```

y como resultado, `res` contiene las 801 probabilidades asociadas a todos los posibles valores de $\tilde{s}$.

Los resultados obtenidos con la anterior técnica es equivalente a lo obtenido usando la función `lbeta` que computa el logaritmo natural de la función beta. Así, para calcular la probabilidad en (2.2.6), simplemente usamos el siguiente código

```
choose(800,500)*exp(lbeta(1350,950)-lbeta(850,650))

## [1] 0.0005969
```

Nótese que esta probabilidad es la misma contenido en `res`, puesto que

```
res[501]

## [1] 0.0005969
```

Finalmente, se observa que la distribución predictiva (2.2.5) corresponde a una distribución Beta-binomial con parámetros $s+\alpha$ y $\beta - s + n$. Y el paquete `VGAM` en `R` (Yee 2012) contiene funciones que calculan la función de densidad, función de distribución, percentiles, además de generar números aleatorios para la distribución Beta-binomial. Las probabilidades puntuales de $\tilde{s}$ se puede calcular con la función `dbetabinom`, teniendo en cuenta que los parámetros utilizados son $\mu = (s+\alpha)/(n+\alpha+\beta)$ y $\rho = 1/(1+n+\alpha+\beta)$. Con el siguiente código, podemos calcular las probabilidades para todos los posibles valores de $\tilde{s}$.

```
library(VGAM)
mu<-(s+alfa)/(n+alfa+beta)
rho<-1/(1+n+alfa+beta)
res2<-rep(NA,(1+n.mono))
for(i in 1:length(res2)){
  res2[i]<-dbetabinom(i-1,size=n.mono,prob=mu,rho=rho)
}
```

Podemos ver que

```
res2[501]

## [1] 0.0005969
```

Lo cual es idéntico a lo obtenido anteriormente. Adicionalmente, al escribir la distribución predictiva de (2.2.5) como la función de densidad de una distribución Beta-binomial, se puede encontrar la esperanza de esta distribución, la cual está dada por

$$E(\tilde{S}|S) = \tilde{n}\frac{s+\alpha}{n+\alpha+\beta}$$

Nótese que la esperanza en la anterior expresión corresponde simplemente al tamaño $\tilde{n}$ de la nueva muestra multiplicado por la estimación bayesiana del parámetro θ. Adicionalmente, la esperanza de $\tilde{S}$ también se puede obtener muliplicando todos los posibles valores de $\tilde{S}$ con su respectiva probabilidad, y sumand al final, como se muestra a continuacion.

```
sum(res*c(0:n.mono))

## [1] 453.3

n.mono*(s+alfa)/(n+alfa+beta)

## [1] 453.3
```

Retomando el ejemplo 2.1.1, suponga que la encuesta de opinión electoral se lleva a cabo en diferentes ciudades de un determinado país, en este caso, para cada ciudad se tiene una muestra de variables con distribución Bernoulli o equivalentemente una variable binomial; de esta forma, se dispone de una muestra de variables con distribución Binomial. La distribución posterior del parámetro θ para estos casos se encuentra en el siguiente resultado.

Resultado 2.2.4. *Cuando se tiene una sucesión de variables aleatorias $S_1, \ldots, S_i, \ldots, S_k$ independientes y con distribución Binomial(n_i, θ) para $i = 1, \ldots, k$, entonces la distribución posterior del parámetro de interés θ es*

$$\theta \mid S_1, \ldots, S_k \sim Beta\left(\sum_{i=1}^{k} s_i + \alpha, \beta + \sum_{i=1}^{k} n_i - \sum_{i=1}^{k} s_i\right)$$

Prueba.

$$\begin{aligned} p(\theta \mid S_1, \ldots, S_k) &\propto \prod_{i=1}^{k} p(S_i \mid \theta) p(\theta \mid \alpha, \beta) \\ &\propto \prod_{i=1}^{k} \theta^{\sum_{i=1} s_i} \theta^{\alpha-1} (1-\theta)^{\beta-1} (1-\theta)^{\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} s_i} I_{[0,1]}(\theta) \\ &= \theta^{\sum_{i=1}^{k} s_i + \alpha - 1} (1-\theta)^{\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} s_i + \beta} I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de densidad de la distribución $Beta\left(\sum_{i=1}^{k} s_i + \alpha, \beta + \sum_{i=1}^{k} n_i - \sum_{i=1}^{n} s_i\right)$. ■

Ejemplo 2.2.1. El siguiente conjunto de datos fue estudiado inicialmente por Efron & Morris (1975) y se ha convertido en uno de los ejemplos prácticos más citados en la historia de la estadística moderna. Se trata de los porcentajes de bateo en una muestra de 18 jugadores profesionales en la temporada regular de béisbol en Estados Unidos en el año 1970. Wikipedia (2011*b*) establece que, en términos generales, este valor representa la razón entre la cantidad de *hits*[3] y el número de turnos al bate. La fórmula para calcular esta estadística es s/n, donde s es el número de *hits* y n es el total de turnos. Este conjunto de datos está disponible en el paquete '`pscl` de `R` y se puede cargar mediante el siguiente código computacional.

[3]Wikipedia (2011*a*) afirma que se anota como *hit* la conexión efectuada por el bateador que coloca la pelota dentro del terreno de juego, permitiéndole alcanzar al menos una base, sin que se produzca un error de defensa del equipo contrario. Para lograr un hit, el bateador debe llegar a primera base antes de que ningún jugador defensivo lo toque con la bola en el trayecto del home a la inicial, o que el jugador de la defensa que tenga la bola pise la primera base antes que el bateador llegue a la misma.

```
library(pscl)
data(EfronMorris)
head(EfronMorris)

##               name  team league  r     y   n     p
## 1 Roberto Clemente Pitts     NL 18 0.400 367 0.346
## 2   Frank Robinson  Balt     AL 17 0.378 426 0.298
## 3     Frank Howard  Wash     AL 16 0.356 521 0.276
## 4    Jay Johnstone   Cal     AL 15 0.333 275 0.222
## 5        Ken Berry   Chi     AL 14 0.311 418 0.273
## 6      Jim Spencer   Cal     AL 14 0.311 466 0.270
```

En las primeras cuatro columnas, se encuentran información sobre el número, el nombre de los jugadores, así como el equipo y la liga al cual pertenecen. La variable r denota el número de *hits* en los primeros 45 turnos al bate de la temporada del 1970, la variable y corresponde al porcentaje de *hits* en estos 45 turnos Las dos últimas variables hacen referencia al resto de la temporada: n denota el número de bates y p porcentaje de *hits*.

Suponga que, partiendo de la muestra de los 18 jugadores, el objetivo es estimar el porcentaje de *hits*, denotado como θ en el año de 1970. En primera instancia es plausible considerar que cada uno de los jugadores se comporta de manera independiente y que el porcentaje de *hits* es común a todos, puesto que pertenecen a ligas similares de un mismo país. Por lo tanto, es posible establecer que el número de hits s_i $(i = 1, \ldots, 18)$ para cada jugador tiene la siguiente distribución

$$S_i \sim Binomial(n_i, \theta) \qquad i = 1, \ldots, 18.$$

Utilizando un enfoque bayesiano, es posible sacar provecho de la información recolectada al principio de la temporada, en cuanto a los resultados de los primeros 45 turnos al bate. En esta instancia, se tuvieron $18 + 17 + \cdots + 8 + 7 = 215$ hits para un total de $45 \times 18 = 810$ turnos al bate. Con esta información, se define la caracterización estructural de la distribución previa que, siguiendo las recomendaciones anteriores, está dada por una $Beta(\alpha = 215, \beta = 810 - 215) = Beta(\alpha = 215, \beta = 595)$. Del resultado 2.2.4, y teniendo en cuenta que al final de la temporada se obtuvieron $\sum S_i = 1825$ hits para un total de $\sum n_i = 6649$ turnos al bate, se tiene que la distribución posterior para este ejemplo es una $Beta(1825 + 215, 6649 - 1825 + 595) = Beta(2040, 5419)$. Por lo tanto, utilizando la distribución posterior, se estima que el porcentaje de bateo en la liga profesional en el año de 1970 es de $\frac{2040}{2040+5419} = \frac{2040}{7459} = 0.273$. Este valor corresponde a la media de la distribución posterior. Por otro lado, los límites del intervalo de credibilidad corresponde a los percentiles teóricos de la distribución $Beta(2040, 5419)$, esto es, $(0.263,\ 0.284)$.

La figura 2.6 muestra el comportamiento de las distribuciones previa y posterior para este ejemplo.

Finalmente, notamos que los mismos resultados se encuentran cuando se analiza este conjunto de datos en `JAGS`, mediante el siguiente código computacional.

```
k <- nrow(EfronMorris)

Bin.model <- function(){
for(i in 1 : k)
{
s[i] ~ dbin(theta, n[i])
}
theta ~ dbeta(215, 595)
}
```

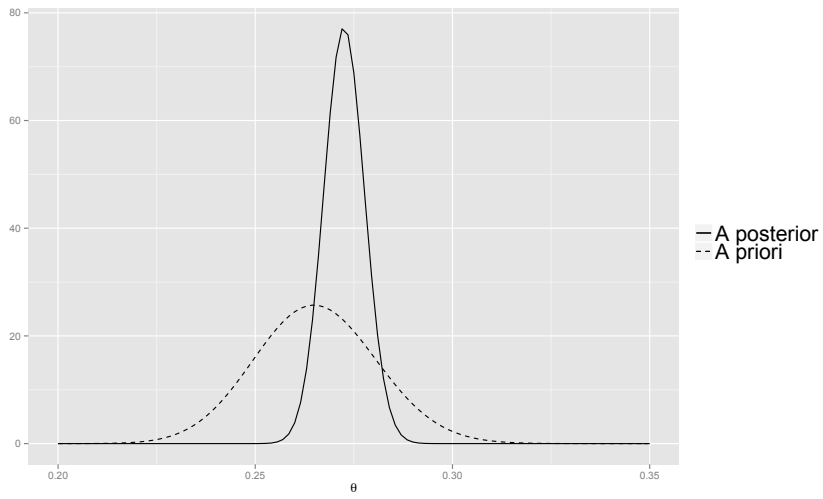


Figura 2.6: *Función de densidad previa y función de densidad posterior para el ejemplo de bateo.*

```
s <- round(EfronMorris$n*EfronMorris$p)
n <- EfronMorris$n

Bin.data <- list("s","n","k")
Bin.param <- c("theta")
Bin.inits <- function(){
list("theta"=c(0.5))
}

Bin.fit <- jags(data=Bin.data, inits=Bin.inits, Bin.param, n.iter=10000,
                n.burnin=1000, model.file=Bin.model)

print(Bin.fit)
```

Al ejecutar los anteriore comandos, se encuentran que la estimación bayesiana posterior del parámetro θ está dada por 0.274, mientras que el intervalo de credibilidad al 95 % es $(0.264,\ 0.284)$.

Es posible analizar este conjunto de datos desde otra perspectiva al suponer que los jugadores no constituyen una muestra aleatoria y cada uno de ellos tiene un promedio de bateo diferente. Sin embargo, este análisis se deja como ejercicio en un capítulo posterior.

Ejemplo 2.2.2. Continuando con el conjunto de datos de Efron y Morris, suponga que el entrenador de un equipo de las ligas inferiores está interesado en adquirir los servicios de Max Alvis. Este jugador no tuvo un buen promedio de bateo en la temporada (está situada en el último lugar) y no tuvo muchos turnos al bate (solo 70 turnos en el resto de la temporada). El entrenador quiere conocer cuál será el número más probable de hits que anotará en la siguiente temporada. Teniendo en cuenta que es un jugador que viene de la liga profesional, lo más conveniente es que tenga muchos turnos al bate, digamos 400.

Para resolver este cuestionamiento, es conveniente recurrir a la función predictiva posterior, dada en el resultado 2.2.3. Para este análisis, se define la caracterización estructural de la distribución previa del jugador que está dada por una $Beta(\alpha = 7, \beta = 38)$ (puesto que en la temporada del 1970, obtuvo 7 *hits* de 45 turnos). La siguiente función en R permite obtener la distribución predictiva para este

jugador, que se muestra en la figura 2.7.

```
n <- 70
s<- 14
alp <-7
bet <- 38
n.ast <- 400

predictiva <- rep(NA, n.ast+1)
for(k in 0:n.ast)
{
predictiva[(k+1)] <-
  choose(n.ast, k) * beta(k+s+alp, bet-k-s+n.ast+n) / beta(s+alp, bet-s+n)
}
```

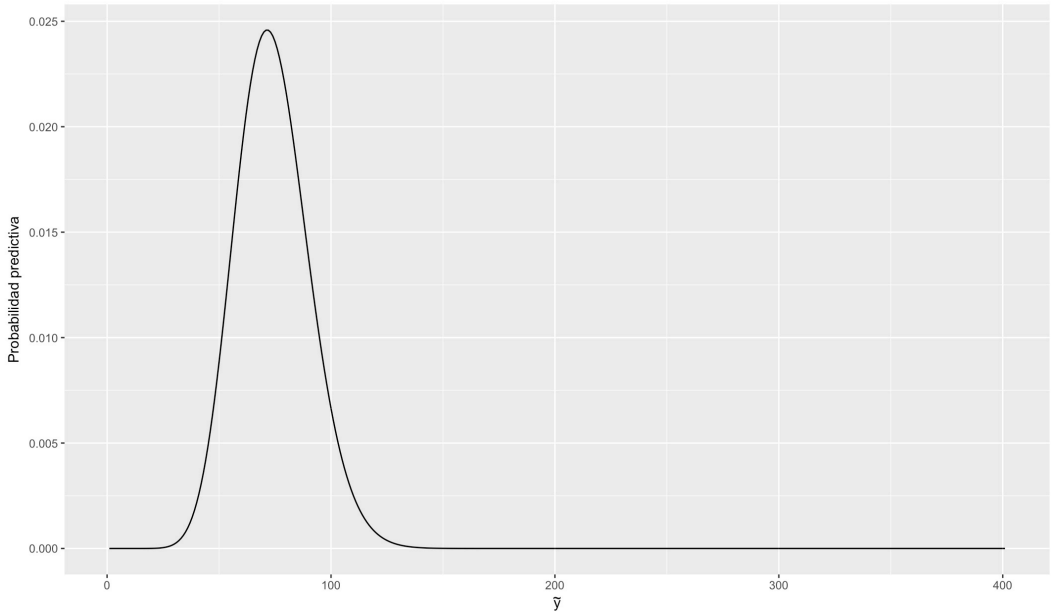


Figura 2.7: *Función de densidad predictiva posterior para el jugador Max Alvis.*

En la gráfica de la distribución predictiva posteiror, se observa que esta distribución se encuentra concentrado principalmenten en valores inferiores al 100 (el cual representa un porcentaje de *hits* del 0.25). Adicionalmente, examinando los valores numéricos se puede encontrar que lo más probable es que el jugador realice 71 hits en 400 turnos al bate, cifra que no convence al entrenador para adquirir los servicios del jugador.

2.3 Modelo Binomial negativa

La distribución binomial negativa describe el número de ensayos necesarios para alcanzar un número determinado y fijo de éxitos k en una secuencia independiente de experimentos tipo Bernoulli. ésta distribución es particularmente útil cuando el porcentaje θ que se quiere estimar es muy pequeño, como la proporción de una población que padece de alguna enfermedad. La razón por la que no se utiliza la distribución binomial es que al fijar el número de ensayos n, como el porcentaje θ es muy pequeño, es muy probable que en en la muestra de tamaño n no se encuetre ningún paciente con la enfermedad; mientras que al utilizar la distribución binomial negativa, de antemano se garantiza que se obtendrá k pacientes con la enfermedad en la muestra.

Suponga que Y es una variable aleatoria cuya distribución es Binomial negativa que representa el número de ensayos necesarios y para alcanzar un número determinado y fijo de éxitos k en un experimento. La forma funcional de esta distribución es la siguiente

$$p(Y \mid \theta) = \binom{y-1}{k-1}\theta^k(1-\theta)^{y-k}I_{\{k,k+1,\ldots,\}}(y), \tag{2.3.1}$$

Así como en la distribución Bernoulli y Binomial, el parámetro θ está restringido al espacio $\Theta = [0,1]$. Luego, es admisible proponer que θ siga una distribución Beta. Por tanto, la distribución previa del parámetro θ está dada por la expresión (2.1.2). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.3.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid Y \sim Beta(\alpha + k, \beta + y - k)$$

Prueba.

$$\begin{aligned} p(\theta \mid Y) &\propto p(Y \mid \theta)p(\theta \mid \alpha, \beta) \\ &\propto \theta^{\alpha+k-1}(1-\theta)^{y+\beta-k-1}I_{[0,1]}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se llega a una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Beta(\alpha + k, \beta + y - k)$. ■

En algunas situaciones se puede encontrar una muestra de variables con distribución binomial negativa, por ejemplo, la entrevista de pacientes para encontrar cierta enfermedad puede llevarse a cabo en diferentes puntos de atención médica o en diferentes ciudades del país. Así en cada punto de atención, se tendrá el dato correspondiente a una variable con distribución binomial negativa. El procedimiento inferencial bayesiano para estas situaciones se describe a continuación:

Resultado 2.3.2. *Cuando se tiene una sucesión de variables aleatorias $Y_1, \ldots, Y_n$ independientes y con distribución $BinomialNegativa(k_i, \theta)$ $(i = 1, \ldots, n)$, entonces la distribución posterior del parámetro de interés es*

$$\theta \mid Y_1, \ldots, Y_n \sim Beta(\alpha + \sum_{i=1}^{n} k_i, \beta + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} k_i) \tag{2.3.2}$$

Prueba. Se deja como ejercicio para el lector. ■

Ejemplo 2.3.1. Una franquicia de investigación farmacéutica ha desarrollado un nuevo tratamiento farmacológico sobre pacientes diabéticos que padezcan, a su vez, de enfermedades cardíacas, mejor conocidas como cardiopatías, entre las que se pueden encontrar la angina de pecho, infarto de miocardio, insuficiencia mitral, estenosis mitral, entre otras. Para evaluar el nuevo tratamiento, es necesario

seleccionar una muestra, mediante el diseño de un experimento clínico, de pacientes que tienen estas características.

Por otro lado, se sabe que la proporción de personas que padecen de diabetes y que además tienen algún tipo de condición cardiaca es muy baja y es necesario obtener una estimación precisa de la proporción de personas con estas condiciones. Con base en lo anteriormente expuesto, se puede pensar en seleccionar una grande de personas y utilizar un acercamiento binomial para estimar esta proporción. Sin embargo, dado que la prevalencia de esta condición es bastante baja, es posible que el número de personas en la muestra que presenten estas enfermedades sea nulo; por consiguiente, la estimación binomial no será, de ninguna forma, precisa.

Por lo tanto, el diseño clínico está supeditado al uso de la distribución Binomial Negativa, en donde se entrevistarán pacientes, de una base de datos de un hospital de la ciudad asociado con la franquicia, hasta conseguir una muestra de cinco pacientes que padezcan de estas condiciones. Después de varios meses de entrevistas, se encontró el quinto paciente en la entrevista número 1106.

Mediante el análisis bayesiano, suponiendo una distribución previa $Beta(0.5, 0.5)$, se llega a que la distribución posterior del parámetrtos θ es $Beta(0.5 + 5, 0.5 + 1106 - 5) = Beta(5.5, 1101.5)$. Por lo tanto, la estimación puntual del parámetro de interés, que corresponde a la media de la distribución posterior, es 0.0049, que equivale una proporción de 0.49 % de personas con estas enfermedades. El siguiente código computacional muestra cómo se puede llegar a las mismas conclusiones con `JAGS`

```
BinNeg.model <- function(){
y ~ dnegbin(theta,5)
theta~dbeta(0.5, 0.5)
}

#BinNeg.data <-
#list(y =1106)
BinNeg.param <- c("theta")
BinNeg.inits <- function(){
list("theta"=0.5)
}

BinNeg.fit <- jags(data=list(y=1106), inits=BinNeg.inits, BinNeg.param,
                   n.iter=10000, n.burn=1000, model.file=BinNeg.model)

print(BinNeg.fit)
```

Después de cinco mil iteraciones, la salida del anterior código muestra la estimación puntual dada por 0.005 y un intervalo de credibilidad al 95 %, dado por $(0.002, 0.01)$.

Ejemplo 2.3.2. Continuando con la temática del ejemplo anterior, suponga que la franquicia llevó a cabo la misma investigación en las 31 ciudades con mayor densidad poblacional de país. Como en la mayoría de los casos, debido al condicionamiento presupuestal, el experimento difirió en el número de éxitos en cada caso. En total, se tuvieron 29620 entrevistas para un total de éxitos de 152, tal como se muestra a continuación.

Ciudad	y	k
BOGOTA	1001	4
MEDELLIN	978	6
CALI	999	5
BARRANQUILLA	860	4

CARTAGENA	1155	4
CUCUTA	585	6
BUCARAMANGA	1030	3
IBAGUE	960	5
SOLEDAD	1002	6
SANTA MARTA	763	7
SOACHA	1036	5
PASTO	779	5
MONTERIA	1158	4
VILLAVICENCIO	1017	5
BELLO	888	6
MANIZALES	977	4
VALLEDUPAR	1256	6
BUENAVENTURA	1349	6
NEIVA	1047	5
PALMIRA	1088	5
ARMENIA	649	3
POPAYAN	765	4
FLORIDABLANCA	699	5
SINCELEJO	1042	4
ITAGUI	1212	5
BARRANCABERMEJA	660	5
TULUA	671	5
ENVIGADO	835	6
DOSQUEBRADAS	997	5
RIOHACHA	1146	4
SINCELEJO	1016	5

Mediante el análisis bayesiano, suponiendo una distribución previa[4] no informativa $Beta(0.5, 0.5)$, se llega a que la distribución posterior del parámetrtos θ es $Beta(0.5 + 152, 0.5 + 29620 - 152) = Beta(152.5, 29468.5)$. Por lo tanto, la estimación puntual del parámetro de interés, que corresponde a la media de la distribución posterior, es 0.0051, que equivale una proporción de 0.51 % de personas con estas enfermedades. El siguiente código computacional muestra cómo se puede llegar a las mismas conclusiones con JAGS

```
BinNeg2.model <- function(){
for(i in 1:31){
y[i]~dnegbin(theta,k[i])
}
theta ~ dbeta(0.5, 0.5)
}

y <- c(1001, 978, 999, 860, 1155, 585, 1030, 960, 1002, 763, 1036, 779, 1158, 1017,
       888, 977, 1256, 1349, 1047, 1088, 649, 765, 699, 1042, 1212, 660, 671, 835, 997, 1146, 1016)
k <- c(4, 6, 5, 4, 4, 6, 3, 5, 6, 7, 5, 5, 4, 5, 6, 4, 6, 6, 5, 5, 3, 4, 5, 4,
       5, 5, 5, 6, 5, 4, 5)

BinNeg2.data <- list("y", "k")
```

[4]Nótese que es posible también asignar una previa informativa $Beta(5.5, 1101.5)$, que da cuenta de la información del estudio del ejemplo anterior.

```
BinNeg2.param <- c("theta")
BinNeg2.inits <- function(){
list("theta"=0.5)
}
BinNeg2.fit <- jags(data=BinNeg2.data, inits=BinNeg2.inits, BinNeg2.param, n.iter=10000, n.burnin=100

print(BinNeg2.fit)
```

Después de cinco mil iteraciones, la salida del anterior código muestra la estimación puntual dada por 0.005 y un intervalo de credibilidad al 95 %, dado por $(0.004, 0.006)$, mucho más estrecho que el intervalo de credibilidad del anterior ejercicio.

Una vez observados los datos actuales y encontrada la distribución posterior, se puede encontrar la distribución predictiva posterior de una nueva variable con distribución binomial negativa. Es decir, se puede definir el mecanismo probabilístico para el número de ensayos necesarios para encontrar $\tilde{k}$ éxitos.

Resultado 2.3.3. *Después de la recolección de datos, la distribución predictiva posterior para una nueva variable $\tilde{Y}$ está dada por*

$$p(\tilde{Y}|Y_1,\cdots,Y_n) = \binom{\tilde{y}-1}{\tilde{k}-1}\frac{Beta(\alpha+\tilde{k}+\sum k_i, \beta+\tilde{y}-\tilde{k}+\sum y_i-\sum k_i)}{Beta(\alpha+\sum k_i, \beta+\sum y_i-\sum k_i)} I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y})$$

Prueba.

$$\begin{aligned}
&p(\tilde{Y}|Y_1,\cdots,Y_n)\\
&= \int p(\tilde{Y}|\theta)p(\theta|Y_1,\cdots,Y_n)d\theta\\
&= \int_0^1 \binom{\tilde{y}-1}{\tilde{k}-1}\theta^{\alpha+\tilde{k}}(1-\theta)^{\beta+\tilde{y}-\tilde{k}} I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y})\frac{\theta^{\sum k_i-1}(1-\theta)^{\sum y_i-\sum k_i-1}}{Beta(\alpha+\sum k_i, \beta+\sum y_i-\sum k_i)}d\theta\\
&= \binom{\tilde{y}-1}{\tilde{k}-1}\frac{I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y})}{Beta(\alpha+\sum k_i, \beta+\sum y_i-\sum k_i)}\int_0^1 \theta^{\alpha+\tilde{k}+\sum k_i-1}(1-\theta)^{\beta+\tilde{y}-\tilde{k}+\sum y_i-\sum k_i-1}d\theta\\
&= \binom{\tilde{y}-1}{\tilde{k}-1}\frac{Beta(\alpha+\tilde{k}+\sum k_i, \beta+\tilde{y}-\tilde{k}+\sum y_i-\sum k_i)}{Beta(\alpha+\sum k_i, \beta+\sum y_i-\sum k_i)} I_{\{\tilde{k},\tilde{k}+1,\cdots\}}(\tilde{y})
\end{aligned}$$

■

Ejemplo 2.3.3. Siguiendo con los datos del Ejemplo 2.3.2, suponga que se quiere recolectar información de 3 pacientes con cardiopatía en cierta ciudad, y se quiere conocer acerca del número de entrevistas necesarias. Utilizando la distribución previa $Beta(0.5, 0.5)$ y los datos de las 31 ciudades del ejemplo, se tiene que la distribución predictiva para el número de entrevistas necesarias para encontrar 3 pacientes está dada por

$$\begin{aligned}
&p(\tilde{Y}|Y_1,\cdots,Y_n)\\
&= \binom{\tilde{y}-1}{4}\frac{Beta(0.5+5+152, 0.5+\tilde{y}-5+29620-152)}{Beta(0.5+152, 0.5+29620-152)} I_{\{5,6,\cdots\}}(\tilde{y})\\
&= \binom{\tilde{y}-1}{4}\frac{Beta(157.5, \tilde{y}+29463.5)}{Beta(152.5, 29468.5)} I_{\{5,6,\cdots\}}(\tilde{y})
\end{aligned}$$

Con los siguientes códigos se puede calcular la anterior función predictiva.

```
predict<-function(y,alfa,beta,s,n,k){
  choose(y-1, k-1) * exp(lbeta(alfa+k+s, beta+y-k+n-s) - lbeta(alfa+s, beta+n-s))
}

alfa <- beta <- 0.5
s <- 152; n <- 29620; k <- 5
fun <- rep(NA)
for(y in 5:5000){
   fun[y-4] <- predict(y, alfa, beta, s, n, k)
}
sum(fun)

## [1] 1
```

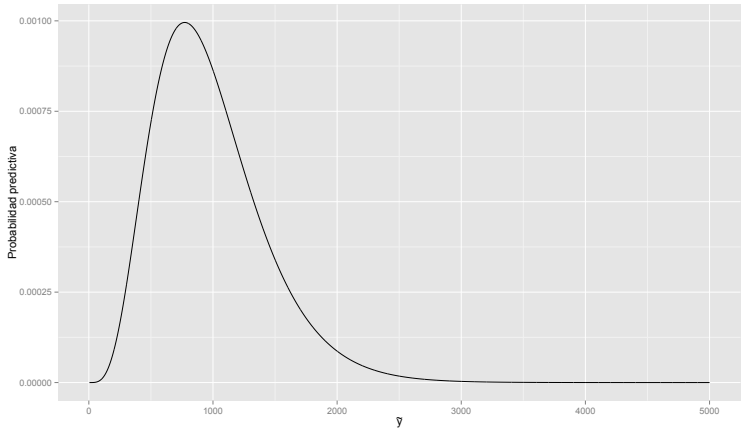


Figura 2.8: *Distribución predictiva posterior para el número de entrevistas necesarias para encontrar 5 pacientes usando los datos del ejemplo 2.3.2.*

Se puede ver que el número de entrevistas que tiene mayor probabilidad asociadas es el valor 772, usando el comando

```
which(fun==max(fun))+4

## [1] 772
```

También, se puede calcular la probabilidad de que en menos de 500 entrevistas se encuentren los 5 pacientes es de solo el 12 % usando el comando

```
sum(fun[1:(500-4)])

## [1] 0.1201
```

2.4 Modelo Poisson

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ es una muestra aleatoria de variables con distribución Poisson con parámetro θ, la función de distribución conjunta o la función de verosimilitud está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} I_{\{0,1,\ldots\}}(y_i) \\ &= \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n) \end{aligned}$$

donde $\{0, 1 \ldots\}^n$ denota el producto cartesiano n veces sobre el conjunto $\{0, 1 \ldots\}$. Por otro lado, como el parámetro θ está restringido al espacio $\Theta = (0, \infty)$, entonces es posible formular varias opciones para la distribución previa del parámetro. Algunas opciones se encuentran considerando la distribución exponencial o la distribución chi-cuadrado o la distribución Gamma. Nótese que las dos primeras distribuciones son casos particulares de la última. Por lo tanto, la distribución previa del parámetro θ está dada por

$$p(\theta \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta). \tag{2.4.1}$$

Bajo este marco de referencia se tienen el siguiente resultado con respecto a la distribucion posterior del parámetro de interés θ.

Resultado 2.4.1. *La distribución posterior del parámetro θ está dada por*

$$\theta \mid \mathbf{Y} \sim Gamma\left(\sum_{i=1}^{n} y_i + \alpha, n + \beta\right)$$

Prueba.

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^{n} y_i!} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} \theta^{\sum_{i=1}^{n} y_i} e^{-\beta\theta} e^{-n\theta} I_{(0,\infty)}(\theta) \\ &\propto \theta^{\sum_{i=1}^{n} y_i + \alpha - 1} e^{-(\beta+n)\theta} I_{(0,\infty)}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma(\sum_{i=1}^{n} y_i + \alpha, n + \beta)$. ■

Utilizando el resultado anterior, se tiene que la estimación Bayesiana del parámetro θ está dada por

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i + \alpha}{n + \beta}.$$

La anterior expresión sugiere tomar los parámetros de la distribución previa α y β de la siguiente manera: β representa el número de observaciones en la información previa, mientras que α representa la suma de los datos de la información previa. De esta forma, α/β representa la estimación previa del parámetro θ. Y la estimación Bayesiana de θ se puede escribir como

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^{n} y_i + \alpha}{\beta + n} \\ &= \frac{n}{n+\beta} * \frac{\sum y_i}{n} + \frac{\beta}{n+\beta} * \frac{\alpha}{\beta} \\ &= \frac{n}{n+\beta} * \hat{\theta_C} + \frac{\beta}{n+\beta} * \hat{\theta_P} \end{aligned}$$

Es decir, la estimación Bayesiana de θ es un promedio ponderado entre la estimación clásica y la estimación previa del parámetro θ, donde los pesos dependen directamente del tamaño muestral de la información actual y de la información previa.

A continuación estudiamos las distribuciones predictivas previa y posterior de una nueva observación

Resultado 2.4.2. *La distribución predictiva previa para una observación* $\mathbf{y} = \{y_1, \ldots, y_n\}$ *de la muestra aleatoria está dada por*

$$p(\mathbf{Y}) = \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}} \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!} \tag{2.4.2}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}
p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta)\ d\theta \\
&= \int_0^\infty \frac{e^{-n\theta}\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n) \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\ d\theta \\
&= \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}} \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!} \\
&\qquad \times \int_0^\infty \frac{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}}{\Gamma(\sum_{i=1}^n y_i + \alpha)} \theta^{\sum_{i=1}^n y_i + \alpha - 1} e^{-(\beta+n)\theta}\ d\theta \\
&= \frac{\Gamma(\sum_{i=1}^n y_i + \alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(n+\beta)^{\sum_{i=1}^n y_i + \alpha}} \frac{I_{\{0,1,\ldots\}^n}(y_1, \ldots, y_n)}{\prod_{i=1}^n y_i!}
\end{aligned}$$

■

En el caso en que la muestra aleatoria estuviera constituida por una sola variable aleatoria, entonces $n = 1$ y si, en particular, los hiper-parámetros de la distribución previa fuesen $\alpha = \beta = 1$, entonces no es difícil ver, utilizando la definición de la función matemática Gamma, que la función de distribución predictiva (2.4.2) estaría dada por

$$\begin{aligned}
p(Y) &= \frac{\Gamma(y+1)}{\Gamma(1)} \frac{1}{2^{y+1}} \frac{I_{\{0,1,\ldots\}}(y)}{y!} \\
&= \frac{1}{2^{y+1}} I_{\{0,1,\ldots\}}(y)
\end{aligned} \tag{2.4.3}$$

Para chequear la convergencia de la anterior distribución es necesario recurrir a los resultados del análisis matemático (Apostol 1957, p. 361). Dado que el espacio de muestreo de la variable aleatoria Y es $\{0, 1, \ldots\}$, entonces la suma infinita converge a uno lo que conlleva a que, en este caso particular, $P(Y)$ sea una auténtica función de densidad de probabilidad.

$$\sum_{y=0}^\infty p(Y = y) = \sum_{y=0}^\infty \left(\frac{1}{2}\right)^{y+1} = \frac{1}{2} \sum_{y=0}^\infty \left(\frac{1}{2}\right)^y = \frac{1}{2} \frac{1}{1 - 1/2} = 1$$

y podemos afirmar que la expresion (2.4.3) sí representa una función de densidad de una variable discreta. Ahora, consideramos la distribución predictiva poseterior de una muestra aleatoria, esta distribución se presenta en el siguiente resultado.

Resultado 2.4.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva posible observación* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$*, de tamaño* n^**, está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{\Gamma(\sum_{i=1}^{n^*} \tilde{y}_i + \sum_{i=1}^{n} y_i + \alpha)}{\Gamma(\sum_{i=1}^{n} y_i + \alpha)} \frac{(\beta+n)^{\sum_{i=1}^{n} y_i + \alpha}}{(n^*+\beta+n)^{\sum_{i=1}^{n^*} \tilde{y}_i + \sum_{i=1}^{n} y_i + \alpha}} \times \frac{I_{\{0,1,\ldots\}^{n^*}}(\tilde{y}_1, \ldots, \tilde{y}_{n^*})}{\prod_{i=1}^{n^*} \tilde{y}_i!} \tag{2.4.4}$$

Prueba. De la definición de función de distribución predictiva, y haciendo uso del mismo razonamiento en la demostración del Resultado 2.4.2, se tiene la prueba inmediata. ■

La anterior distribución corresponde a una distribucion multivariada que nos permite calcular probabilidades predictivas para cualesquieras valores de $\tilde{y}_1, \cdots, \tilde{y}_{n^*}$; sin embargo, en algunas situaciones, como por ejemplo, cuando θ representa el número promedio de algún suceso en una región geográfica, entonces al momento de la predicción, podemos estar interesados en predecir el número total o el número promedio de sucesos en la nueva muestra aleatoria de regiones geográficos. Es decir, podemos estar más interesados en la distribución de $\sum_{i=1}^{n^*} \tilde{y}_i$ o de $\sum_{i=1}^{n^*} \tilde{y}_i/n^*$ en vez de la distribución conjunta de $\tilde{y}_1, \cdots, \tilde{y}_{n^*}$. La distribución predictiva de $\sum_{i=1}^{n^*} \tilde{y}_i$ se presenta en el siguiente resultado, y con esta distribución se puede obtener fácilmente probabilidades predictivas para $\sum_{i=1}^{n^*} \tilde{y}_i/n^*$.

Resultado 2.4.4. *Después de la recolección de los datos, la distribución predictiva posterior para la suma de un vector de observaciones nuevas* $(\tilde{y}_1, \ldots, \tilde{y}_{n^*})$*,* $\tilde{s} = \sum_{y=1}^{n^*} \tilde{y}_i$*, está dada por:*

$$p(\tilde{\mathbf{s}} \mid \mathbf{Y}) = \frac{\Gamma(\tilde{s} + \sum_{i=1}^{n} y_i + \alpha)}{\Gamma(\sum_{i=1}^{n} y_i + \alpha)} \frac{(n+\beta)^{\sum_{i=1}^{n} y_i + \alpha}}{(n^*+n+\beta)^{\tilde{s} + \sum_{i=1}^{n} y_i + \alpha}} \frac{(n^*)^{\tilde{s}} I_{\{0,1,\ldots\}}(\tilde{s})}{\tilde{s}!} \tag{2.4.5}$$

Prueba. Usando el hecho de que $\theta|\mathbf{Y} \sim Gamma(\sum_{i=1}^{n} y_i + \alpha, n+\beta)$ y $\tilde{s}|\theta \sim Poisson(n^*\theta)$ se procede a calcular $\tilde{s}/p(\mathbf{y})$, así:

$$\begin{aligned}
& p(\tilde{s}|\mathbf{y}) \\
&= \int_\Omega p(\tilde{s}|\theta) p(\theta|\mathbf{y}) d\theta \\
&= \int_\Omega \frac{(n^*\theta)^{\tilde{s}} e^{-n^*\theta}}{\tilde{s}!} I_{\{0,1,\ldots\}}(\tilde{s}) (\beta+n)^{\sum_{i=1}^{n} y_i + \alpha} \frac{\theta^{\tilde{s} + \sum_{i=1}^{n} y_i + \alpha - 1}}{\Gamma(\sum_{i=1}^{n} y_i + \alpha)} e^{-(\beta+n)\theta} I_{(0,\infty)}(\theta) d\theta \\
&= \frac{(n^*)^{\tilde{s}} (\beta+n)^{\sum_{i=1}^{n} y_i + \alpha}}{\tilde{s}! \Gamma(\sum_{i=1}^{n} y_i + \alpha)} I_{\{0,1,\ldots\}}(\tilde{s}) \int_0^\infty \theta^{\sum_{i=1}^{n} y_i + \alpha - 1} e^{-(n^*+\beta+n)\theta} d\theta
\end{aligned}$$

Agrupando las constantes para obtener la integral de una distribución gamma con $\alpha = \tilde{s} + \sum_{i=1}^{n} y_i + \alpha$ y $\beta = n^* + n + \beta$ se obtiene el resultado. ■

En la práctica, evaluar directamente la expresión (2.4.5) puede ocasionar problemas numéricas, por la presencia de la función Gamma y las potencias. Para evitar dicha dificultad, podemos usar la siguiente expresión equivalente cuando $\tilde{s} = 1, 2, \cdots$:

$$p(\tilde{\mathbf{s}} \mid \mathbf{Y}) = \frac{\Gamma(\tilde{s})}{B(\tilde{s}, \sum_{i=1}^{n} y_i + \alpha)} \left(\frac{n+\beta}{n^*+n+\beta}\right)^{\sum_{i=1}^{n} y_i + \alpha} \frac{(n^*)^{\tilde{s}}}{(n^*+n+\beta)^{\tilde{s}} \tilde{s}!}$$

Cuando $\tilde{s} = 0$, la distribución predictiva es simplemente:

$$p(\tilde{\mathbf{s}} \mid \mathbf{Y}) = \left(\frac{n+\beta}{n^*+n+\beta}\right)^{\sum_{i=1}^n y_i+\alpha}$$

Ahora, debido a la complejidad de la expresión en (2.4.5), es prácticamente imposible comprobar analíticamente $\sum_{i=0}^{\infty} p(\tilde{s}=i)=1$, y también muy difícil encontrar una expresión matemática cerrada de la esperanza de la variable $\tilde{\mathbf{s}}$, sin embargo, en situaciones prácticas, se puede usar aproximaciones numéricas tal como se verá en el ejemplo al final de esta sección.

En el ejemplo 1.5.3 se consideró la situación cuando no se tiene ninguna información previa, la distribución previa que se debe usar está dada por

$$p(\theta) \propto \theta^{-1/2},$$

que corresponde a una distribución previa impropia, puesto que $\int_0^{\infty} \theta^{-1/2} = \infty$. Sin embargo, este hecho no afecta que la inferencia posterior se pueda llevar a cabo, puesto que la distribución posterior está dada por

$$\theta|\mathbf{Y} \sim Gamma(\sum y_i + 1/2, n)$$

y la estimación Bayesiana del parámetro θ viene dada por

$$\hat{\theta} = \frac{\sum y_i + 1/2}{n}.$$

la cual es muy similar a la estimación clásica de θ dada por $\bar{Y}$.

Cuando se utiliza la distribución previa no informativa de Jeffreys, la distribución predictiva para nuevas observaciones $\tilde{y} = \tilde{y}_1, \cdots, \tilde{y}_{n^*}$ y $\tilde{s} = \sum_{i=1}^{n^*} \tilde{y}_i$ están dadas por

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{\Gamma(\sum_{i=1}^{n^*} \tilde{y}_i + \sum_{i=1}^n y_i + 0.5)}{\Gamma(\sum_{i=1}^n y_i + 0.5)} \frac{n^{\sum_{i=1}^n y_i+0.5}}{(n^*+n)^{\sum_{i=1}^n \tilde{y}_i+\sum_{i=1}^n y_i+0.5}} \frac{I_{\{0,1,\ldots\}^{n^*}}(\tilde{y}_1,\ldots,\tilde{y}_{n^*})}{\prod_{i=1}^{n^*} \tilde{y}_i!} \tag{2.4.6}$$

y

$$p(\tilde{\mathbf{s}} \mid \mathbf{Y}) = \frac{\Gamma(\tilde{s} + \sum_{i=1}^n y_i + 0.5)}{\Gamma(\sum_{i=1}^n y_i + 0.5)} \frac{n^{\sum_{i=1}^n y_i+0.5}}{(n^*+n)^{\tilde{s}+\sum_{i=1}^n y_i+0.5}} \frac{I_{\{0,1,\ldots\}}(\tilde{s})}{\tilde{s}!} \tag{2.4.7}$$

Ejemplo 2.4.1. Por políticas gubernamentales, los alcaldes las ciudades están obligados a realizar un seguimiento exhaustivo al comportamiento de la accidentalidad en las vías urbanas y medirlo en términos del número de accidentes de tránsito. Lo anterior es necesario para evaluar la gestión de las autoridades administrativas y evaluar las políticas públicas que el gobierno de la ciudad ha implementado para disminuir esta cifra.

Suponga que la alcaldía de una ciudad quiere implementar una estrategia educativa para disminuir el número de accidentes de tránsito, generados por manejar en estado de embriaguez. Para esto, se registraron durante diez días 30 días el número de accidentes de tránsito por ebriedad del conductor. Los datos para cada uno de los días son 22, 9, 9, 20, 10, 14, 11, 14, 11, 11, 19, 12, 8, 9, 16, 8, 13, 8, 14, 12, 14, 11, 14, 13, 11, 14, 13, 11, 7, 12.

Es posible modelar la variable aleatoria número de accidentes de tránsito en un día mediante una distribución de Poisson puesto que el promedio muestral y la varianza muestral de los datos son semejantes. Para este conjunto de datos, el promedio equivale a 12.33, mientras que la varianza es de 12.51. El histograma de los valores observados se puede ver en la figura 2.9.

En primera instancia, es posible realizar un análisis no informativo, al formular una distribución previa de Jeffreys, utilizando el resultado del ejemplo 2.9 de la página 49, que indica que una distribución previa no informativa es proporcional a $\theta^{-1/2}$, para lo cual la distribución posterior $Gamma(\sum_{i=1}^n y_i +$

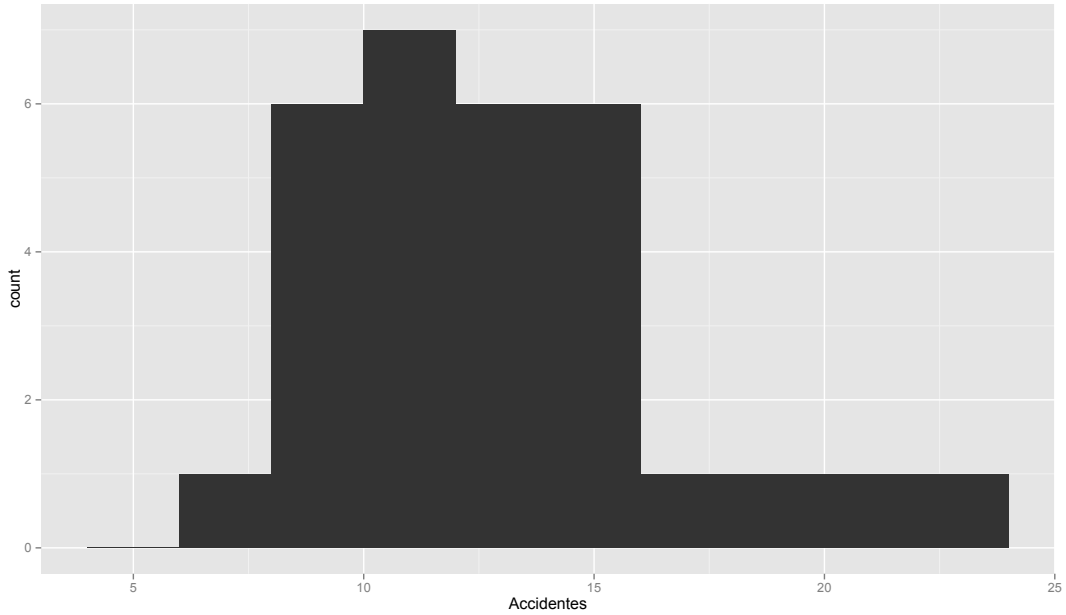


Figura 2.9: *Histograma para los datos de accidentes de tránsito.*

$1/2, n)$. De esta manera, la distribución posterior del parámetro de interés es $Gamma(370.5, 30)$. Por lo tanto, un estimador de θ está dado por la media de la distribución posterior que es $370.5/30 = 12.35$, muy cercano al valor del estimador de máxima verosimilitud correspondiente al promedio muestral. La figura 2.10 (lado izquierdo) muestra el comportamiento de las distribuciones de Jeffreys y posterior para este ejemplo.

Por otro lado, basándose en datos históricos, la alcaldía observó que, en el mismo periodo del año anterior, ocurrieron 37 accidentes en 9 días de observación. Luego, una distribución previa informativa[5] está dada por $Gamma(\alpha = 38, \beta = 9)$. Luego, apelando al resultado 2.4.1, la distribución posterior corresponde a una $Gamma(370+38, 30+9) = Gamma(408, 39)$. Para este caso, un estimador de θ está dado por la media de la distribución posterior que es $480/39 = 12.31$. La figura 2.10 (lado derecho) muestra el comportamiento de las distribuciones previa (informativa) y posterior para este ejemplo.

A continuación se examina la distribución predictiva. En la figura 2.11 se grafica la distribución predictiva para una nueva observación cuando se usa la previa no informativa y la previa informativa. Los códigos para el cálculo cuando se usa la previa no informativa es como siguen:

```
Trans <- c(22, 9, 9, 20, 10, 14, 11, 14, 11, 11, 19, 12, 8, 9, 16, 8, 13, 8, 14, 12,
14, 11, 14, 13, 11, 14, 13, 11, 7, 12 )
n <- length(Trans)
pre.Transito.NoInf <- function(s){
if(s>0){
  val <- gamma(s)*(n/(n+1))^(sum(Trans)+0.5)/
      (beta(s,sum(Trans)+0.5)*prod(1:s)*(n+1)^s)
}
if(s==0){
 val <- (n/(n+1))^(sum(Trans)+0.5)
```

[5]En la práctica, se recomienda que los valores de los hiperparámetros α y β correspondan a la suma del número de eventos más uno y número de observaciones, respectivamente.

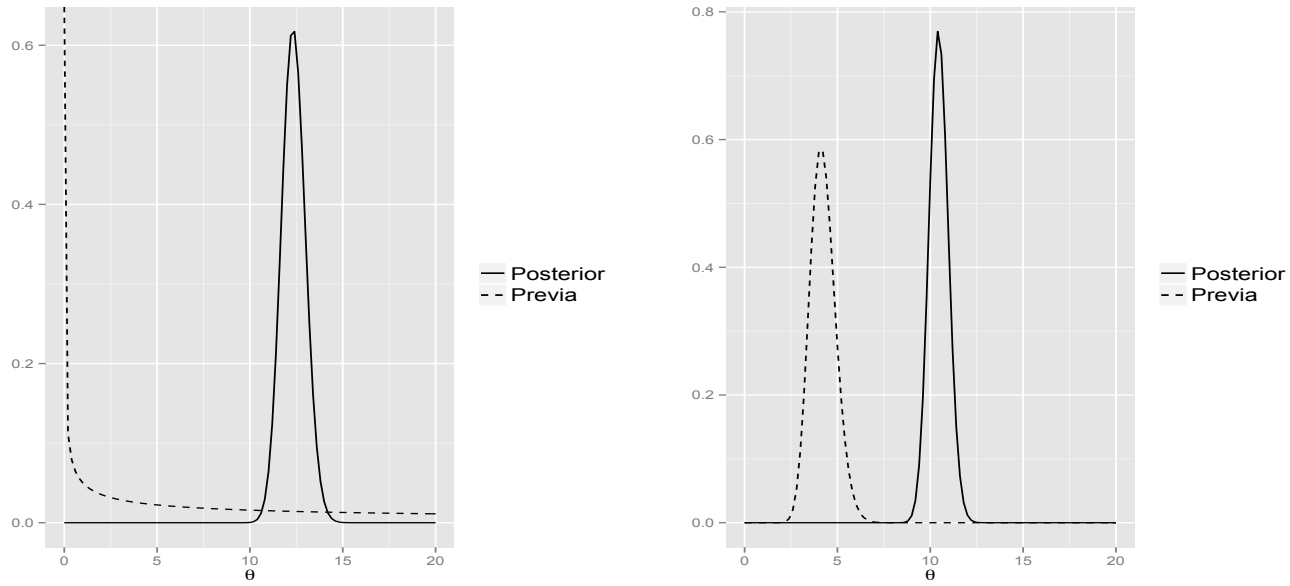


Figura 2.10: *Distribución previa y distribución posterior para el ejemplo del tránsito con dos distribuciones previas diferentes (el lado izquierdo representa el caso cuando se usa la previa no informativa, el lado derecho la previa informativa).*

```
}
val
}
s.max <- 40; s.val <- 0:s.max; pre.NoInf.val <- c()
for(i in 1:length(s.val)){
pre.NoInf.val[i] <- pre.Transito.NoInf(s.val[i])
}
sum(pre.NoInf.val)

## [1] 1
```

Nótese que en los anteriores códigos, se usó como valor máximo de 40 para la variable $\tilde{\mathbf{s}}$ a pesar de que ésta toma valores infinitos, pero al ver que la suma de las probabilidades desde el valor 0 hasta el máximo de 40 es igual a 1, podemos concluir que la probabilidad de que $\tilde{\mathbf{s}}$ tome valores mayores a 40 es prácticamente nula.

Adicionalmente, podemos tener una aproximación de la esperanza de la variable $\tilde{\mathbf{s}}$ como

```
sum(pre.NoInf.val*s.val)

## [1] 12.35
```

Finalmente, en la figura 2.4.1 se observa la distribución predictiva posterior usando dos diferentes distribuciones previas.

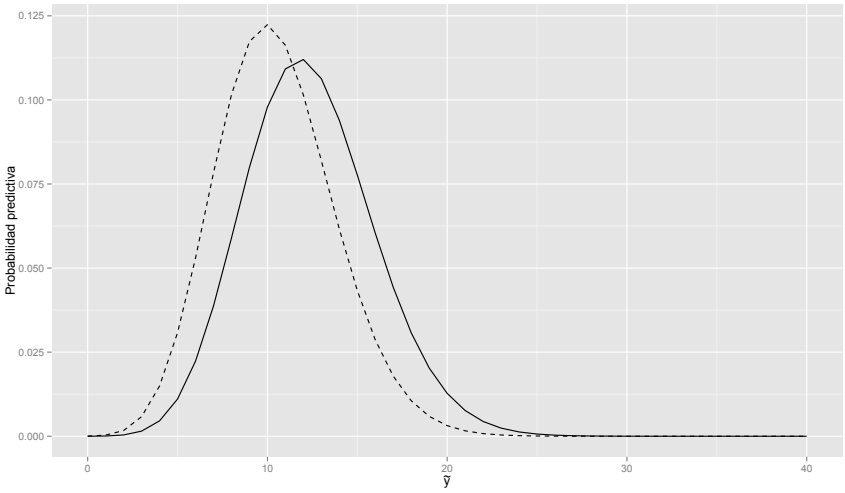


Figura 2.11: *Distribución predictiva posterior para $n* = 1$ para el ejemplo del tránsito. La línea sólida denota la distribución predictiva obtenida con la previa no informativa, la línea continua denota la obtenida con la previa $Gamma(\alpha = 38, \beta = 9)$.*

2.5 Modelo exponencial

Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ corresponde a una muestra de variables aleatorias con distribución Exponencial. Luego, la función de distribución conjunta o verosimilitud está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \theta e^{-\theta y} I_{(0,\infty)}(y_i) \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} y_i} I_{(0,\infty)^n}(y_1, \ldots, y_n) \end{aligned} \tag{2.5.1}$$

Donde $(0, \infty)^n$ denota el producto cartesiano n veces sobre el intervalo $(0, \infty)$. Por otro lado, como el parámetro θ está restringido al espacio $\Theta = (0, \infty)$, entonces es posible formular varias opciones para la distribución previa del parámetro, al igual que en la distribución Poisson. Así mismo, suponga que la distribución previa para el parámetro de interés es la distribución Gamma tal como aparece en la expresión (2.4.1). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.5.1. *La distribución posterior del parámetro θ sigue una distribución*

$$\theta \mid \mathbf{Y} \sim Gamma\left(\alpha + n, \beta + \sum_{i=1}^{n} y_i\right)$$

Prueba.

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta) \\ &= \theta^n e^{-\theta \sum_{i=1}^{n} y_i} I_{(0,\infty)^n}(y_1, \ldots, y_n) \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)} I_{(0,\infty)}(\theta) \\ &\propto \theta^{\alpha+n-1} e^{-(\beta + \sum_{i=1}^{n} y_i)} I_{(0,\infty)}(\theta) \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Gamma(\alpha + n, \beta + \sum_{i=1}^{n} y_i)$. ■

Resultado 2.5.2. *La distribución predictiva previa para una observación* $\mathbf{y} = \{y_1, \ldots, y_n\}$ *de la muestra aleatoria está dada por*

$$p(\mathbf{Y}) = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}} I_{(0,\infty)^n}(y_1, \ldots, y_n) \tag{2.5.2}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned} p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \theta) p(\theta \mid \alpha, \beta)\, d\theta \\ &= \int_0^\infty \theta^n e^{-\theta \sum_{i=1}^n y_i} I_{(0,\infty)^n}(y_1, \ldots, y_n) \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\, d\theta \\ &= \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}} I_{(0,\infty)^n}(y_1, \ldots, y_n) \\ &\qquad \times \int_0^\infty \frac{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(n+\alpha)} \theta^{\alpha+n-1} e^{-(\beta+\sum_{i=1}^n y_i)\theta}\, d\theta \\ &= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta + \sum_{i=1}^n y_i)^{\alpha+n}} I_{(0,\infty)^n}(y_1, \ldots, y_n) \end{aligned}$$

■

En el caso en que la muestra aleatoria estuviera constituida por una sola variable aleatoria, entonces no es difícil ver, utilizando la definición de la función matemática Gamma, que la función de distribución predictiva (2.5.2) estaría dada por

$$\begin{aligned} p(Y) &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta+y)^{\alpha+1}} I_{(0,\infty)}(y) \\ &= \frac{\alpha\beta^\alpha}{(\beta+y)^{\alpha+1}} I_{(0,\infty)}(y) \end{aligned}$$

Para chequear la convergencia de la anterior distribución es necesario recurrir a los resultados del cálculo integral. Dado que el espacio de muestreo de la variable aleatoria Y es el intervalo $(0, \infty)$, entonces la integral a uno lo que conlleva a que, en este caso particular, $P(Y)$ sea una auténtica función de densidad de probabilidad.

$$\int_0^\infty p(Y)\, dy = \int_0^\infty \frac{\alpha\beta^\alpha}{(\beta+y)^{\alpha+1}}\, dy = \beta^\alpha \left[\frac{(\beta+y)^{-\alpha}}{-\alpha} \right]_0^\infty = 1$$

Volviendo al caso general en donde se tiene una muestra aleatoria, se tiene el siguiente resultado.

Resultado 2.5.3. *Después de la recolección de los datos, la distribución predictiva posterior para una conjunto de nuevas variables aleatorias* $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{n^*}\}$*, de tamaño* n^**, está dada por*

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{\Gamma(n+\alpha+n^*)}{\Gamma(n+\alpha)} &\frac{(\beta + \sum_{i=1}^n y_i)^{n+\alpha}}{(\sum_{i=1}^{n^*} \tilde{y}_i + \beta + \sum_{i=1}^n y_i)^{n^*+\alpha+n}} \\ &\times I_{(0,\infty)^{n^*}}(\tilde{y}_1, \ldots, \tilde{y}_{n^*}) \end{aligned} \tag{2.5.3}$$

Prueba. De la definición de función de distribución predictiva, y haciendo uso del mismo razonamiento en la demostración del Resultado 2.5.2, se tiene la prueba inmediatamente. ■

El anterior resultado permite calcular la distribución predictiva conjunta de variables aleatorias por observar. En algunas situaciones, lo que se quiere pronosticar es el comportamiento probabilístico de promedio muestral de este conjunto de variables aleatorias, es decir, $\bar{Y}^* = \sum_{i=1}^{n^*} \tilde{Y}_i/n^*$. En el siguiente resultado se presenta la distribución predictiva de esta variable aleatoria.

Resultado 2.5.4. *Después de la recolección de los datos, la distribución predictiva posterior para el promedio muestral de un nuevo conjunto de variables aleatorias* $\bar{Y}^* = \sum_{i=1}^{n^*} \tilde{Y}_i/n^*$ *está dada por*

$$p(\bar{Y}^*) = \frac{n^*\Gamma(n^*+\alpha+n)}{\Gamma(n^*)\Gamma(\alpha+n)} \frac{(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{(n^*\bar{Y}^*+\beta+\sum y_i)^{n^*+\alpha+n}} (n^*\bar{Y}^*)^{n^*-1} I_{(0,\infty)}(\bar{Y}^*)$$

Prueba. En primer lugar se halla la distribución predictiva posterior de la variable $\tilde{S} = \sum_{i=1}^{n^*} \tilde{Y}_i$, teniendo en cuenta que $\tilde{S}|\theta \sim Gamma(n^*, \theta)$, de esta forma

$$\begin{aligned}
p(\tilde{S}|\mathbf{Y}) &= \int p(\tilde{S}|\theta)p(\theta|\mathbf{Y})\ d\theta \\
&= \int_0^\infty \frac{\theta^{n^*}}{\Gamma(n^*)} \tilde{S}^{n^*-1} e^{-\theta\tilde{S}} I_{(0,\infty)}(\tilde{S}) \frac{(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(\alpha+n)} \theta^{\alpha+n-1} e^{-(\beta+\sum y_i)\theta} d\theta \\
&= \frac{\tilde{S}^{n^*-1}(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(n^*)\Gamma(\alpha+n)} I_{(0,\infty)}(\tilde{S}) \int_0^\infty \theta^{n^*+\alpha+n-1} e^{-(\tilde{S}+\beta+\sum y_i)\theta}\ d\theta \\
&= \frac{\tilde{S}^{n^*-1}(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{\Gamma(n^*)\Gamma(\alpha+n)} \frac{\Gamma(n^*+\alpha+n)}{(\tilde{S}+\beta+\sum y_i)^{n^*+\alpha+n}} I_{(0,\infty)}(\tilde{S})
\end{aligned}$$

Al aplicar el teorema de transformación a la distribución predictiva, se puede hallar la distribución de $\bar{Y}^*$, dada por

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{n^*\Gamma(n^*+\alpha+n)}{\Gamma(n^*)\Gamma(\alpha+n)} \frac{(\beta+\sum_{i=1}^n y_i)^{\alpha+n}}{(n^*\bar{Y}^*+\beta+\sum y_i)^{n^*+\alpha+n}} (n^*\bar{Y}^*)^{n^*-1} I_{(0,\infty)}(\bar{Y}^*)$$

■

En la práctica puede ocurrir que alguos de los valores de n, n^*, $\sum_{i=1}^n y_i$ y $n^*\bar{Y}^*$ son muy grandes, y evaluar directamente la expresión anterior puede ocasionar problemas numéricos. Realizando algunas operaciones algebráicas, se encuentra la siguiente expresión equivalente para la distribución predictiva posterior de $\bar{Y}^*$ que evita problemas numéricas:

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{1}{\bar{Y}^* Beta(n, n^*)} \left(\frac{\beta+\sum_{i=1}^n y_i}{\beta+\sum_{i=1}^n y_i + n^*\bar{Y}^*}\right)^{\alpha+n} \left(\frac{n^*\bar{Y}^*}{\beta+\sum_{i=1}^n y_i + n^*\bar{Y}^*}\right)^{n^*} I_{(0,\infty)}(\bar{Y}^*) \quad (2.5.4)$$

Por otro lado, se puede verificar que la distribución previa no informativa de Jeffrey está dada por $p(\theta) \propto \theta^{-1}$, la cual combinada con la función de verosimilitud arroja la distribución posterior $Gamma(n, \sum_{i=1}^n y_i)$. También se puede ver que al utilizar la distribución previa no informativa de Jeffrey, la distribución predictiva posterior de $\bar{Y}^*$ está dada por

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{n^*\Gamma(n^*+n)}{\Gamma(n^*)\Gamma(n)} \frac{(\sum_{i=1}^n y_i)^n}{(n^*\bar{Y}^*+\sum y_i)^{n^*+n}} (n^*\bar{Y}^*)^{n^*-1} I_{(0,\infty)}(\bar{Y}^*) \quad (2.5.5)$$

La cual es equivalente a la siguiente expresión que en ocasiones puede ser útil para evitar problemas numéricos

$$p(\bar{Y}^*|\mathbf{Y}) = \frac{1}{\bar{Y}^* Beta(n, n^*)} \left(\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n y_i + n^*\bar{Y}^*}\right)^{n} \left(\frac{n^*\bar{Y}^*}{\sum_{i=1}^n y_i + n^*\bar{Y}^*}\right)^{n^*} I_{(0,\infty)}(\bar{Y}^*) \quad (2.5.6)$$

Ejemplo 2.5.1. Crowley & Hu (1977) reportan un conjunto de datos que da cuenta de los tiempos de sobrevivencia de $n = 69$ miembros del programa de transplante de corazón de Stanford. Los tiempos se reportan en días después del transplante. Los datos pueden ser encontrados en el paquete `survival` (Therneau & Lumley 2011) de `R`. A continuación se cargan los datos, y se calcula la variable de tiempo de sobrevida, computando la diferencia entre el tiempo de entrada y tiempo de salida de los individuos que efectivamente recibieron transplante.

```
require(survival)
data(heart)
attach(heart)

## The following object is masked from package:survival:
##
##    transplant

surv <- stop[transplant==1] - start[transplant==1]
```

A continuación se muestran los primeros datos de este estudio. Se recuerda que el total de pacientes atendidos en este estudio fue de $n = 69$ y la suma de los tiempos de sobrevida es de $\sum_{i=1}^{n} y_i = 25998.5$.

```
head(data.frame(heart[transplant==1,c("start","stop")], surv))

##    start stop surv
## 4      1   16   15
## 6     36   39    3
## 10    51  675  624
## 14    12   58   46
## 16    26  153  127
## 19    17   81   64
```

Estos tiempos pueden ser modelados mediante una distribución exponencial. Además de inferir acerca del parámetro de esta distribución, también es posible inferir acerca del tiempo promedio de sobrevivencia de un individuo sometido a este tipo de transplantes. Luego, dadas las implicaciones del estudio, se debe ser muy cuidadosos en la asignación de los parámetros de la distribución previa. Una forma de hacerlo es asignar valores muy pequeños a estos parámetros. Otra forma de hacerlo es utilizando la distribución previa de Jeffreys, que corresponde a una distribución impropia y conduce a resultados muy cercanos a los del enfoque anterior. Nótese que la distribución no informativa utilizada corresponde a la distribución Gamma con valores pequeños en ambos parámetros, puesto que la función de densidad de $Gamma(\alpha, \beta)$ se asemeja a la previa no informativa de Jeffreys: $p(\theta) \propto \theta^{-1}$ a medida que $\alpha \to 0$ y $\beta \to 0$.

De esta forma, la distribución posterior del parámetro de interés es $Gamma(69, 25998.5)$. Como es bien sabido, una estimación bayesiana para el parámetro θ está dada por la media de esta distribución posterior, la cual equivale a $69/25998.5 = 0.0026$. Ahora, como la esperanza de la distribución exponencial es $1/\theta$, entonces el tiempo promedio de sobrevivencia es de $1/0.0026 = 376.78$ días. El siguiente código computacional en JAGS puede ser usado para realizar inferencias sobre el parámetro θ, sobre el tiempo promedio y el tiempo mediano. De la misma forma, es posible obtener intervalos de credibilidad para estos parámetros.

```
Exp.model <- function(){
for(i in 1:n)
{
```

```
y[i] ~ dexp(theta)
}
theta ~ dgamma(0.1,0.1)
mean <- 1/theta
}

n <-69
y <- surv

Exp.data <- list("n", "y")
Exp.param <- c("theta", "mean")
Exp.inits <- function(){
list("theta"=0.5)
}

Exp.fit <- jags(data=Exp.data, inits=Exp.inits, Exp.param, n.iter=10000, n.burnin=1000, model.file=Ex

print(Exp.fit)
```

Después de diez mil iteraciones, los resultados de este código arrojan una estimación para θ de 0.003. Mientras que para el tiempo promedio de sobrevivencia $1/\theta$, se tiene una estimación puntual de 382.3 con un intervalo de credibilidad de (301.6, 482.9). La mediana se estimó en 378 días de sobrevivencia.

Suponga ahora que se va a realizar el transplante de corazón a 5 pacientes, y se quiere conocer el comportamiento probabilístico del tiempo promedio de sobrevida en estos 5 pacientes. Aplicando la distribución predictiva obtenida en el resultado 2.5.4, usando la distribución previa no informativa de Jeffrey, se tiene que

$$\begin{aligned} p(\bar{Y}^*|\mathbf{Y}) &= \frac{5\Gamma(5+69)}{\Gamma(5)\Gamma(69)} \frac{25998.5^{69}}{(5\bar{Y}^* + 25998.5)^{5+69}} (5\bar{Y}^*)^4 \\ &= \frac{1}{\bar{Y}^* Beta(5,69)} \left(\frac{25998.5}{5\bar{Y}^* + 25998.5}\right)^{69} \left(\frac{5\bar{Y}^*}{5\bar{Y}^* + 25998.5}\right)^5 \end{aligned}$$

El cálculo de esta función predictiva se puede llevar a cabo con el siguiente código en R, además de comprobar que la integral de la función es 1.

```
pred_exp<-function(x){
((s/(s+x*n.mono))^n)*((x*n.mono/(s+x*n.mono))^n.mono)/(x*beta(n,n.mono))
}

alfa<-beta<-0
s<-25998.5
n<-69; n.mono<-5
integrate(pred_exp, 0.0001, 10000)

## 1 with absolute error < 0.00000000032
```

La distribución predictiva de esta función se puede visualizar en la Figura 2.12, donde se puede ver que la mayor masa de la función se acumula alrededor del valor 260 días. También podemos calcular probabilidades de interés relacionadas con el promedio de viviencia de estos cinco pacientes, por ejemplo, la probabilidad de que en promedio sobrevivan más de 800 días es de 2.6 % usando el siguiente comando:

```
integrate(pred_exp,800,10000)

## 0.02645 with absolute error < 0.0000000026
```

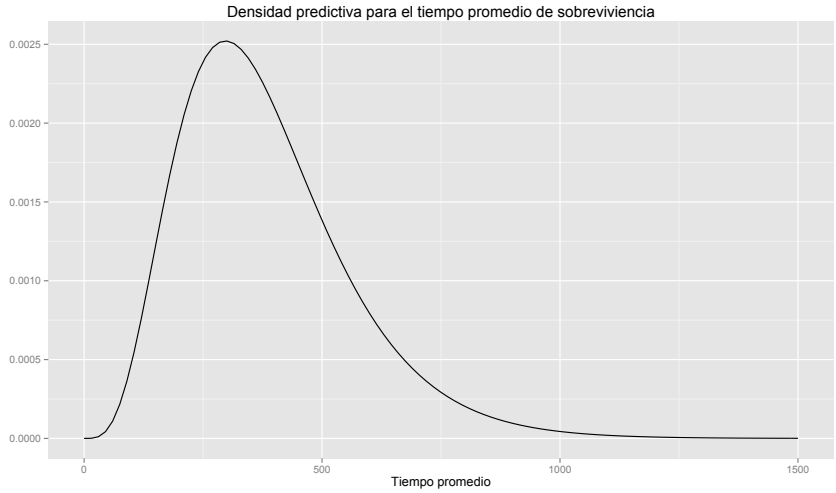


Figura 2.12: *Distribución predictiva posterior para el tiempo promedio de sobrevivencia de transplante de corazón.*

2.6 Modelo normal con media desconocida y varianza conocida

En esta dos últimas secciones de este capítulo, se considera datos que pueden ser descritos adecuadamente con la distribución normal, la cual se diferencia de las anteriores distribuciones consideradas pues tiene dos parámetros. En el siguiente capítulo se considera el caso general cuando ambos parámetros son desconocidos. En esta parte, se asume que la varianza teórica es conocica, y el objetivo es estimar la media teórica.

Suponga que $Y_1, \cdots, Y_n$ son variables independientes e idénticamente distribuidos con distribución $Normal(\theta, \sigma^2)$ con θ desconcido pero σ^2 conocido. De esta forma, la función de verosimilitud de los datos está dada por

$$\begin{aligned} p(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\} I_{\mathbb{R}}(y) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \end{aligned} \tag{2.6.1}$$

Como el parámetro θ puede tomar cualquier valor en los reales, es posible asignarle una distribución previa $\theta \sim Normal(\mu, \tau^2)$. Bajo este marco de referencia se tienen los siguientes resultados

Resultado 2.6.1. *La distribución posterior del parámetro de interés θ sigue una distribución*

$$\theta|\mathbf{Y} \sim Normal(\mu_n, \tau_n^2).$$

En donde

$$\mu_n = \frac{\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \tag{2.6.2}$$

Prueba.

$$\begin{aligned}
p(\theta \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta) p(\theta \mid \mu, \tau^2) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2 - \frac{1}{2\tau^2}(\theta - \mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}(y_i - \theta)^2}{\sigma^2} + \frac{(\theta - \mu)^2}{\tau^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2}{\sigma^2} - \frac{2\theta\sum_{i=1}^{n} y_i}{\sigma^2} + \frac{\theta^2}{\tau^2} - \frac{2\theta\mu}{\tau^2}\right]\right\} \\
&= \exp\left\{-\frac{\theta^2}{2}\left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right] + \theta\left[\frac{n\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2}\right]\right\} \\
&= \exp\left\{-\frac{\theta^2}{2\tau_n^2} + \frac{\theta\mu_n}{\tau_n^2}\right\} \\
&= \exp\left\{-\frac{1}{2\tau_n^2}(\theta^2 - 2\theta\mu_n)\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta^2 - 2\theta\mu_n + \mu_n^2)\right\} \\
&= \exp\left\{-\frac{1}{2\tau_n^2}(\theta - \mu_n)^2\right\}
\end{aligned}$$

Por lo tanto, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_n, \tau_n^2)$. ■

Observando la forma de μ_n, que corresponde a la estimación bayesiana del parámetro θ, podemos concluir que éste es una combinación convexa entre el estimador clásico de máxima verosimlitud $\hat{\theta}_C = \bar{y}$ y el estimador previo $\hat{\theta}_P = \mu$, puesto que:

$$\begin{aligned}
\hat{\theta}_B = \mu_n &= \frac{\frac{n}{\sigma^2}\bar{Y} + \frac{1}{\tau^2}\mu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \\
&= \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\bar{Y} + \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\mu \\
&= \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\hat{\theta}_C + \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\hat{\theta}_P
\end{aligned}$$

De donde se puede concluir que para una distribución previa fija, entre mayor sea el tamaño muestral n, más peso tendrá el estimador clásico $\hat{\theta}_C$ en el cálculo del estimador bayesiano. De la misma forma, para un conjunto fijo de datos $\mathbf{Y}$, entre menor sea la varianza previa, τ^2, más certeza tenemos sobre la información previa y por consiguiente la estimación bayesiana μ_n se acercará más a la estimación previa. En la Figura 2.13 se observa la función de densidad previa, función de verosimilitud y función de densidad posterior con $\mu = 5$, $\tau^2 = 0.01$, $\bar{y} = 2$, $\sigma^2 = 1$ y $n = 5, 10, 50, 200$. Podemos observar que a medida que el tamaño muestral n aumente, la función de verosimilitud (vista como la función del parámetro θ) se vuelve más concentrada alrededor del valor de $\bar{y}$, y a consecuencia, la función de

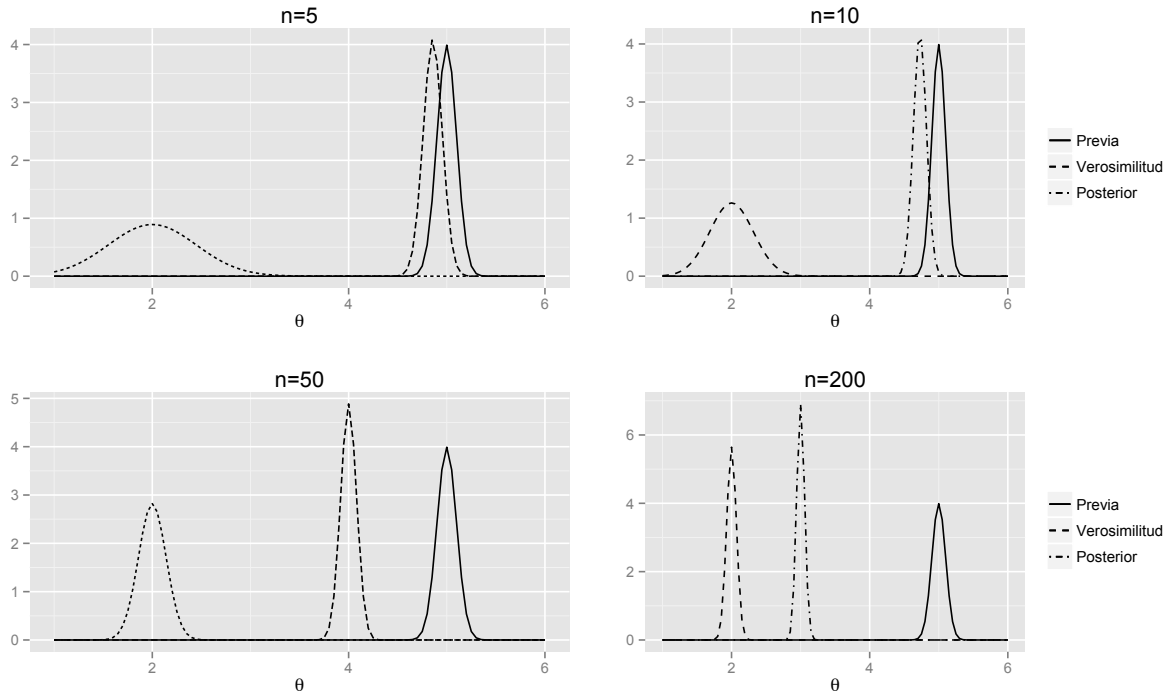


Figura 2.13: *Distribución previa, función de verosimilitud y distribución posterior del parámetro θ con $\mu = 5$, $\tau^2 = 0.01$, $\bar{y} = 2$, $\sigma^2 = 1$ y $n = 5, 10, 50, 200$.*

densidad posterior de θ se sitúa más cercana a la función de verosimilitud, y la estimación bayesiana se acerca más a la estimación clásica $\bar{y}$.

Distribución previa no informativa para θ

Por otro lado, nótese que en el caso en donde de antemano se desconozca el comportamiento estructural de θ, es posible hacer su distribución previa tan plana y vaga como sea posible. Para esto, basta con hacer tender al parámetro de precisión de la distribución previa hacia infinito. Es decir $\tau^2 \longrightarrow \infty$, en este caso, la distribución previa de θ corresponde a una distribución impropia, $p(\theta) \propto cte$. Esta idea intuitiva para representar la falta de la información previa corresponde a la previa no informativa de Jeffreys, puesto que la información de Fisher del parámetro θ en una variable con distribución normal está dada por

$$I(\theta) = 1/\sigma^2$$

De donde se puede concluir qua la previa no informativa de Jeffreys está dada por

$$p(\theta) \propto 1/\sigma \propto cte$$

Se puede ver que cuando se utiliza la anterior distribución previa para θ, la distribución posterior está dada por (Ejercicio 2)

$$\theta \mid \mathbf{Y} \sim Normal\left(\bar{y},\ \frac{\sigma^2}{n}\right)$$

Finalmente, comparemos los resultados inferenciales obtenidos con la previa no informativa de Jeffreys con el enfoque inferencial clásico en términos de la estimación puntua y el intervalo de credibilidad y de confianza.

- En cuanto a la estimación puntual, es claro que ambos enfoques conducen al mismo estimador $\hat{\theta} = \bar{Y}$.

- Con respecto al intervalo para el parámetro θ, al usar el enforque bayesiano con la previa no informativa de Jeffreys, un intervalo de credibilidad de $(1-\alpha)\times 100\,\%$ están dadas por los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución posterior de θ: $Normal(\bar{y}, \sigma^2/n)$, denotaremos estos percentiles como a y b, respectivamente. Por definición tenemos que, si $X \sim N(\bar{y}, \sigma^2/n)$,

$$\begin{aligned}\alpha/2 &= Pr(X < a)\\ &= Pr(\frac{X-\bar{y}}{\sigma/\sqrt{n}} < \frac{a-\bar{y}}{\sigma/\sqrt{n}})\\ &= Pr(Z < \frac{a-\bar{y}}{\sigma/\sqrt{n}})\end{aligned}$$

  Estos es, $\frac{a-\bar{y}}{\sigma/\sqrt{n}}$ es el percentil $\alpha/2$ de la distribución normal estándar $z_{\alpha/2}$ ó equivalentemente $-z_{1-\alpha/2}$. De esta forma, tenemos que $a = \bar{y} - z_{1-\alpha/2}\sigma/\sqrt{n}$. Análogamente tenemos que $b = bary + z_{1-\alpha/2}\sigma/\sqrt{n}$, y podemos concluir que un intervalo de crediblidad de $(1-\alpha)\times 100\,\%$ está dada por $\bar{y} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$, el cual coincide con el intervalo de confianza para θ usando el enfoque de la inferencia clásica (Zhang & Gutiérrez 2010).

Ahora revisamos las diferentes formas de hallar la distribución previa para θ. En primer lugar, consideramos el caso cuando la información previa se encuentra en un conjunto de datos $x_1, \cdots, x_m$ que corresponden a mediciones de la variable de estudio Y en otro punto de tiempo, en otro punto geográfico, o inclusive en otra población de estudio. En este caso, podemos tomar la media de la distribución previa μ como $\bar{X}$ y la varianza de la distribución previa τ^2 como S_X^2.

En el caso de que no se dispongan de datos como información previa, sino que ésta está contenida en alguna estimación que se haya realizado sobre θ. Por ejemplo, si se dispone de algún modelamiento estadístico que se haya hecho previamente sobre θ, podemos fácilmente obtener el valor estimado de θ y el error estándar de esta estimación, y naturalmente, estos dos valores serían nuestros parámetros de la distribución previa: μ y τ^2.

Finalmente, si la estimación previa de θ se presentada en forma de un intervalo, por ejemplo, si se sabe que un intervalo de confianza para θ es $(15.3,\ 24.7)$, entonces podemos usar μ como el punto medio de este intervalo, es decir, $\mu = 20$ y para escoger el valor de τ^2, se tiene en cuenta que en muchas ramas de la estadística, un intervalo de confianza se puede aproximar por $\hat{\theta} \pm 2\sqrt{var(\hat{\theta})}$. De esta forma, podemos usar $\tau^2 = \left(\frac{24.7-20}{2}\right)^2 \approx 5.5$

En cuanto a las distribuciones predictivas, a continuación se encuentran las expresiones correspondientes para una observación o una nueva muestra.

Resultado 2.6.2. *La distribución predictiva previa para una observación y es*

$$y \sim Normal(\mu, \tau^2 + \sigma^2)$$

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}p(Y) &= \int p(Y \mid \theta) p(\theta \mid \mu, \tau^2)\ d\theta\\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\} d\theta\end{aligned}$$

Berger (1985) desarrolló las siguientes igualdades

$$\begin{aligned}
&\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2}+\frac{(y-\theta)^2}{\sigma^2}\right]\\
&=\frac{1}{2}\left[\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\theta^2-2\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)\theta+\left(\frac{\mu^2}{\tau^2}+\frac{y^2}{\sigma^2}\right)\right]\\
&=\frac{1}{2\tau_1^2}\left[\theta^2-2\tau_1^2\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)\theta+\tau_1^4\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)^2\right]+\frac{1}{2}\left(\frac{\mu^2}{\tau^2}+\frac{y^2}{\sigma^2}\right)-\frac{\tau_1^2}{2}\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)^2\\
&=\frac{1}{2\tau_1^2}\left[\theta-\tau_1^2\left(\frac{\mu}{\tau^2}+\frac{y}{\sigma^2}\right)\right]^2+\frac{1}{2}\left[\left(\frac{1}{\sigma^2}-\frac{\tau_1^2}{\sigma^4}\right)y^2-2\frac{\mu\tau_1^2}{\tau^2\sigma^2}y+\left(\frac{\mu^2}{\tau^2}-\frac{\mu^2\tau_1^2}{\tau^4}\right)\right]\\
&=\frac{1}{2\tau_1^2}\left[\theta-\mu_1\right]^2+\frac{1}{2}\left[\frac{1}{\sigma^2+\tau^2}y^2-2\frac{\mu}{\sigma^2+\tau^2}y+\frac{\mu^2}{\sigma^2+\tau^2}\right]\\
&=\frac{1}{2\tau_1^2}\left[\theta-\mu_1\right]^2+\frac{1}{2(\sigma^2+\tau^2)}(y-\mu)^2.
\end{aligned}$$

Entonces

$$\begin{aligned}
p(Y)&=\int_{-\infty}^{\infty}\frac{1}{2\pi\sigma\tau}\exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right\}\exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}(y-\mu)^2\right\}d\theta\\
&=\frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\tau_1^2}}}\exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}(y-\mu)^2\right\}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\tau_1^2}}\exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right\}d\theta\\
&=\frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}}\exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}(y-\mu)^2\right\}
\end{aligned}$$

■

Una vez recolectados los datos $\mathbf{Y}=\{Y_1,\cdots,Y_n\}$, se obtiene la distribución predictiva posterior dada en el siguiente resultado. La demostración es similar al del resultado anterior, y se deja como ejercicio para los lectores.

Resultado 2.6.3. *La distribución predictiva posterior para una nueva observación $\tilde{y}$ es*

$$\tilde{y}\mid\mathbf{Y}\sim Normal(\mu_n,\tau_n^2+\sigma^2)$$

cuando se tiene una distribución previa informativa para θ.

Cuando se utiliza la distribución previa no informativa de Jeffreys, la distribución predictiva posterior para una nueva observación $\tilde{y}$ es

$$\tilde{y}\mid\mathbf{Y}\sim Normal\left(\bar{y},\left(1+\frac{1}{n}\right)\sigma^2\right)$$

Prueba. Se deja como ejercicio para los lectores. ■

En algunas situaciones, se quiere conocer el comportamiento probabilístico de más de una nueva observación, digamos $Y_1^*,\cdots,Y_{n^*}^*$, en este caso, lo ideal sería obtener la distribución conjunta predictiva posterior de la nueva muestra, $p(Y_1^*,\cdots,Y_{n^*}^*|\mathbf{Y})$. Sin embargo, esta distribución no es fácil de hallar, y procedemos a hallar la distribución predictiva posterior de la media de esta nueva muestra $\bar{Y}^*$, la cual es dada en el siguiente resultado.

Resultado 2.6.4. *La distribución predictiva posterior para la media muestral $\bar{Y}^*$ de una nueva muestra es*

$$\bar{Y}^*|\mathbf{Y} \sim N\left(\mu_n, \frac{\sigma^2}{n^*} + \tau_n^2\right)$$

cuando se tiene una previa informativa para θ, μ_n y τ_n^2 fueron definidos en (2.6.2).

Cuando se utiliza la distribución previa no informativa de Jeffreys para θ, la distribución predictiva posterior para la media muestral $\bar{Y}^*$ de una nueva muestra es

$$\bar{Y}^*|\mathbf{Y} \sim N\left(\bar{y}, \left(\frac{1}{n} + \frac{1}{n^*}\right)\sigma^2\right).$$

Prueba. Primero, consideramos cuando la distribución previa de θ es $N(\mu, \tau^2)$, tenemos que:

$$\begin{aligned}
p(\bar{Y}^*|\mathbf{Y}) &= \int_{-\infty}^{\infty} p(\bar{Y}^*|\theta)p(\theta|\mathbf{Y})\, d\theta \\
&= \int_{-\infty}^{\infty} (2\pi\frac{\sigma^2}{n^*})^{-1/2} \exp\left\{-\frac{n^*}{2\sigma^2}(\bar{y}^* - \theta)^2\right\} (2\pi\tau_n^2)^{-1/2} \exp\left\{-\frac{1}{2\tau_n^2}(\theta - \mu_n)^2\right\}\, d\theta \\
&= \int_{-\infty}^{\infty} (2\pi)^{-1} (\frac{\sigma^2}{n^*}\tau_n^2)^{-1/2} \exp\left\{-\frac{1}{2}\left[\frac{(\bar{y}^* - \theta)^2}{\sigma^2/n^*} + \frac{(\theta - \mu_n)^2}{\tau_n^2}\right]\right\}\, d\theta \\
&= \underbrace{\int_{-\infty}^{\infty} (2\pi\frac{1}{n^*/\sigma^2 + 1/\tau_n^2})^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{n^*}{\sigma^2} + \frac{1}{\tau_n^2}\right)\left(\theta - \frac{\bar{y}^*/(\sigma^2/n^*) + \mu_n/\tau_n^2}{n^*/\sigma^2 + 1/\tau_n^2}\right)^2\right\}\, d\theta}_{\text{igual a } 1} \\
&\quad (2\pi)^{-1/2} (\frac{\sigma^2}{n^*}\tau_n^2)^{-1/2} (\frac{n^*}{\sigma^2} + \frac{1}{\tau_n^2})^{-1/2} \exp\left\{-\frac{1}{2(\sigma^2/n^* + \tau_n^2)}(\bar{y}^* - \mu_n)^2\right\} \\
&= (2\pi)^{-1/2} (\frac{\sigma^2}{n^*}\tau_n^2)^{-1/2} (\frac{n^*}{\sigma^2} + \frac{1}{\tau_n^2})^{-1/2} \exp\left\{-\frac{1}{2(\sigma^2/n^* + \tau_n^2)}(\bar{y}^* - \mu_n)^2\right\} \\
&= (2\pi)^{-1/2} (\frac{\sigma^2}{n^*} + \tau_n^2)^{-1/2} \exp\left\{-\frac{1}{2(\sigma^2/n^* + \tau_n^2)}(\bar{y}^* - \mu_n)^2\right\}
\end{aligned}$$

Los desarrollos para cuando se utiliza la distribución previa no informativa de Jeffreys es similar, y se deja para los lectores (Ejercicio 4). ■

Del anterior resultado, podemos observar que: (1) la esperanza de la distribución de $\bar{Y}^*|\mathbf{Y}$ es igual a la esperanza de $\theta|\mathbf{Y}$, y (2) a diferencia de la varianza de $\theta|\mathbf{Y}$, la varianza de $\bar{Y}^*|\mathbf{Y}$ tiene un componente adicional: σ^2/n^*, de esta forma, tenemos tres fuentes de incertidumbre al momento de pronosticar $\bar{Y}^*$: la incertidumbre en la información previa, la incertidumbre en la muestra observada y la incertidumbre en la nueva muestra.

Ejemplo 2.6.1. En Zhang & Gutiérrez (2010, Ejemplo 2.3.6) reportan datos sobre el grosor de láminas de vidrio templado de 3cm para controlar la calidad de los vidrios producidos por una línea de producción con el fin de conocer el grosor real de las láminas producidas. Estos datos son: 3.56, 3.36, 2.99, 2.71, 3.31, 3.68, 2.78, 2.95, 2.82, 3.45, 3.42 y 3.15, con el promedio de 3.18cm. Suponga que por especificaciones técnicas, se conoce que la varianza del grosor es de 0.1cm^2. Por otro lado como información previa, se conoce que en la última inspección de calidad, se conoce que el grosor promedio fue de 2.8cm con una desviación estándar de 0.23cm.

De la anterior información, se puede decir que el parámetro de interés θ sería el grosor promedio de las láminas. También podemos afirmar que $\sigma^2 = 0.1cm^2$, $\bar{y} = 3.18cm$, $n = 12$, y los parámetros de

la distribución previa están dados por $\mu = 2.8cm$ y $\tau = 0.45cm$, de esta forma, podemos calcular los parámetros de la distribución posterior

$$\mu_n = \frac{\frac{12}{0.1} \times 3.18 + \frac{1}{0.23^2} \times 2.8}{\frac{12}{0.1} + \frac{1}{0.23^2}} = 3.13cm$$
$$\tau_n^2 = \left(\frac{12}{0.1} + \frac{1}{0.23^2}\right)^{-1} = 0.007cm^2$$
$$\tau^2 = \sqrt{0.007cm^2} = 0.084cm$$

Entonces la distribución posterior del grosor promedio es $N(\mu_n = 3.13cm,\ \tau_n^2 = 0.007cm^2)$. Podemos concluir que la estimación bayesiana del parámetro corresponde a $3.13cm$, mientras que para calcular un intervalo de credibilidad de 95 % para el parámetro de interés, se debe calcular los percentiles 2.5 % y 97.5 % de la distribución posterior de θ, y el intervalo de credibilidad queda dado por $(2.966cm,\ 3.293cm)$.

A continuación se ilustra el uso de `JAGS` para obtener la estimación bayesiana del parámetro θ.

```
Norm.model <- function(){
for(i in 1 : n)
{
y[i] ~ dnorm(theta, 1/0.1)
}
theta ~ dnorm(2.8, 1/(0.23^2) )
}

n <- 12
y <- c(3.56, 3.36, 2.99, 2.71, 3.31, 3.68, 2.78, 2.95, 2.82, 3.45, 3.42, 3.15)

Norm.data <- list("y","n")
Norm.param <- c("theta")
Norm.inits <- function(){
list("theta"=c(3.2))
}

Norm.fit <- jags(data=Norm.data, inits=Norm.inits, Norm.param, n.iter=10000,
n.burnin=1000, model.file=Norm.model)

## module glm loaded

## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 12
##    Unobserved stochastic nodes: 1
##    Total graph size: 45
##
## Initializing model

print(Norm.fit)
```

```
## Inference for Bugs model at "/var/folders/n7/01szs8_x7pq1bvpwwnvq7w_w0000gn/T//Rtmph2sfkH/model15e
##  3 chains, each with 10000 iterations (first 1000 discarded), n.thin = 9
##  n.sims = 3000 iterations saved
##          mu.vect sd.vect  2.5%   25%   50%   75%  97.5%  Rhat n.eff
## theta      3.132   0.085 2.965 3.073 3.132 3.190  3.297 1.001  3000
## deviance   7.331   1.583 6.170 6.285 6.710 7.762 11.824 1.001  3000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 1.3 and DIC = 8.6
## DIC is an estimate of expected predictive error (lower deviance is better).
```

De donde podemos ver que la estimación bayesiana de θ es $3.132cm$ con un error estándar de $0.085cm$, mientras que un intervalo de credibilidad del 95 % es $(2.965cm,\ 3.297cm)$, resultados muy similares a lo obtenido calculando directamente μ_n y τ_n.

Finalmente, ilustramos el uso de la distribución predictiva posterior. Suponga que la fábrica debe hacer un despacho de 8 láminas, y se quiere conocer sobre el grosor promedio del despacho $\bar{y}^*$. Usando el resultado 2.6.4, tenemos que la disribución de $\bar{Y}^*$ condicionado en los 12 datos observados está dado por

$$\bar{Y}^*|\mathbf{Y} \ \sim N\left(\mu_n,\ \frac{\sigma^2}{n^*}+\tau_n^2\right) = N\left(3.13cm,\ \frac{0.1}{8}+0.007\ cm^2\right) = N(3.13cm,\ 0.0195cm^2)$$

De esta forma, podemos afirmar que el grosor promedio del despacho es de 3.13cm con un intervalo de 95 % dado por los percentiles 2.5 % y 97.5 % de la anterior distribución: (2.85cm, 3.40cm). Nótese que el intervalo para $\bar{Y}^*$ es más ancho que el intervalo para θ, pues este tiene una varianza mayor a la varianza de la distribución posterior de θ.

También podemos calcular el intervalo para $\bar{Y}^*$ desde el punto de vista de simulación, simulando valores de θ desde su distribución posterior, y luego simulando valores de $\bar{Y}^*$ desde $p(\bar{Y}^* \mid \theta)$. Los siguientes códigos implemtan este enforque con 5000 iteraciones. Podemos ver que los resultados obtenidos son casi idénticos a los calculados anteriormente.

```
y.bar <- c()
mu.n <- 3.13; tau2.n <- 0.007; sigma2 <- 0.1
for(i in 1:5000){
  theta <- rnorm(1, mu.n, sqrt(tau2.n))
  y.bar[i] <- rnorm(1, theta, sqrt(sigma2/8))
}
mean(y.bar)

## [1] 3.127

quantile(y.bar, c(0.025,0.975))

##  2.5% 97.5%
## 2.853 3.405
```

2.7 Modelo normal con varianza desconocida y media conocida

En esta sección consideramos variables independientes e idénticamente distribuidas $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$. Asumimos que θ es conocida y el parámetro de interés es σ^2. En la práctica es inusual encontrar situaciones donde este supuesto se cumpla, sin embargo, los desarrollos presentados en esta sección serán útiles en el siguiente capítulo cuando se aborda la distribución normal con ambos parámetros desconocidos.

Para desarrollar la estimación bayesiana para σ^2, el primer paso es asignarle una distribución previa que sea acorde con los valores que toma el parámetro, y en lo posible inducir una distribución posterior conjugada. De esta forma, observamos que el parámetro σ^2 es estrictamente positivo, y la distribución previa debe tener soporte únicamente los valores positivos. Si bien la distribución que primero viene a mente sería la distribución Gamma, al observar la función de verosimilitud dada en 2.6.1, es claro que esta opción no inducirá una distribución posterior conjugado. Por lo tanto recurrimos a la distribución Inversa-Gamma que tiene la siguiente función de densidad:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left\{-\frac{\beta}{x}\right\} \tag{2.7.1}$$

para $x > 0$. $\alpha > 0$ es el parámetro de forma y $\beta > 0$ es el parámetro de escala, y usamos la notacion $X \sim Inversa - Gamma(\alpha, \beta)$. Se tiene que para esta distribución, $E(X) = \frac{\beta}{\alpha-1}$ cuando $\alpha > 1$ y $Var(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ cuando $\alpha > 2$.

En la literatura se acostumbra usar la siguiente distribución previa para el parámetro σ^2:

$$\sigma^2 \sim Inversa - Gamma(n_0/2,\ n_0\sigma_0^2/2)$$

La esperanza previa σ^2 viene dada por

$$E(\sigma^2) = \frac{\frac{n_0\sigma_0^2}{2}}{\frac{n_0}{2} - 1} = \frac{n_0\sigma_0^2}{n_0 - 2} \approx \sigma_0^2$$

y

$$Var(\sigma^2) = \frac{\left(\frac{n_0\sigma_0^2}{2}\right)^2}{\left(\frac{n_0}{2} - 1\right)\left(\frac{n_0}{2} - 2\right)} \approx \frac{2\sigma_0^4}{n_0}$$

De donde podemos escoger el parámetro σ_0^2 como el valor que se cree apropiado para σ^2 con base en la información previa. n_0 denota el número de datos en la información previa, el cual determina el grado de certidumbre del investigador sobre la información previa, pues entre mayor sea n_0, mayor cantidad de datos representa la distribución previa, la varianza previa de σ^2 se hace menor, lo cual representa menor incertidumbre en la distribución previa.

En la figura 2.14 podemos observar la forma de la distribución previa para σ^2 para diferentes valores de σ_0^2 y n_0. Es claro que la distribución está concentrada alrededor del valor de σ_0^2; y para un mismo valor de σ_0^2, entre mayor sea n_0 más concentrada está la función alrededor de σ_0^2.

Una vez definida la distribución previa de σ^2, procedemos a encontrar la distribución posterior.

Resultado 2.7.1. *La distribución posterior de* σ^2 *es*

$$\sigma^2 \mid \mathbf{Y} \sim Inversa - Gamma\left(\frac{n_0 + n}{2}, \frac{v_0}{2}\right)$$

En donde $v_0 = n_0\sigma_0^2 + n\hat{\sigma}_C^2$ *donde* $\hat{\sigma}_C^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \theta)^2$

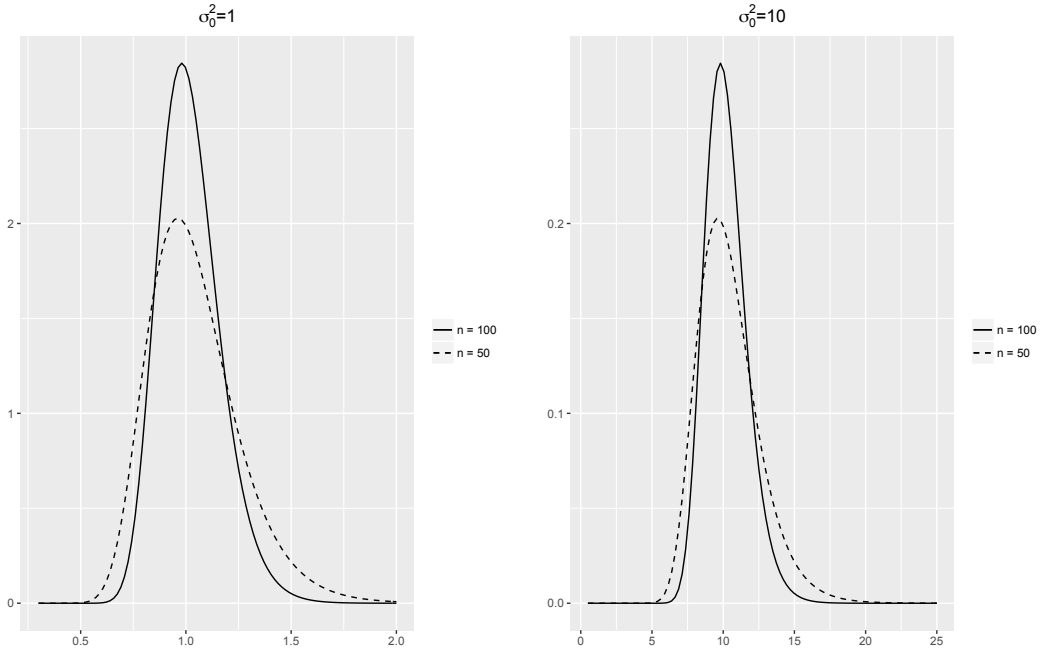


Figura 2.14: *Función de densidad Inversa-Gamma para diferentes valores de* σ_0^2 *y* n_0.

Prueba. Acudiendo a la distribución posterior conjunta e incorporando los términos que no dependen de θ en la constante de proporcionalidad, se tiene que

$$\begin{aligned} p(\sigma^2 \mid \mathbf{Y}) &\propto p(\sigma^2)p(\mathbf{Y} \mid \sigma^2) \\ &\propto (\sigma^2)^{-n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2}{2\sigma^2}\right\} (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \\ &= (\sigma^2)^{-(n_0+n)/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + n\hat{\sigma}_C^2\right]\right\} \end{aligned}$$

donde $\hat{\sigma}_C^2$ es el estimador clásico de σ^2 cuando θ es conocido, definido como $\hat{\sigma}_C^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$. De esta forma se encuentra la distribución posterior de σ^2. ■

De la distribución posterior de σ^2 se puede observar que la estimación bayesiana de σ^2 está dada por

$$\begin{aligned} \hat{\sigma}_B^2 &= \frac{\frac{v_0}{2}}{\frac{n_0+n}{2}} \\ &= \frac{n_0\sigma_0^2 + n\hat{\sigma}_C^2}{n_0+n-2} \\ &\approx \frac{n_0\sigma_0^2 + n\hat{\sigma}_C^2}{n_0+n} \\ &= \frac{n_0}{n_0+n}\sigma_0^2 + \frac{n}{n_0+n}\hat{\sigma}_C^2 \end{aligned}$$

Esto es, la estimación bayesiana viene siendo un promedio ponderado entre la estimación previa σ_0^2 y la estimación clásica $\hat{\sigma}_C^2$, y las ponderaciones depende directa y únicamente del número de datos

de las dos fuentes de información: n_0 y n. En la figura 2.15 se muestra la función de densidad de la distribución previa, distribución posterior y la función de verosimilitud vista como una función de σ^2 con $n_0 = 20$, $\sigma_0^2 = 10$, $\hat{\sigma}_C^2 = 50$, el tama??o muestral es $n = 5, 20, 50, 100$. Podemos ver que a medida que a medida que el tamaõ muestral n aumenta, la función de verosimilitud se concentra más alrededor del valor de $\hat{\sigma}_C^2$, y como consecuencia, la función de densidad posterior de σ^2 se acerca más a la función de verosimilitud, y la estimación bayesiana también se asemeja más a la estimación clásica $\hat{\sigma}_C^2$.

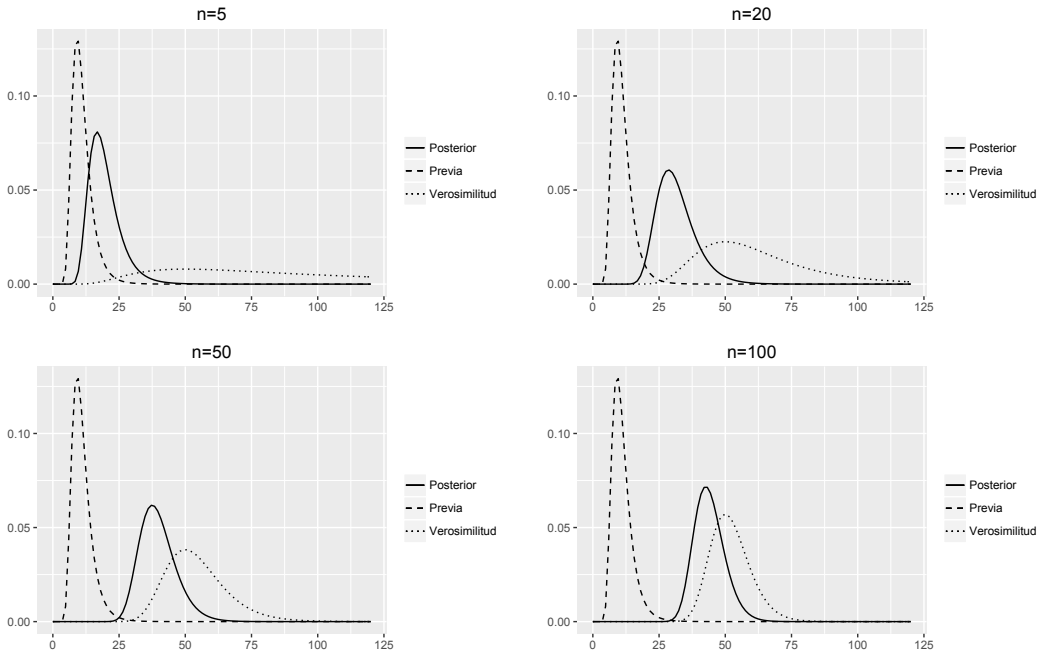


Figura 2.15: *Distribución previa, función de verosimilitud y distribución posterior de* σ^2 *con* $n_0 = 20$, $\sigma_0^2 = 10$, $\hat{\sigma}_C^2 = 50$ *y* $n = 5, 20, 50, 100$.

Distribución previa no informativa para σ^2

Recurriendo a las distribución previa no informativa de Jeffreys que establece que $p(\theta) \propto I(\theta)^{1/2}$, donde θ es el parámetro de interés, y $I(\theta)$ es la información de Fisher acerca del parámetro θ. Después de algunos cálculos se puede ver que (para más detalles, ver Zhang & Gutiérrez (2010, Sec.2.4)) $I(\sigma^2) = \frac{n}{2\sigma^4}$

De esta forma, la distribución previa no informativa de Jeffreys está dada por

$$p(\sigma^2) \propto \sigma^{-2}$$

para $\sigma^2 > 0$. En `JAGS` se acostumbra a usar valores pequeños de α y β para denotar esta distribución no informativa pues en este caso, la función de densidad (2.7.1) es similar a $p(\sigma^2) \propto \sigma^{-2}$.

Ahora, si bien esta distribución previa es una distribución impropia (pues $\int_0^\infty p(\sigma^2) = \infty$), al combinar con la función de verosimilitud, se puede concluir que la distribución posterior de σ^2 corresponde a una distribución Inversa-Gamma con el parámetro de forma $n/2$, y parámetro de escala $n\hat{\sigma}_C^2/2$ (se deja como ejercicio a los lectores). En la figura 2.16 se puede observar la función de densidad previa de Jeffreys (éste ha sido escalado para la óptima visualización), la función de verosimilitud con $n = 50$, $\hat{\sigma}_C^2 = 10$ y la función de densidad posterior resultante. Se puede ver claramente que la forma de la distribución posterior es muy similar a la función de verosimilitud, de donde se puede intuir que la estimación bayesiana resulta similar a la estimación clásica de σ^2. Efectivamente, podemos ver

que cuando se usa la distribución previa no informativa de Jeffreys, la estimación bayesiana de σ^2 corresponde a

$$\begin{aligned}\hat{\sigma}_B^2 &= \frac{\frac{n\hat{\sigma}_C^2}{2}}{\frac{n}{2}-1} \\ &= \frac{n}{n-2}\hat{\sigma}_C^2 \\ &\approx \hat{\sigma}_C^2\end{aligned}$$

y un intervalo de credibilidad del 95 % queda dado por los percentiles 2.5 % y 97.5 % de la distribución $Inversa-Gamma(\frac{n}{2}, \frac{n\hat{\sigma}_C^2}{2})$.

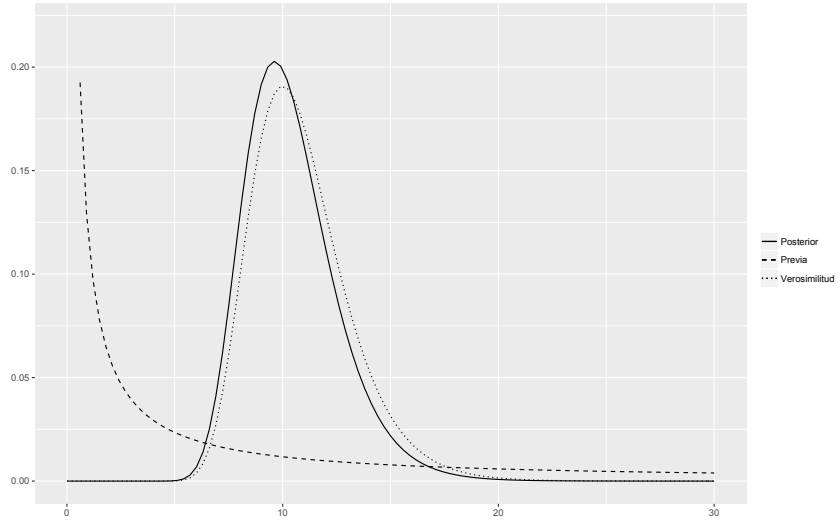


Figura 2.16: *Distribución previa no informativa de Jeffreys, función de verosimilitud y distribución posterior de* σ^2 *con* $n = 50$ *y* $\hat{\sigma}_C^2 = 10$.

Ahora, nos enfocamos en la distribución predictiva posterior para una nueva observación $\tilde{y}$ cuya expresión se muestra en el siguiente resultado.

Resultado 2.7.2. *La distribución predictiva posterior de una nueva observación* $\tilde{y}$*, cuando se utiliza una distribución previa informativa para* σ^2*, es la distribución* t *no estandarizado con grado de libertad* n_0+n*, el parámetro de localización* θ *y el parámetro de escala* $v_0/(n_0+n)$*, donde* $v_0 = n_0\sigma_0^2+\sum_{i=1}^n(y_i-\theta)^2$*, esto es,*

$$\tilde{y} \mid \mathbf{Y} \sim t_{n_0+n}\left(\theta, \frac{v_0}{n_0+n}\right).$$

Cuando se utiliza la distribución previa no informativa de Jeffreys para σ^2*, la distrubción predictiva de una nueva observación* $\tilde{y}$ *es* $t_n(\theta, \hat{\sigma}_C^2)$*, con* $\hat{\sigma}_C^2 = \frac{1}{n}\sum_{i=1}^n(y_i-\theta)^2$.

Prueba. Cuando la distribución previa para σ^2 es $Inversa-Gamma(n_0/2,\ n_0\sigma_0^2/2)$, la distribución

posterior de σ^2 es $Inversa-Gamma((n_0+n)/2,\ (n_0\sigma_0^2+n\hat{\sigma}_C^2)/2)$, tenemos que

$$\begin{aligned}
p(\tilde{y}\mid\mathbf{Y}) &= \int_0^\infty p(\tilde{y}\mid\sigma^2)p(\sigma^2\mid\mathbf{Y})\,d\sigma^2 \\
&= \int_0^\infty (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right\}\frac{(\frac{v_0}{2})^{(n_0+n)/2}(\sigma^2)^{-(n_0+n)/2-1}}{\Gamma(\frac{n_0+n}{2})}\exp\left\{-\frac{v_0}{2\sigma^2}\right\}\,d\sigma^2 \\
&= \frac{(2\pi)^{-1/2}(\frac{v_0}{2})^{(n_0+n)/2}}{\Gamma(\frac{n_0+n}{2})}\int_0^\infty (\sigma^2)^{-\frac{n_0+n+1}{2}-1}\exp\left\{-\frac{1}{\sigma^2}\left[\frac{v_0}{2}+\frac{1}{2}(\tilde{y}-\theta)^2\right]\right\}\,d\sigma^2 \\
&= \frac{(2\pi)^{-1/2}\Gamma(\frac{n_0+n+1}{2})}{\Gamma(\frac{n_0+n}{2})}\left(\frac{v_0}{2}\right)^{(n_0+n)/2}\left(\frac{v_0}{2}+\frac{1}{2}(\tilde{y}-\theta)^2\right)^{-\frac{n_0+n+1}{2}} \\
&= \frac{(2\pi)^{-1/2}\Gamma(\frac{n_0+n+1}{2})}{\Gamma(\frac{n_0+n}{2})}\left(\frac{v_0}{2}\right)^{(n_0+n)/2}\left(\frac{v_0}{2}\right)^{-(n_0+n+1)/2}\left(1+\frac{(\tilde{y}-\theta)^2}{v_0}\right)^{-(n_0+n+1)/2} \\
&= \frac{(2\pi)^{-1/2}\Gamma(\frac{n_0+n+1}{2})}{\Gamma(\frac{n_0+n}{2})}\left(\frac{v_0}{2}\right)^{-1/2}\left(1+\frac{(\tilde{y}-\theta)^2}{v_0}\right)^{-(n_0+n+1)/2} \\
&= \frac{(\pi v_0)^{-1/2}\Gamma(\frac{n_0+n+1}{2})}{\Gamma(\frac{n_0+n}{2})}\left(1+\frac{(\tilde{y}-\theta)^2}{v_0}\right)^{-(n_0+n+1)/2}
\end{aligned}$$

la cual corresponde a la función de densidad de la distribución $t_{n_0+n}\left(\theta,\frac{v_0}{n_0+n}\right)$. La distribución predictiva cuando se se usa la distribución previa de Jeffreys para σ^2 se deja como ejercicio (Ejercicio 6). ■

Del anterior resultado, tenemos que

$$\begin{aligned}
E(\tilde{Y}\mid\mathbf{Y}) &= \theta \\
Var(\tilde{Y}\mid\mathbf{Y}) &= \frac{v_0}{n_0+n}\frac{n_0+n}{n_0+n-2} = \frac{n_0\sigma_0^2+n\hat{\sigma}_C^2}{n_0+n-2}
\end{aligned}$$

Podemos ver que la esperanza de la nueva observación es θ, al igual que las variables de la muestra observada; mientras que la varianza de la nueva observación viene siendo aproximadamente un promedio ponderado entre la estimación previa de la varianza σ_0^2 y la estimación clásica $\hat{\sigma}_C^2$, los pesos dependen de los tamaños de las dos fuentes de información: n_0 y n.

Una forma alterna para aproximar el comportamiento aleatorio de $\tilde{Y}$ es usando la simulación. Se simulan valores de σ^2 desde su distribución posterior, y luego se simulan valores de $\tilde{Y}$ desde $p(\tilde{Y}\mid\sigma^2)$, es decir, de la distribución $Normal(\theta,\sigma^2)$. En la figura 2.7 se observa el histograma de 10 mil valores de $\tilde{Y}$ simulados con este procedimiento, en la misma gráfica se observa la función de densidad de la distribución predictiva de $\tilde{Y}$. Podemos que los valores obtenidos de esta forma concide plenamente con la distribución objetiva.

Cuando es de interés conocer el comportamiento de nuevas variables aleatorias $Y_1^*,\cdots,Y_{n^*}^*$, podemos obtener la distribución predictiva del valor promedio de estas nuevas medicicones: $\bar{Y}^*$. Esta distribución se encuenta en el siguiente resultado.

Resultado 2.7.3. *La distribución predictiva posterior de la media $\bar{Y}^*$ en una nueva muestrea $Y_1^*,\cdots,Y_{n^*}^*$, cuando se utiliza una distribución previa informativa para σ^2, es la distribución t no estandarizado con grado de libertad n_0+n, el parámetro de localización θ y el parámetro de escala $\frac{v_0}{n^*(n_0+n)}$, donde $v_0=n_0\sigma_0^2+\sum_{i=1}^n(y_i-\theta)^2$, esto es,*

$$\bar{Y}^*\mid\mathbf{Y}\sim t_{n_0+n}\left(\theta,\frac{v_0}{n^*(n_0+n)}\right).$$

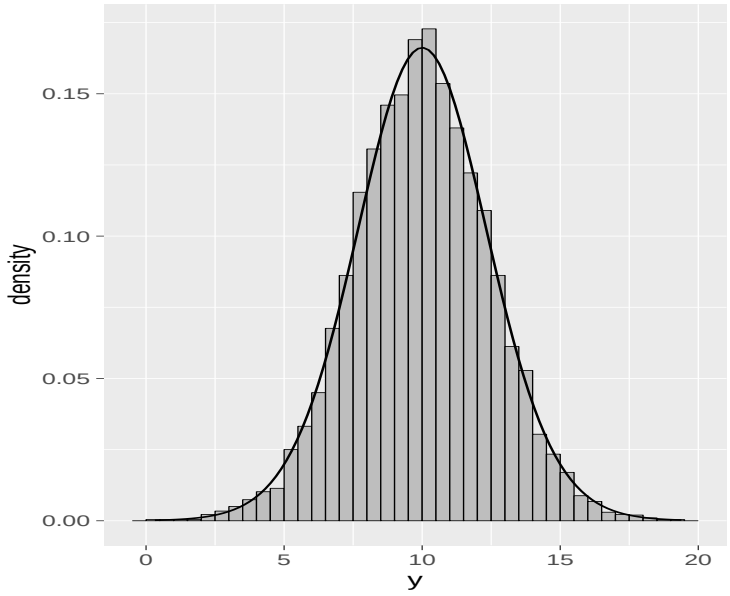


Figura 2.17: *Histograma de 10 mil valores simulados de $\tilde{y}$ y la función de densidad de la distribución predictiva* $t_{n_0+n}(\theta, \frac{v_0}{n_0+n})$.

Cuando se utiliza la distribución previa no informativa de Jeffreys para σ^2*, la distrubción predictiva de una nueva observación* $\bar{Y}^*$ *es*

$$\bar{Y}^* \mid \mathbf{Y} \sim t_n\left(\theta, \frac{\hat{\sigma}_C^2}{n^*}\right)$$

con $\hat{\sigma}_C^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$.

Prueba. La demostración del resultado es similar a la del resultado 2.7.2, y se deja como ejercicio para los lectores. ■

Ejemplo 2.7.1. Retomamos los datos usados en el ejemplo 2.6.1 donde se cuenta con las siguientes 12 mediciones a láminas de vidrios templados 3.56cm, 3.36cm, 2.99cm, 2.71cm, 3.31cm, 3.68cm, 2.78cm, 2.95cm, 2.82cm, 3.45cm, 3.42cm y 3.15cm. Los resultados inferenciales del ejemplo dejan entrever que el grosor promedio de las láminas de esta línea de producción se puede asumir de 3cm, y queremos conocer un poco acerca de la varianza σ^2 de esta línea de producción, ya que un valor grande de σ^2 no es deseable ya que los productos sería muy "disparejos" entre ellos, y representa una falla para el proceso industrial.

Al asumir la media conocida $\theta = 3cm$, se tiene que la estimación clásica de σ^2 viene dada por $\hat{\sigma}_C^2 = 0.13cm^2$. Al utilizar la distribución previa no informativa de Jeffreys, la estimación bayesiana de σ^2 viene dada por $\frac{n}{n-2}\hat{\sigma}_C^2 = 0.157cm^2$. Para calcular un intervalo de credibilidad de 95 %, podemos usar el siguiente código

```
library(pscl)
qigamma(0.025, alpha=12/2, beta=12*0.13/2)

## [1] 0.06685
```

```
qigamma(0.975, alpha=12/2, beta=12*0.13/2)

## [1] 0.3542
```

de donde el intervalo para σ^2 está dada por $(0.0668cm^2, 0.3542cm^2)$. Para facilitar la interpretación de la estimación, es preferible estimar la desviación estándar σ. La estimación bayesiana de este parámetro sería $\hat{\sigma}_B = \sqrt{0.157cm^2} \approx 0.4cm$, y un intervalo de credibilidad para σ viene dada por $(\sqrt{0.0668cm^2}, \sqrt{0.3542cm^2}) = (0.25cm, 0.59cm)$.

A continuación se ilustra el uso de `JAGS` para obtener la estimación bayesiana del parámetro σ^2. En `JAGS` no se admite el comando `dflat()` para especificar una previa no informativa como lo hace `WinBugs`, por cual se debe definir la distribución previa escogiendo los parámetros que representa la falta de información previa; por otro lado, se acostumbra a usar la distribución Gamma como la distribución previa del parámetro de precisión $\tau = 1/\sigma^2$, tal como se muestra a continuación.

```
IG.model <- function(){
for(i in 1 : n)
{
   y[i] ~ dnorm(3, tau)
}
  sigma <- 1/sqrt(tau)
  tau ~ dgamma(0.001, 0.001)
}

n <- 12
y <- c(3.56, 3.36, 2.99, 2.71, 3.31, 3.68, 2.78, 2.95, 2.82, 3.45, 3.42, 3.15)

IG.data <- list("y","n")
IG.param <- c("sigma")
IG.inits <- function(){
  list("tau"=c(1))
}

IG.fit <- jags(data=IG.data, inits=IG.inits, IG.param, n.iter=10000,
    n.burnin=1000, model.file=IG.model)

print(IG.fit)
```

De los anteriores códigos, se obtiene que la estimación bayesiana resultante para σ es de 0.388cm, con un intervalo de credibilidad del 95 % de $(0.258cm, 0.605cm)$, resultados muy similares a lo presentado anteriormente.

2.8 Ejercicios

1. En una muestra aleatoria de variables con distribución *Exponencial* con paramémetro θ, verifique que la distribución previa no informativa de Jeffreys está dada por

$$p(\theta) \propto \theta^{-1} \tag{2.8.1}$$

2. Desarrolla los cálculos necesarios para comprobar que para datos normales con varianza conocida, la distribución posterior de la media θ es $Normal(\bar{y}, \sigma^2/n)$ cuando la distribución previa es $p(\theta) \propto cte$.

3. Demuestre el resultado 2.6.3.

4. Demuestre la segunda parte del resultado 2.6.4.

5. Demuestre que en datos con distribución normal: $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$ con θ conocido, al utilizar la distribución previa no informativa de Jeffreys para σ^2, la distrubción de posterior de σ^2 viene dada por $Inversa - Gamma(\frac{n}{2}, \frac{n\hat{\sigma}_C^2}{2})$ con $\hat{\sigma}_C^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$.

6. Demuestre que en datos con distribución normal: $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$ con θ conocido, al utilizar la distribución previa no informativa de Jeffreys para σ^2, la distrubción predictiva de una nueva observación $\tilde{y}$ es $t_n(\theta, \hat{\sigma}_C^2)$, con $\hat{\sigma}_C^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$.

7. Demuestre el resultado 2.7.3.

Capítulo 3

Modelos multiparamétricos

En este capítulo, discutimos situaciones donde se requieren estimar simultáneamente más de un parámetro, es decir, los datos que enfrentamos se ajustan a una distrubución de probabilidad que involucre a multiples parámetros. Específicamente, se estudiará las siguientes distribuciones

- La distribución normal univariada que tiene dos parámetros: la media θ y la varianza σ^2,
- La distribución normal mutivariada con vector de medias $\boldsymbol{\theta}$ y la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, y
- La distribución multinomial cuyo parámetro constituye en el vector de probabilidades $\boldsymbol{\theta}$.

En el contexto de la estimación bayesiana, es necesario hallar la distribución posterior conjunta de estos parámetros, y encontrar la estimación por alguna de las siguientes dos formas: (1) hallar teóricamente la esperanza de la distribución posterior conjunta, ó (2) simular valores de la distribución posterior conjunta, de donde se puede obtener la estimación puntual y por intervalo.

3.1 Normal univariada con media y varianza desconocida

Supongamos que se dispone de realizaciones de un conjunto de variables independientes e idénticamente distribuidos $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$, cuando se desconoce tanto la media como la varianza de la distribución, es necesario plantear diversos enfoques y situarse en el más conveniente, según el contexto del problema. En términos de la asignación de las distribuciones previa para θ y σ^2 es posible:

- Suponer que la distribución previa $p(\theta)$ es independiente de la distribución previa $p(\sigma^2)$ y que ambas distribuciones son informativas.
- Suponer que la distribución previa $p(\theta)$ es independiente de la distribución previa $p(\sigma^2)$ y que ambas distribuciones son no informativas.
- Suponer que la distribución previa para θ depende de σ^2 y escribirla como $p(\theta \mid \sigma^2)$, mientras que la distribución previa de σ^2 no depende de θ y se puede escribir como $p(\sigma^2)$.

A continuación, analizamos cada uno de estos planteamientos, y desarrollamos los resultados necesrios para la estimación de θ y σ^2.

3.1.1 Parámetros independientes

El primer enfoque que considermos para el análisis de los parámetros de interés θ y σ^2 en una distribución normal univariada es suponer que las distribuciones previa de cada uno de los parámetros son independientes pero al mismo tiempo son informativas. Gelman, Carlin, Stern & Rubin (2003) afirma que este supuesto de independencia es atractivo en problemas para los cuales la información previa para θ no toma la forma de un número fijo de observaciones con varianza σ^2. Adicionalmente, este supuesto de independencia es coeherente con el hecho de que en la teoría clásica de estimación los estimadores insesgados de varianza mínima de θ y σ^2 son independientes (ver Zhang & Gutiérrez (2010, Sec.2.4)).

En este orden de ideas, y siguiendo la argumentación del capítulo anterior, la distribución previa para el parámetro θ es

$$\theta \sim Normal(\mu, \tau^2)$$

y la distribución previa para el parámetro σ^2 es

$$\sigma^2 \sim Inversa - Gamma(n_0/2, n_0\sigma_0^2/2)$$

Asumiendo independencia previa, la distribución previa conjunta está dada por

$$p(\theta, \sigma^2) \propto (\sigma^2)^{-n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\} \tag{3.1.1}$$

Una vez que se conoce la forma estructural de la distribución previa conjunta, es posible establecer la distribución posterior conjunta puesto que la verosimilitud de los datos, $p(\mathbf{Y} \mid \theta, \sigma^2)$, está dada por la expresión (2.6.1) y

$$p(\theta, \sigma^2 \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \theta, \sigma^2) p(\theta, \sigma^2)$$

Resultado 3.1.1. *La distribución posterior conjunta de los parámetros de interés está dada por*

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto (\sigma^2)^{-(n+n_0)/2-1} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2\right] - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \end{aligned} \tag{3.1.2}$$

Prueba. Tenemos que

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta, \sigma^2) p(\theta, \sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} (\sigma^2)^{-n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\ &= (\sigma^2)^{-(n+n_0)/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + \sum_{i=1}^{n}(y_i-\theta)^2\right] - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \\ &\propto (\sigma^2)^{-(n+n_0)/2-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2\right] - \frac{1}{2\tau^2}(\theta-\mu)^2\right\} \end{aligned}$$

dond la última expresión se obtiene al sumar y restar $\bar{y}$ dentro de $(y_i - \theta)^2$. ■

Nótese que la distribución posterior conjunta no tiene una forma estructural conocida y por lo tanto no es posible realizar el método de integración analítica para obtener una constante de integración (Migon & Gamerman 1999). Sin embargo, sí es posible obtener las distribuciones condicionales posterior de θ y de σ^2, notando que

$$p(\theta \mid \sigma^2, \mathbf{Y}) \propto p(\theta, \underbrace{\sigma^2}_{fijo} \mid \mathbf{Y}) \qquad \text{y} \qquad p(\sigma^2 \mid \theta, \mathbf{Y}) \propto p(\underbrace{\theta}_{fijo}, \sigma^2 \mid \mathbf{Y})$$

Es decir, para encontrar la distribución posterior marginal de θ dado σ^2, se utiliza la distribución posterior conjunta y los términos que no dependan de θ se incorporan en la constante de proporcionalidad. El mismo razonamiento se aplica para el parámetro σ^2.

Resultado 3.1.2. *La distribución posterior condicional de θ es*

$$\theta \mid \sigma^2, \mathbf{Y} \sim Normal(\mu_n, \tau_n^2) \tag{3.1.3}$$

En donde las expresiones para μ_n y τ_n^2 están dadas por 2.6.2. Por otro lado, la distribución posterior condicional de σ^2 es

$$\sigma^2 \mid \theta, \mathbf{Y} \sim Inversa - Gamma\left(\frac{n_0+n}{2}, \frac{v_0}{2}\right) \tag{3.1.4}$$

con $v_0 = n_0\sigma_0^2 + (n-1)S^2 + n(\bar{y}-\theta)^2$.

Prueba. Acudiendo a la distribución posterior conjunta e incorporando los términos que no dependen de θ en la constante de proporcionalidad, se tiene que

$$p(\theta \mid \sigma^2, \mathbf{Y}) \propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right\}$$

Completando los cuadrados y siguiendo el razonamiento de la demostración del resultado 2.6.1, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_n, \tau_n^2)$. Para la distribución posterior condicional de σ^2, consultar el resultado 2.7.1. ■

Una vez encontradas las distribuciones posteriores condicionales de θ y σ^2, se puede obtener la estimación de estos parámetros usando métodos de Monte Carlo, específicamente el muestreo de Gibbs, que puesto en el contexto de este capítulo, se resume en los siguientes pasos:

(1) Fijar un valor inicial para θ, lo denotamos por $\theta_{(1)}$

(2) Simular un valor de la distribución de $\sigma^2|\theta, \mathbf{Y}$ en (3.1.4) donde el parámetro v_0 que depende de θ, debe ser reemplazado por $\theta_{(1)}$ del paso anterior. Este valor simulado se denotará por $\sigma^2_{(1)}$

(3) Simlar un valor de la distribución de $\theta|\sigma^2, \mathbf{Y}$ en (3.1.3) donde en μ_n y τ_n^2 se debe reemplazar σ^2 por $\sigma^2_{(1)}$. Este valor simulado se denota por $\theta_{(2)}$.

(4) Se repite los pasos (2) y (3) hasta completar un número de iteraciones suficientes para alcanzar la convergencia en ambos parámetros

Después de ejecutar el muestreador de Gibbs, se eliminan los primeros valores simulados para descartar influencia del valor inicial y posiblemente se deba efectuar el *thinning* para eliminar correlaciones que pueden estar presentes. Posterior a eso, tenemos los valores finales simulados de θ y σ^2, de donde podemos calcular la estimación tomando los promedios respectivos, y calcular intervalos de credibilidad como los percentiles muestrales de los valores simulados.

En cuanto a la distribución predictiva para una nueva observación $\tilde{y}$, esta está dada por

$$p(\tilde{y} \mid \mathbf{Y}) = \int_0^\infty \int_{-\infty}^\infty p(\tilde{y} \mid \theta,\ \sigma^2) p(\theta,\ \sigma^2 \mid \mathbf{Y}) d\theta\ d\sigma^2$$

Hallar esta distribución de forma exacta no es fácil, y podemos optar por conocer el comportamiento probabilístico de $\tilde{y}$ por medio de la simulación. Tal como se explicó en el capítulo anterior, se debe simular en primer lugar valores de θ y de σ^2 de la distribución posterior $p(\theta,\ \sigma^2 \mid \mathbf{Y})$ usando el muestreador de Gibbs y posteriormente se simula valores de $\tilde{y}$ de la distribución $p(\tilde{y} \mid \theta,\ \sigma^2)$.

Ahora, si se quiere conocer el comportamiento de una nueva muestra aleatoria $Y_1^*, \cdots, Y_{n^*}^*$, lo podemos hacer por medio de la distribución predictiva de la media $\bar{Y}^*$ de la siguiente forma: se debe simular en primer lugar valores de θ y de σ^2 de la distribución posterior $p(\theta,\ \sigma^2 \mid \mathbf{Y})$ usando el muestreador de Gibbs y posteriormente se simula valores de $\bar{Y}^*$ de la distribución $N(\theta, \frac{\sigma^2}{n^*})$.

Ejemplo 3.1.1. Efron (2010) consideró un conjunto de datos que muestran la función renal de 157 individuos que se sometieron a una prueba médica exhaustiva en un hospital. Los resultados de la prueba renal están en un intervalo de -6 puntos a 4 puntos. Entre más alto sea el resultado, se concluye que el riñón del individuo es más sano. Nótese que estas pruebas son importantes para predecir el comportamiento de un riñón donado a un paciente con problemas renales. Los datos son extraídos de la siguiente página WEB (`http://statweb.stanford.edu/~ckirby/brad/LSI/datasets-and-programs/datasets.html`) y para este ejemplo sólo se utilizaron los primeros 15 datos del archivo.

En principio, es de interés para el investigador conocer la media y la dispersión de estos datos, para poder analizar a fondo la situación de los pacientes que esperan un transplante.

Dado que se trata de una primera aproximación, se prefiere utilizar distribuciones previas no informativas para los parámetros de la media y varianza. Lo anterior se logra en `JAGS` definiendo las distribuciones previas de `mu ~ dnorm(0,0.001)` y de `tau ~ dgamma(0.001,0.001)` donde `tau` corresponde al parámetro de precisión que resulta ser el inverso de la varianza σ^2. De esta forma, la distribución previa de *mu* está centrada en cero, pero con una varianza muy grande al igual que la distribución de la varianza, los cuales representan distribuciones previas no informativas.

El siguiente código en `JAGS` muestra cómo se lleva a cabo la inferencia.

```
Model <- function(){
  for (i in 1:n)
  {
    y[i] ~ dnorm(theta,tau)
  }
  theta ~ dnorm(0,0.001);
  sigma <- 1/sqrt(tau)
  tau ~ dgamma(0.001, 0.001)
}

n <- 15
y <- c(1.69045085, -1.41076082, -0.27909483, -0.91387987, 3.21868429, -1.47282460,
       -0.96524353, -2.45084934, 1.03838153, 1.79928679, 0.97826621, 0.67463830,
       -1.08665864, -0.00509027, 0.43708128)

Model.data <- list("y","n")
Model.param <- c("theta", "sigma")
Model.inits <- function(){
  list("theta"=c(0), "tau"=c(1))
}

Model.fit <- jags(data=Model.data, inits=Model.inits, Model.param, n.iter=10000,
                  n.burnin=1000, model.file=Model)

print(Model.fit)
```

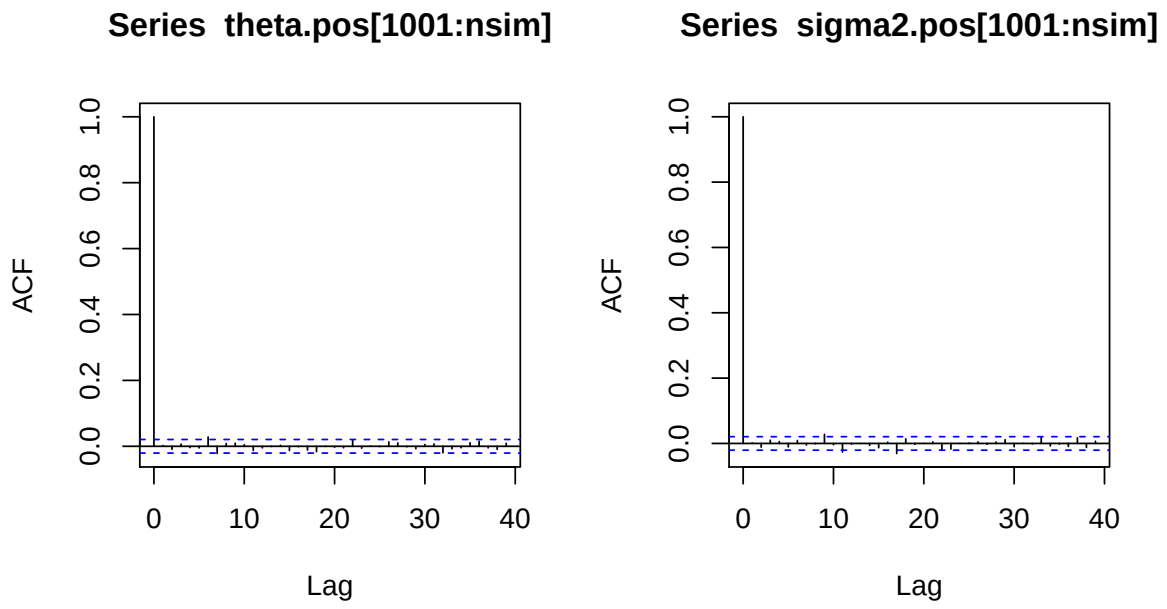
Después de ejecutar diez mil iteraciones, la salida del anterior código muestra una estimación puntual para la esperanza de Y de 0.087 con un intervalo de credibilidad del 95 % dado por (-0.75, 0.95).

Por otro lado, la estimación puntual de la desviación estándar de Y es de 1.557 con un intervalo de credibilidad del 95 % dado por (1.12, 2.39).

A continuación se ilustra el uso de `R` el algoritmo de Gibbs para los datos del ejemplo. Se recalca que se utiliza la librería `MCMCpack` (Martin, Quinn & Park 2011) para generar las realizaciones de la distribución Inversa-Gamma.

```
set.seed(123456)
library(MCMCpack)

## Loading required package: MASS
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
## ## Copyright (C) 2003-2018 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
## ##
## ## Support provided by the U.S. National Science Foundation
## ## (Grants SES-0350646 and SES-0350613)
## ##

y <- c(1.69045085, -1.41076082, -0.27909483, -0.91387987, 3.21868429, -1.47282460,
       -0.96524353, -2.45084934, 1.03838153, 1.79928679, 0.97826621, 0.67463830,
       -1.08665864, -0.00509027, 0.43708128)

n <- length(y)

#parametros previos de theta
mu <- 0; tau2 <- 1000
#parametros previos de sigma2
a <- 0.001; b <- 0.001

nsim <- 10000
theta.pos <- rep(NA,nsim)
sigma2.pos <- rep(NA,nsim)

# Valor inicial de theta
theta.pos[1] <- 0

#parametros posteriores de sigma2
a.n <- a+n/2
b.n <- b+((n-1)*var(y)+n*(mean(y)-theta.pos[1]))/2
#simulacion de la distribucion posterior condicional de theta
sigma2.pos[1] <- rinvgamma(1, a.n, b.n)
########################
# Muestreador de Gibbs #
########################
for(i in 2:nsim){
  #parametros posteriores de theta
  tau2.n <- 1 / ((n/sigma2.pos[i-1])+(1/tau2))
  mu.n <- tau2.n * (mean(y) * (n/sigma2.pos[i-1])+mu/tau2)
  #simulacion de la distribucion posterior condicional de theta
  theta.pos[i] <- rnorm(1, mean=mu.n, sd=sqrt(tau2.n))
```

```
  #parametros posteriores de sigma2
  a.n <- a + n/2
  b.n <- b + ((n-1) * var(y) + n * (mean(y)-theta.pos[i])) / 2
  #simulacion de la distribucion posterior condicional de theta
  sigma2.pos[i] <- rinvgamma(1, a.n, b.n)
}
par(mfrow=c(1,2))
acf(theta.pos[1001:nsim])
acf(sigma2.pos[1001:nsim])
```



Al observar que no existen correlaciones importantes en los valores simulados de θ y σ^2 (después de descartar las primeras 1000 iteraciones), se concluye que se puede utilizar directamente estos valores para la obtención de las estimaciones. Por cual, se calcula el promedio y los percentiles muestrales de los valores simulados de la siguiente forma:

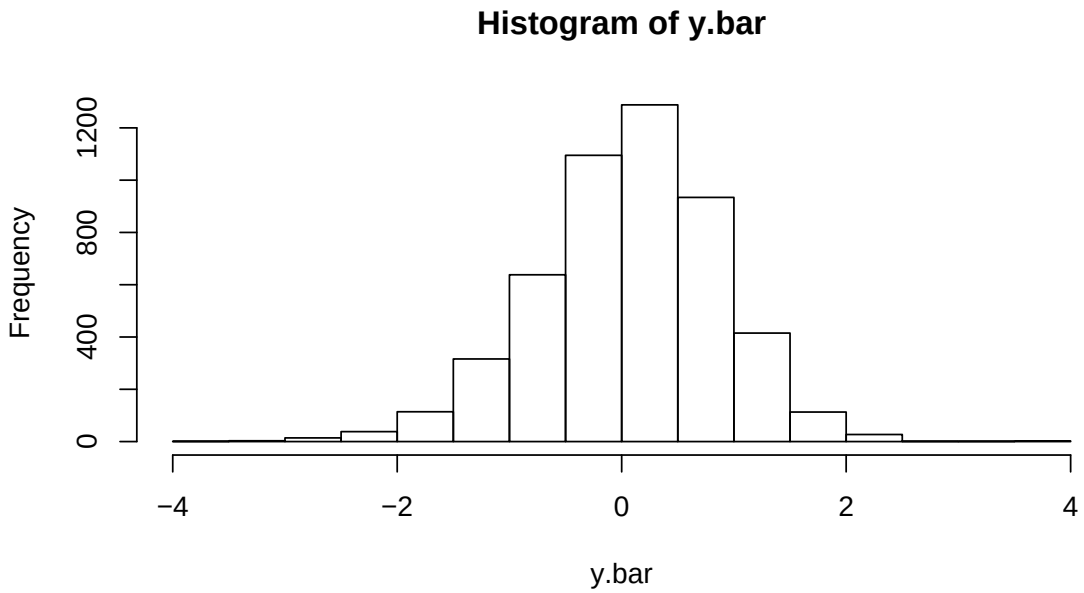
```
mean(theta.pos[1001:nsim])

## [1] 0.08

quantile(theta.pos[1001:nsim], c(0.025,0.975))

##  2.5%   98%
## -0.72  0.90

mean(sqrt(sigma2.pos[1001:nsim]))

## [1] 1.5

quantile(sqrt(sigma2.pos[1001:nsim]), c(0.025,0.975))
```

```
## 2.5%  98%
##  1.0  2.3
```

De donde podemos concluir que una estimación puntual para la esperanza de Y de 0.08 con un intervalo de credibilidad del 95 % dado por (-0.73, 0.89). Por otro lado, la estimación puntual de la desviación estándar de Y es de 1.5 con un intervalo de credibilidad del 95 % dado por (1.0, 2.3), resultados muy similares a lo obtenido don `JAGS`.

Finalmente ilustramos la forma de obtener la distribución predictiva para el promedio muestral de 5 nuevos pacientes.

```
n.ast <- 5; y.bar <- c()
for(i in 1:(nsim/2)){
  y.bar[i] <- rnorm(1,theta.pos[i+nsim/2],sqrt(sigma2.pos[i+nsim/2]/n.ast))
}
hist(y.bar)
```




```
mean(y.bar)

## [1] 0.068

sd(y.bar)

## [1] 0.81

quantile(y.bar,c(0.025,0.975))

## 2.5%  98%
## -1.6  1.6
```

Podemos ver que se espera que el promedio de las pruebas en 5 nuevos pacientes es de 0.068, con un intervalo del 95 % de (-1.6, 1.6), este intervalo es mucho más ancho que el de θ, pues naturalmente $\bar{Y}$ tiene mayor incertidumbre que los parámetros del modelo, y en segundo lugar, el tamaño de nuevos datos es de 5 que es bastante pequeño, lo cual hace que el pronóstico para $\bar{Y}^*$ no sea muy preciso.

3.1.2 Parámetros dependientes

En algunas situaciones es muy útil asumir una distribución previa conjugada, y para lograr eso no es posible establecer que los parámetros tengan distribuciones previa independientes. Bajo esta situación, la inferencia posterior de los parámetros de interés debe ser llevada a cabo en dos etapas: En la primera, se debe establecer la distribución previa conjunta para ambos parámetros siguiendo la sencilla regla que afirma que

$$p(\theta, \sigma^2) = p(\sigma^2)p(\theta \mid \sigma^2)$$

En la segunda etapa ya es posible analizar propiamente cada uno de los parámetros de interés siguiendo otra sencilla regla que afirma que

$$p(\theta, \sigma^2 \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \theta, \sigma^2)p(\theta, \sigma^2)$$

La anterior formulación conlleva a asignar una distribución previa para θ dependiente del parámetro σ^2. Esto quiere decir que en la distribución $p(\theta \mid \sigma^2)$, el valor de σ^2 se considera una constante fija y conocida, esta distribución previa está dada por[1]

$$p(\theta \mid \sigma^2) \sim Normal(\mu, \sigma^2/c_0)$$

donde c_0 es una constante. Por otro lado, y siguiendo los argumentos de la sección 2.7, una posible opción para la distribución previa de σ^2, que no depende de θ, corresponde a

$$p(\sigma^2) \sim Inversa - Gamma(n_0/2, n_0\sigma_0^2/2)$$

De esta forma, podemos encontrar la distribución conjunta previa de θ y σ^2 como sigue:

Resultado 3.1.3. *La distribución conjunta previa de los parámetros θ y σ^2 está dada por una distribución*

$$\theta, \sigma^2 \sim Normal - Inversa - Gamma\left(\mu, c_0, \frac{n_0+1}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

Prueba.

$$\begin{aligned} p(\theta, \sigma^2) &= p(\sigma^2)p(\theta \mid \sigma^2) \\ &\propto (\sigma^2)^{-\frac{n_0}{2}-1} \exp\left\{-\frac{n_0\sigma_0^2}{2\sigma^2}\right\} (\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{c_0}{2\sigma^2}(\theta-\mu)^2\right\} \\ &= (\sigma^2)^{-\frac{n_0+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + c_0(\theta-\mu)^2\right]\right\} \end{aligned}$$

la cual corresponde a la forma de la función de densidad de la distribución Normal-Inversa-Gamma. ■

Una vez encontrada la distribución conjunta previa, procedemos a encontrar la distribución conjunta posterior, y así poder encontrar las estimaciones de θ y σ^2.

[1] La forma como la distribución previa de θ dependa de σ^2 es coherente con la información de Fisher sobre θ que es igual a σ^{-2}.

Resultado 3.1.4. *La distribución posterior conjunta de los parámetros θ y σ^2 está dada por*

$$\theta, \sigma^2 \mid \mathbf{Y} \sim Normal - Inversa - Gamma\left(\mu_n, c_0 + n, \frac{n_0 + n + 1}{2}, \beta\right).$$

con

$$\beta = \frac{1}{2}\left(n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0 + n}(\mu - \bar{y})^2\right)$$

y

$$\mu_n = \frac{\frac{n}{\sigma^2}\bar{Y} + \frac{c_0}{\sigma^2}\mu}{\frac{n}{\sigma^2} + \frac{c_0}{\sigma^2}} = \frac{n\bar{Y} + c_0\mu}{n + c_0}$$

Prueba. En primer lugar, recordamos que la función de verosimilitud de la muestra está dada por

$$p(\mathbf{Y} \mid \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y} - \theta)^2\right]\right\} \tag{3.1.5}$$

Por otro lado, se tiene que

$$\begin{aligned} p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \theta, \sigma^2)p(\theta, \sigma^2) \\ &\propto (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + c_0(\theta - \mu)^2 + (n-1)S^2 + n(\bar{y} - \theta)^2\right]\right\} \\ &= (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \qquad (3.1.6) \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + (c_0 + n)(\theta - \mu_n)^2 + \frac{c_0 n}{c_0 + n}(\mu - \bar{y})^2\right]\right\} \qquad (3.1.7) \end{aligned}$$

puesto que

$$c_0(\theta - \mu)^2 + n(\bar{y} - \theta)^2 = (c_0 + n)(\theta - \mu_n)^2 + \frac{c_0 n}{c_0 + n}(\mu - \bar{y})^2$$

■

Para encontrar las distribuciones marginales posterior de cada uno de los parámetros se procede de la siguiente forma:

1. Para hallar la distribución posterior condicional de θ, dada por $P(\theta \mid \sigma^2, \mathbf{Y})$, se debe considerar que σ^2 es una constante fija y conocida tal como se consideró al principio de esta sección. Basado en lo anterior, es posible utilizar la siguiente regla de probabilidad

   $$P(\theta \mid \sigma^2, \mathbf{Y}) = \frac{p(\theta, \sigma^2 \mid \mathbf{Y})}{p(\sigma^2, \mathbf{Y})} p(\mathbf{Y}) \propto p(\theta, \sigma^2 \mid \mathbf{Y})$$

   Lo anterior sugiere que la distribución marginal posterior de θ, $p(\theta \mid \sigma^2, \mathbf{Y})$, se encuentra utilizando la distribución posterior conjunta, $p(\theta, \sigma^2 \mid \mathbf{Y})$, suponiendo que todas las expresiones que involucren al valor σ^2 se pueden incluir en la constante de proporcionalidad

2. Dado que σ^2 no depende de ningún otro parámetro entonces, utilizando la distribución posterior conjunta, es posible encontrar su distribución marginal posterior de la siguiente forma

   $$p(\sigma^2 \mid \mathbf{Y}) = \int p(\theta, \sigma^2 \mid \mathbf{Y}) d\theta$$

   Lo propio es posible hacer con θ, utilizando la distribución posterior conjunta, es posible encontrar su distribución marginal posterior de la siguiente forma

   $$p(\theta \mid \mathbf{Y}) = \int p(\theta, \sigma^2 \mid \mathbf{Y}) d\sigma^2$$

Resultado 3.1.5. *La distribución posterior de θ condicional a $\sigma^2, \mathbf{Y}$ está dada por*

$$\theta \mid \sigma^2, \mathbf{Y} \sim Normal(\mu_n, \sigma^2/(n+c_0))$$

con $\mu_n = \dfrac{n\bar{y} + c_0\mu}{n+c_0}$.

Prueba. Acudiendo a la distribución posterior conjunta dada en (3.1.6), tenemos que

$$\begin{aligned} p(\theta \mid \sigma^2, \mathbf{Y}) &\propto p(\theta, \sigma^2 \mid \mathbf{Y}) \\ &\propto (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \\ &\qquad \times \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + (c_0+n)(\theta-\mu_n)^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(c_0+n)(\theta-\mu_n)^2\right\} \end{aligned}$$

la cual corresponde a la forma de la función de densidad de la distribución $Normal(\mu_n, \sigma^2/(n+c_0))$. ■

En el anterior resultado, la media de la distribución condicional posterior μ_n se puede escribir como $\mu_n = \frac{n}{n+c_0}\bar{y} + \frac{c_0}{n+c_0}\mu$, promedio ponderado entre la estimación clásica $\bar{y}$ y la estimación previa μ. Observando la forma que toman los pesos $\frac{n}{n+c_0}$ y $\frac{c_0}{n+c_0}$, se puede pensar a c_0 como el número de observaciones en la información previa, y así, los pesos de la estimación clásica y la estimación previa dependen directamente de los tamaños muestrales respectivos.

Resultado 3.1.6. *La distribución marginal posterior del parámetro σ^2 es*

$$\sigma^2 \mid \mathbf{Y} \sim Inversa - Gamma\left(\frac{n+n_0}{2}, \frac{(n+n_0)\sigma_n^2}{2}\right)$$

Donde $(n+n_0)\sigma_n^2 = n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2$ corresponde a una suma ponderada de la varianza previa, la varianza muestral y la diferencia entre la media muestral y la media previa.

Prueba. De la distribución posterior conjunta (3.1.6) e integrando con respecto a θ, se tiene que

$$\begin{aligned} p(\sigma^2 \mid \mathbf{Y}) &= \int p(\theta, \sigma^2 \mid \mathbf{Y})\, d\theta \\ &\propto (\sigma^2)^{-\frac{n_0+n+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \\ &\qquad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{n+c_0}{2\sigma^2}(\theta-\mu_n)^2\right\}\, d\theta \\ &\propto (\sigma^2)^{-\frac{n_0+n}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2\right]\right\} \\ &\qquad \times \int_{-\infty}^{\infty} \frac{\sqrt{n+c_0}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{n+c_0}{2\sigma^2}(\theta-\mu_n)^2\right\}\, d\theta \\ &\propto (\sigma^2)^{-\frac{n_0+n}{2}-1} \exp\left\{-\frac{(n+n_0)\sigma_n^2}{2\sigma^2}\right\} \end{aligned}$$

la cual corresponde a la forma de la función de densidad de la distribución $Inversa{-}Gamma(\frac{n+n_0}{2}, \frac{(n+n_0)\sigma_n^2}{2})$.
■

Dadas las distribuciones $p(\sigma^2 \mid \mathbf{Y})$ y $p(\theta \mid \sigma^2, \mathbf{Y})$, podemos proceder de la siguiente forma para obetener valores simulados de θ y σ^2 y así, obtener las estimaciones. Si el número de iteraciones se fija como G, entonces se procede a:

(1) Simular G valores de la distribución de $\sigma^2|\mathbf{Y}$, es decir, de la distribución $Inversa - Gamma$ encontrada en el anterior resultado, estos valores se denotan por $\sigma^2_{(1)}, \sigma^2_{(2)}, \cdots, \sigma^2_{(G)}$.

(2) Para cada valor de $\sigma^2_{(g)}$, con $g = 1, \cdots, G$, simlar un valor de la distribución de $\theta|\sigma^2, \mathbf{Y}$, es decir, de la distribución $N(\mu_n, \sigma^2/(n + c_0))$, donde σ^2 se reemplaza por $\sigma^2_{(g)}$. De esta forma, se obtiene los valores $\theta_{(1)}, \theta_{(2)}, \cdots, \theta_{(G)}$.

Es claro que en el anterior algoritmo, no es necesario fijar algún valor inicial para θ o para σ^2, así como tampoco induce correlaciones entre los valores simulados para ningún parámetro. Por lo tanto, se puede usar directamente estos valores para el cálculo de las estimación, y no es necesario descartar los primeros valores simulados, ni realizar el *thinning*.

Ahora bien, existe otra alternativa para obtener la estimación de θ y σ^2: encontrando directamente la distribución posterior de cada parámetro. La distribución posterior de σ^2 ya se encontró en el resultado 3.1.6, resta encontrar la distribución posterior de θ, la cual se presenta en el siguiente resultado.

Resultado 3.1.7. *La distribución posterior del parámetro θ es la distribución t no estandarizado con grado de libertad n_0+n, el parámetro de localización $\mu_n = \dfrac{n\bar{Y} + c_0\mu}{n + c_0}$ y el parámetro de escala $\dfrac{\sigma_n}{\sqrt{c_0 + n}}$ con $(n + n_0)\sigma_n^2 = n_0\sigma_0^2 + (n - 1)S^2 + \frac{c_0 n}{c_0+n}(\mu - \bar{y})^2$. Esto es,*

$$\theta \mid \mathbf{Y} \sim t_{n+n_0}\left(\mu_n, \frac{\sigma_n^2}{c_0 + n}\right)$$

Prueba. Partiendo de la distribución posterior conjunta e integrando con respecto a σ^2, se tiene que

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &= \int_0^\infty p(\theta, \sigma^2 \mid \mathbf{Y})\ d\sigma^2 \\ &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n_0+n+1}{2}+1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n_0 + n)\sigma_n^2 + (c_0 + n)(\theta - \mu_n)^2\right]\right\}\ d\sigma^2 \end{aligned}$$

Haciendo un cambio de variable tal que

$$z = \frac{A}{2\sigma^2}, \qquad \text{donde} \quad A = (n_0 + n)\sigma_n^2 + (c_0 + n)(\theta - \mu_n)^2$$

por tanto

$$d\sigma^2 = -\frac{A}{2z^2}\ dz$$

Entonces, volviendo a la integral en cuestión, se tiene que

$$\begin{aligned}
p(\theta \mid \mathbf{Y}) &\propto \left(\frac{1}{A}\right)^{\frac{n_0+n+1}{2}+1} \int_{\infty}^{0} \frac{-A}{2z^2}(2z)^{\frac{n_0+n+1}{2}+1} e^{-z}\, dz \\
&\propto A^{-\frac{n_0+n+1}{2}} \underbrace{\int_0^{\infty} z^{\frac{n_0+n+1}{2}-1} e^{-z}\, dz}_{Gamma\left(\frac{n_0+n+1}{2},1\right)} \\
&\propto A^{-\frac{n_0+n+1}{2}} \\
&= \left[(n_0+n)\sigma_n^2 + (c_0+n)(\theta-\mu_n)^2\right]^{-\frac{n_0+n+1}{2}} \\
&\propto \left[1 + \frac{(c_0+n)(\theta-\mu_n)^2}{(n_0+n)\sigma_n^2}\right]^{-\frac{n_0+n+1}{2}} \\
&= \left[1 + \frac{1}{n_0+n}\left(\frac{\theta-\mu_n}{\sigma_n/\sqrt{c_0+n}}\right)^2\right]^{-\frac{n_0+n+1}{2}}
\end{aligned}$$

la cual corresponde a la forma de la función de densidad de la distribución deseada. ■

Las distribuciones encontradas en los resultados 3.1.6 y 3.1.7, permite estimar directamente los parámetros θ y σ^2 usando las esperanzas teóricas de las distribuciones posteriores. Esto es, las estimaciones puntuales son:

$$\hat{\theta} = \mu_n = \frac{n\bar{Y} + c_0\mu}{n+n_0}$$

$$\hat{\sigma}^2 = \frac{(n+n_0)\sigma_n^2/2}{(n+n_0)/2-1} = \frac{(n+n_0)\sigma_n^2}{n+n_0-2} \approx \sigma_n^2 = \frac{n_0\sigma_0^2 + (n-1)S^2 + \frac{c_0 n}{c_0+n}(\mu-\bar{y})^2}{n+n_0}$$

Los intervalos de credibilidad de θ y σ^2 de $(1-\alpha)\times 100\,\%$ se construyen usando los percentiles $\alpha/2$ y $1-\alpha/2$ de las respectivas distribuciones posteriores dadas en los resultados mencionados anteriormente.

Ilustramos el uso de la metodología en el siguiente ejemplo.

Ejemplo 3.1.2. Para los datos de función renal (Efron 2010) que se muestran en el Ejemplo 3.1.1, suponga que la información previa está contenida en la medición de función renal en una muestra de 12 pacientes dadas por: -1.3619, -1.1116, -0.4744, -0.5663, 2.2056, 0.9491, 0.2298, -0.7933, 1.0198, -0.9850, 3.5679 y -1.9504. La media y la varianza muestral de estas 12 observaciones corresponden a 0.060775 y 2.598512, así, se toma $\mu = 0.060775$, $\sigma_0^2 = 2.598512$ y $c_0 = n_0 = 12$.

Por otro lado, la media y la varianza muestral de los 15 pacientes en la información actual son $\bar{y} = 0.08349249$ y $S^2 = 2.301684$. De esta forma, los parámetros de las distribuciones marginales posterior de θ y σ^2 se pueden calcular como $\mu_n = \frac{15}{15+12} \times 0.08349249 + \frac{12}{15+12} \times 0.060775 = 0.07339583$ y

$$\sigma_n^2 = \frac{12 * 2.598512 + 14 * 2.301684 + 6.666667 * (0.060775 - 0.08349249)^2}{15+12} = 2.348487$$

En conclusión, las distribuciones marginales posterior de θ y σ^2 están dadas por

$$\theta|\mathbf{Y} \sim t_{27}(0.07339583, 2.348487/27 = 0.086981)$$

y

$$\sigma^2|\mathbf{Y} \sim Inversa - Gamma(27/2 = 13.5,\ 27 * 2.348487/2 = 31.70457)$$

Así, la estimación Bayesiana de θ es $\mu_n = 0.073$ y un intervalo de credibilidad de 95 % para θ se puede calcular como $0.073 \pm t_{27,0.975} * \sqrt{0.086981} = (-0.53,\ 0.68)$. Por otro lado, la estimación Bayesiana de σ^2 está dada por $31.70457/(13.5 - 1) = 2.53$, y un intervalo de credibilidad de 95 % para σ^2 se puede calcular como los percentiles 2.5 % y 97.5 % de la distribución $IG(13.5,\ 31.70457)$, dado por $(1.5, 4.4)$.

Los anteriores cálculos se ilustran en el siguiente código `R`.

```
library(pscl)
# Datos de la informacion previa
x <- c(-1.3619, -1.1116, -0.4744, -0.5663, 2.2056, 0.9491, 0.2298, -0.7933, 1.0198,
          -0.9850, 3.5679, -1.9504)
# Datos de la informacion actual
y <- c(1.69045085, -1.41076082, -0.27909483, -0.91387987, 3.21868429, -1.47282460,
          -0.96524353, -2.45084934, 1.03838153, 1.79928679, 0.97826621, 0.67463830,
          -1.08665864, -0.00509027, 0.43708128)
# Paramatros de la distribucion previa
n0 <- c0 <- 12
mu <- mean(x); sigma2_0 <- var(x)
# Informacion
n <- length(y)
bar.y <- mean(y); S2 <- var(y)
# Algunos paramatros de la distribucion posterior
mu.n <- (n*bar.y + c0*mu)/(n+n0)
sigma2_n <- (n0*sigma2_0+(n-1)*S2+c0*n*(mu-bar.y)^2/(c0+n))/(n+n0)
# Estimacion puntual
theta.hat <- mu.n; sigma2.hat <- (n+n0)*sigma2_n/(n+n0-2)
theta.hat

## [1] 0.073

sigma2.hat

## [1] 2.5

# Intervalo de credibilidad de 95% para theta
mu.n + qt(c(0.025,0.975), df=n+n0)*sqrt(sigma2_n/(n+n0))

## [1] -0.53  0.68

# Intervalo de credibilidad de 95% para sigma2
qigamma(0.025, alpha=(n+n0)/2, beta=(n+n0)*sigma2_n/2)

## [1] 1.5

qigamma(0.975, alpha=(n+n0)/2, beta=(n+n0)*sigma2_n/2)

## [1] 4.4
```

Otra forma de estimar los parámetros θ y σ^2 es utilizando los métodos de Monte Carlos tal como lo expone anteriormente, simulando primero los valores de σ^2 y posteriormente los valores de θ.

```
n.sim <- 20000
sigma2.res <- rinvgamma(n.sim, (n+n0)/2, (n+n0)*sigma2_n/2)
theta.res <- c()
for(i in 1:n.sim){
  theta.res[i] <- rnorm(1,mu.n, sqrt(sigma2.res[i]/(n+c0)))
}
# Visualiza los valores simulados
## par(mfrow=c(1,2))
## ts.plot(theta.res)
## ts.plot(sigma2.res)
## acf(theta.res)
## acf(sigma2.res)
# Estimaciones puntuales
mean(theta.res)

## [1] 0.073

mean(sigma2.res)

## [1] 2.5

# Intervalos de credibilidad del 95%
quantile(theta.res, c(0.025,0.975))

##  2.5%   98%
## -0.53  0.67

quantile(sigma2.res, c(0.025,0.975))

## 2.5%  98%
##  1.5  4.4
```

De las gráficas que arrojan los anteriores códigos, se puede comprobar que (1) no hay necesidad de descartar los primeros valores obtenidos, y (2) los valores simulados no incorrelacionados entre ellos, por lo cual podemos calcular las estimaciones directamente usando los 20 mil valores simulados.

La estimación e intervalo de credibilidad para θ calculada desde los valores simulados corresponden a 0.073 y (-0.53, 0.67); mientras que para σ^2, están dadas por 2.5 y (1.5, 4.4). Podemos ver que los resultados obtenidos en los dos enfoques son muy similares.

3.1.3 Parámetros no informativos

En esta sección consideramos el tratamiento cuando no tenemos información previa disponible. Suponga que $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ corresponde a una muestra de variables aleatorias con distribución $Normal(\theta, \sigma^2)$. Luego, la función de distribución conjunta o verosimilitud está dada por 2.6.1

$$p(\mathbf{Y} \mid \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}$$

En primer lugar suponga que los parámetros tienen distribuciones previa independientes y en esta primera etapa se realizará el análisis suponiendo que estas distribuciones son no informativas. Lo

anterior implica que la distribución previa conjunta de los parámetros de interés está dada por

$$p(\theta, \sigma^2) = p(\theta)p(\sigma^2) \tag{3.1.8}$$

Como la distribución previa de θ es normal, es fácil verificar que ésta empieza a tener las características propias de una distribución no informativa cuando la varianza de la misma se vuelve muy grande, sin importar el valor de la media. Cuando esto sucede, la forma de la distribución previa de θ se torna plana y es lógico pensar que puede ser acercada mediante una distribución constante, tal que

$$p(\theta) \propto cte$$

Por otro lado, Gelman, Carlin, Stern & Rubin (2003) afirma que la distribución Inversa-Gamma, la cual es la distribución previa para el parámetro σ^2, se vuelve no informativa cuando los hiper-parámetros toman valores muy cercanos a cero. De esta forma haciendo tender $\alpha \longrightarrow 0$ y $\beta \longrightarrow 0$, entonces la distribución previa de σ^2 se convierte en

$$p(\sigma^2) \propto \sigma^{-2}$$

la cual coincide con la distribución previa no informativa de Jeffreys discutida en la sección 2.7. Por lo anterior, la distribución previa no informativa conjunta estaría dada por

$$p(\theta, \sigma^2) \propto \sigma^{-2} \tag{3.1.9}$$

Bajo este marco de referencia se tiene el siguiente resultado sobre la distribución posterior de θ

Resultado 3.1.8. *La distribución posterior del parámetro θ sigue una distribución t no estandarizado con grado de libertad $n-1$, el parámetro de localización $\bar{Y}$ y el parámetro de escala $\frac{S^2}{n}$, esto es,*

$$\theta \mid \mathbf{Y} \sim t_{n-1}\left(\bar{y}, \frac{S^2}{n}\right).$$

Donde $(n-1)S^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Esta distribución también puede expresarse como

$$\frac{\theta - \bar{y}}{S/\sqrt{n}} \mid \mathbf{Y} \sim t_{n-1}$$

donde t_{n-1} denota la distribución t estandarizado con grado de libertad $n-1$.

Prueba. En primer lugar nótese que la distribución posterior conjunta de los parámetros de interés es

$$\begin{aligned}
p(\theta, \sigma^2 \mid \mathbf{Y}) &\propto p(\theta, \sigma^2)p(\mathbf{Y} \mid \theta, \sigma^2) \\
&\propto \frac{1}{\sigma^2}\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \theta)^2\right]\right\} \\
&= \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y} - \theta)^2\right]\right\}
\end{aligned} \tag{3.1.10}$$

Ahora, para hallar la distribución marginal posterior de θ es necesario integrar la anterior expresión con respecto a σ^2. Con esto, se tiene que

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &= \int_0^\infty p(\theta, \sigma^2 \mid \mathbf{Y})\ d\sigma^2 \\ &\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\}\ d\sigma^2 \end{aligned}$$

Haciendo un cambio de variable tal que

$$z = \frac{A}{2\sigma^2}, \qquad \text{donde} \quad A = (n-1)S^2 + n(\bar{y}-\theta)^2$$

por tanto

$$d\sigma^2 = -\frac{A}{2z^2}\ dz$$

Entonces, volviendo a la integral en cuestión, se tiene que

$$\begin{aligned} p(\theta \mid \mathbf{Y}) &\propto \left(\frac{1}{A}\right)^{n/2+1} \int_\infty^0 \frac{-A}{2z^2}(2z)^{n/2+1} e^{-z}\ dz \\ &\propto A^{-n/2} \underbrace{\int_0^\infty z^{n/2-1} e^{-z}\ dz}_{Gamma(n/2)} \\ &\propto A^{-n/2} \\ &= [(n-1)S^2 + n(\bar{y}-\theta)^2]^{-n/2} \\ &\propto \left[1 + \frac{n(\bar{y}-\theta)^2}{(n-1)S^2}\right]^{-n/2} = \left[1 + \frac{1}{n-1}\left(\frac{\bar{y}-\theta}{S/\sqrt{n}}\right)^2\right]^{-\frac{(n-1)+1}{2}} \end{aligned}$$

la cual corresponde a la función de densidad de distribución de una variable aleatoria con distribución $t_{n-1}(\bar{y}, S^2/n)$. ■

Resultado 3.1.9. *La distribución posterior del parámetro σ^2 sigue una distribución*

$$\sigma^2 \mid \mathbf{Y} \sim Inversa - Gamma((n-1)/2, (n-1)S^2/2).$$

Prueba. Utilizando el mismo argumento del anterior resultado, se tiene que

$$\begin{aligned} p(\sigma^2 \mid \mathbf{Y}) &= \int_{-\infty}^\infty p(\theta, \sigma^2 \mid \mathbf{Y})\ d\theta \\ &\propto \int_{-\infty}^\infty \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\}\ d\theta \\ &= \left(\frac{1}{\sigma^2}\right)^{n/2+1} \sqrt{2\pi\sigma^2/n} \exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\} \underbrace{\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\}\ d\theta}_{\text{vale 1}} \\ &\propto (\sigma^2)^{-n/2-1/2} \exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\} \\ &= (\sigma^2)^{-\frac{n-1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}(n-1)S^2\right\} \end{aligned}$$

la cual corresponde a la función de densidad de la distribución $Inversa - Gamma((n-1)/2, (n-1)S^2/2)$. ■

De los resultados 3.1.8 y 3.1.9, podemos ver que cuando no se dispone de información previa, la estimación bayesiana de θ y σ^2 están dadas por

$$\hat{\theta}_B = E(\theta \mid \mathbf{Y}) = \bar{Y}$$
$$\hat{\sigma}^2_B = E(\sigma^2 \mid \mathbf{Y}) = \frac{(n-1)S^2/2}{(n-1)/2-1} = \frac{n-1}{n-3}S^2 \approx S^2$$

Podemos concluir que la estimación bayesiana de θ cuando no hay información previa es idéntica a la estimación clásica de θ, mientas que la de σ^2 es muy similar a la estimación clásica.

En cuanto a la estimación por intervalo de credibilidad, podemos ver que un intervalo de crediblidad de $(1-\alpha)\times 100\,\%$ está dado por los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución $t_{n-1}\left(\bar{Y}, \frac{S^2}{n}\right)$, se puede ver que estos corresponden a $\bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$ y $\bar{Y} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}$. En conclusión, un intervalo de credibildad para θ está dado por $\bar{Y} \pm t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}$, el cual es idéntico al intervalo de confianza para θ en la estadística clásica.

En cuanto al intervalo de crediblidad para σ^2, este está dado por los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución $Inversa-Gamma((n-1)/2,\ (n-1)S^2/2)$. En la estadística clásica, el intervalo de confianza para σ^2 está dada por

$$IC(\sigma^2) = \left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}},\ \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right)$$

Aunque la forma de estos dos intervalos son muy diferentes, resultan ser idénticos. A continuación mostramos el porqué. Suponga que a es el percentil $\alpha/2$ de la distribución $Inversa-Gamma((n-1)/2,\ (n-1)S^2/2)$, esto es, si $X \sim Inversa-Gamma((n-1)/2,\ (n-1)S^2/2)$, entonces $Pr(X<a) = \alpha/2$. Ahora por propiedades de la distribución $Inversa-Gamma$, se tiene que $\dfrac{X}{(n-1)S^2} \sim Inversa-Gamma(\frac{n-1}{2},\ \frac{1}{2})$. Por la relación entre la distribución $Gamma$ y la distribución $Inversa-Gamma$, tenemos que $\dfrac{(n-1)S^2}{X} \sim Gamma(\frac{n-1}{2},\ 2)$, es decir, $\dfrac{(n-1)S^2}{X} \sim \chi^2_{n-1}$, de donde tenemos que

$$\begin{aligned}\frac{\alpha}{2} &= Pr(X<a)\\ &= Pr\left(\frac{(n-1)S^2}{X} > \frac{(n-1)S^2}{a}\right)\end{aligned}$$

Esto es, $\dfrac{(n-1)S^2}{a}$ es el percentil $1-\alpha/2$ de la distribución χ^2_{n-1}, esto es, $\dfrac{(n-1)S^2}{a} = \chi^2_{n-1,1-\alpha/2}$, de donde $a = \dfrac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}$, así concluimos que el límite inferior del intervalo de credibildad coincide con el límite inferior del intervalo de confianza. Análogamente se puede ver que también los límites superiores coinciden, y así vemos que el intervalo para σ^2 coincide en la estadística clásica y la estadística bayesiana sin información previa.

Enfoque alterno para estimar θ y σ^2

Existe otra forma de obtener las estimaciones para el parámetro θ, recordando la expresión 3.1.10, podemos afirmar que

$$\theta \mid \sigma^2, \mathbf{Y} \sim Normal(\bar{y}, \sigma^2/n)$$

puesto que

$$\begin{aligned} p(\theta \mid \sigma^2, \mathbf{Y}) &\propto p(\theta, \sigma^2 \mid \mathbf{Y}) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{y}-\theta)^2\right]\right\} \\ &= \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\} \end{aligned}$$

la cual corresponde a la función de densidad de la distribución $Normal(\bar{y}, \sigma^2/n)$. De esta forma, usando las distribución $p(\sigma^2 \mid \mathbf{Y})$ y $p(\theta \mid \sigma^2, \mathbf{Y})$, podemos implementar el siguiente procedimiento para obtener valores simulados de θ y σ^2:

Si el número de iteraciones se fija como G, entonces se procede a:

(1) Simular G valores de la distribución de $\sigma^2|\mathbf{Y}$, es decir, de la distribución $Inversa-Gamma((n-1)/2, (n-1)S^2/2)$, estos valores se denotan por $\sigma^2_{(1)}, \sigma^2_{(2)}, \cdots, \sigma^2_{(G)}$.

(2) Para cada valor de $\sigma^2_{(g)}$, con $g = 1, \cdots, G$, simlar un valor de la distribución de $\theta|\sigma^2, \mathbf{Y}$, es decir, de la distribución $N(\bar{y}, \sigma^2/n)$, donde σ^2 se reemplaza por $\sigma^2_{(g)}$. De sta forma, se obtiene los valores $\theta_{(1)}, \theta_{(2)}, \cdots, \theta_{(G)}$.

Las estimaciones de θ y σ^2 se pueden obteer de los valores obtenidos $\theta_{(1)}, \theta_{(2)}, \cdots, \theta_{(G)}$ y $\sigma^2_{(1)}, \sigma^2_{(2)}, \cdots, \sigma^2_{(G)}$.

Distribución predictiva

La distribución predictiva para una nueva observación $\tilde{Y}$ está dada por

$$\begin{aligned} p(\tilde{y} \mid \mathbf{Y}) &= \int\int p(\tilde{y} \mid \theta, \sigma^2) p(\theta, \sigma^2 \mid \mathbf{Y})\ d\theta\ d\sigma^2 \\ &= \int\int p(\tilde{y} \mid \theta, \sigma^2) p(\theta \mid \sigma^2, \mathbf{Y}) p(\sigma^2 \mid \mathbf{Y})\ d\theta\ d\sigma^2 \\ &= \int \left(\int p(\tilde{y} \mid \theta, \sigma^2) p(\theta \mid \sigma^2, \mathbf{Y})\ d\theta\right) p(\sigma^2 \mid \mathbf{Y})\ d\sigma^2 \end{aligned}$$

En la integral dentro del paréntesis, el parámetro σ^2 permanece fijo, por lo cual, dicha integral es la misma a la del resultado 2.6.3, y corresponde a la distribución $N\left(\bar{y}, \left(1+\frac{1}{n}\right)\sigma^2\right)$. De esta forma, combinando con la distribución posterior de σ^2, tenemos que

$$\begin{aligned} &p(\tilde{y} \mid \mathbf{Y}) \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi(1+\frac{1}{n})\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2(1+\frac{1}{n})}(\tilde{y}-\bar{y})^2\right\} \frac{\left(\frac{(n-1)S^2}{2}\right)^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)} (\sigma^2)^{-\frac{n-1}{2}-1} \exp\left\{-\frac{(n-1)S^2}{2\sigma^2}\right\}\ d\sigma^2 \end{aligned}$$

Después de realizar las manipulaciones algebráicas necesarias, se encuentra que

$$p(\tilde{y} \mid \mathbf{Y}) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \frac{1}{\sqrt{\pi(n-1)}} \left(\left(1+\frac{1}{n}\right)S^2\right)^{-1/2} \left(1 + \frac{1}{n-1}\frac{(\tilde{y}-\bar{y})^2}{\left(1+\frac{1}{n}\right)S^2}\right)^{-n/2} \tag{3.1.11}$$

la cual corresponde a la distribución t no estandarizado con grado de libertad $n-1$, el parámetro de localización $\bar{y}$ y el parámetro de escala $(1+\frac{1}{n})S^2$. De esta forma, podemos ver que los dos primeros

momentos de esta distribución están dados por

$$E(\tilde{Y} \mid \mathbf{Y}) = \bar{y}$$
$$Var(\tilde{Y} \mid \mathbf{Y}) = \frac{n-1}{n-3}\left(1+\frac{1}{n}\right)S^2 = \frac{(n-1)(n+1)}{n(n-3)}S^2$$

Otra manera equivalente de conocer el comportamiento probabilístico de $\tilde{y}$ es por medio de la simulación. Se debe simular en primer lugar valores de θ y de σ^2 de la distribución posterior $p(\theta,\ \sigma^2 \mid \mathbf{Y})$ usando el muestreador de Gibbs y posteriormente se simula valores de $\tilde{y}$ de la distribución $p(\tilde{y} \mid \theta,\ \sigma^2)$. En la figura 3.1.3 se muestran el histograma de 10 mil valores de $\tilde{Y}$ simulados de esta forma, donde los datos muestrales corresponden a 20 datos simulados de la distribución $N(12, 3^2)$. En la misma gráfica se observa también la función de densidad de la distribución t, podemos ver que los valores de simulados de $\tilde{Y}$ efectivamente coincide con la distribución predictiva de $\tilde{Y}$. Por lo anterior, se puede calcular un predictor de $\tilde{Y}$ como el promedio de los 10 mil valores simulados, y calcular el intervalo de predicción usando los percentiles de estos 10 mil valores.

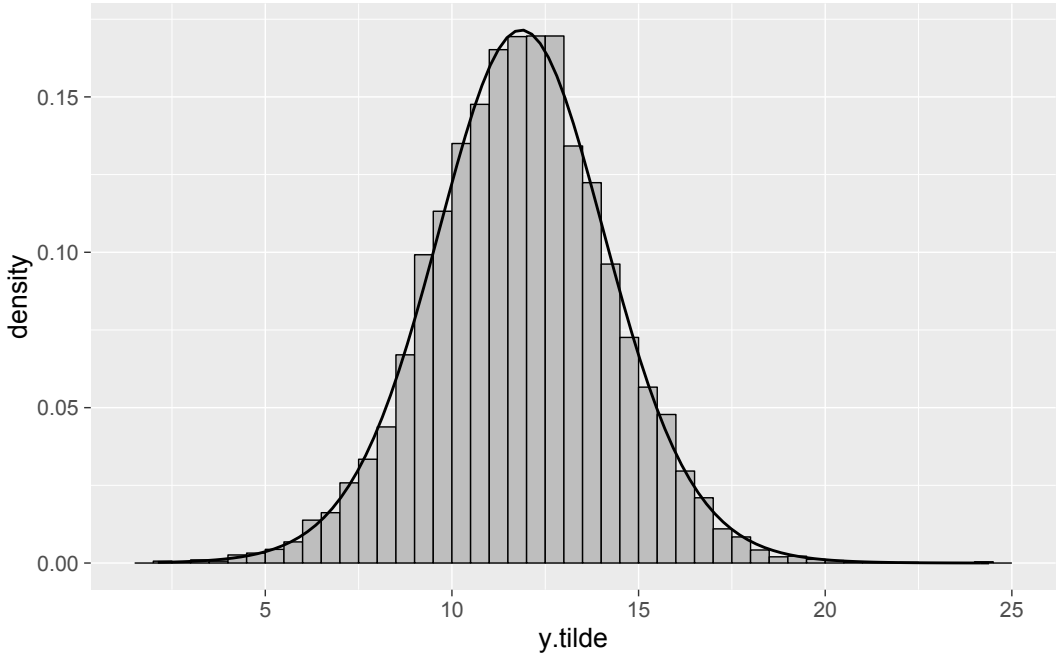


Figura 3.1: *10 mil valores simulados de $\tilde{Y}$ y la función de densidad de la distribución predictiva de $\tilde{Y}$.*

3.2 Normal multivariante con media desconocida y varianza conocida

Cuando la distribución usada para describir el comportamiento de los datos es una distribución normal multivariante, las técnicas de inferencia no se distancian mucho del caso univariado. Se debe tener en cuenta el manejo matricial de las formas cuadráticas y las propiedades básicas del cálculo de matrices. Los desarrollos y resultados derivados de esta sección redundarán en el análisis de los modelos lineales con el enfoque bayesiano.

Sea $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ un vector aleatorio cuya distribución es normal multivariante dada por

$$p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})\right\} \tag{3.2.1}$$

en donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ es el vector que contiene la media de cada uno de los componentes del vector $\mathbf{Y}$ y $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas de orden $p \times p$, simétrica y definida positiva. La verosimilitud para una muestra de n vectores aleatorios independientes e idénticamente distribuidos está dada por

$$p(\mathbf{Y}_1 \ldots, \mathbf{Y}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta})\right\}$$

Los parámetros que requieren estimación corresponden al vector de medias $\boldsymbol{\theta}$ y la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$. Por ahora, se asume que $\boldsymbol{\Sigma}$ es conocida y nos centramos en la estimación del vector de medias $\boldsymbol{\theta}$. Para la distribución previa, considerando que en general no hay restricción sobre los valores de los componentes de $\boldsymbol{\theta}$, asumimos que $\boldsymbol{\theta}$ sigue una distribución previa normal multivariante informativa y parametrizada por los hiper parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Gamma}$

$$p(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}) \propto \mid \boldsymbol{\Gamma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\}$$

En el siguiente resultado, encontramos la distribución posterior del parámetro $\boldsymbol{\theta}$.

Resultado 3.2.1. *La distribución posterior del vector $\boldsymbol{\theta}$ sigue una distribución normal multivariante*

$$\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\Sigma} \sim N_p(\boldsymbol{\mu}_n, \boldsymbol{\Gamma}_n).$$

En donde

$$\boldsymbol{\Gamma}_n = \left(\boldsymbol{\Gamma}^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{3.2.2}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Gamma}_n(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}) \tag{3.2.3}$$

Prueba. En primer lugar, nótese la siguiente identidad

$$\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta}) = \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}}) + n(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta}) \tag{3.2.4}$$

puesto que

$$\begin{aligned}
&\sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}}+\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}}+\bar{\mathbf{Y}}-\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}}) + \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta}) \\
&\qquad + (\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})' + \sum_{i=1}^{n}(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n}(\mathbf{Y}_i-\bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\bar{\mathbf{Y}}) + n(\bar{\mathbf{Y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{Y}}-\boldsymbol{\theta})
\end{aligned}$$

Por otro lado, de la definición de distribución previa, se tiene que

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\Sigma}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \\
&\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[n(\bar{\mathbf{y}} - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-n\bar{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - n\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}} + n\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\theta}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'(\boldsymbol{\Gamma}^{-1} + n\boldsymbol{\Sigma}^{-1})\boldsymbol{\theta} - 2\boldsymbol{\theta}'(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}})\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n + \boldsymbol{\mu}_n\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\mu}_n\right]\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_n)'\boldsymbol{\Gamma}_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_n)\right\}
\end{aligned}$$

La cual corresponde al núcleo de una distribución normal multivariante con el vector de medias $\boldsymbol{\mu}_n$ y matriz de varianzas $\boldsymbol{\Gamma}_n$. ■

Observando los parámetros de la distribución posterior, podemos ver que $\boldsymbol{\Gamma}_n^{-1} = \boldsymbol{\Gamma}^{-1} + (\boldsymbol{\Sigma}/n)^{-1}$. Teniendo en cuenta que la matriz de varianzas y covarianzas es una medida de dispersión de la distribución alrededor de su media, la inversa de dicha matriz se puede ver como una medida de precisión de qué tanto se concentra la distribución alrededor de la media. Así, podemos ver que la precisión posterior viene siendo la suma entre la precisión previa y la precisión de la estaimación clásica del parámetro $\boldsymbol{\theta}$.

En cuanto a la media posterior $\boldsymbol{\mu}_n$, tenemos que

$$\begin{aligned}
\boldsymbol{\mu}_n &= \left(\boldsymbol{\Gamma}^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}\right) \\
&= \left(\mathbf{I} + n\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\right)^{-1}\boldsymbol{\mu} + \left(\frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} + \mathbf{I}\right)^{-1}\bar{\mathbf{y}} \\
&= \underbrace{\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + n\boldsymbol{\Gamma}\right)^{-1}}_{\mathbf{A}_1}\boldsymbol{\mu} + \underbrace{n\boldsymbol{\Gamma}\left(\boldsymbol{\Sigma} + n\boldsymbol{\Gamma}\right)^{-1}}_{\mathbf{A}_2}\bar{\mathbf{y}}
\end{aligned}$$

De donde podemos que ver que la media posterior $\boldsymbol{\mu}_n$ se puede escribir como $\boldsymbol{\mu}_n = \mathbf{A}_1\boldsymbol{\mu} + \mathbf{A}_2\bar{\mathbf{y}}$ donde $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{I}$. Es claro que en el caso univariado, $\mathbf{A}_1$, $\boldsymbol{\mu}$, $\mathbf{A}_2$ y $\bar{\mathbf{y}}$ son todos escalares, y $\boldsymbol{\mu}_n$ es un valor intermedio entre $\boldsymbol{\mu}$ y $\bar{\mathbf{y}}$. Mientras que en caso multivariado, $\mathbf{A}_1\boldsymbol{\mu} + \mathbf{A}_2\bar{\mathbf{y}}$ es similar a una combinación convexa entre los vectores $\boldsymbol{\mu}$ y $\bar{\mathbf{y}}$, pero los coeficientes son matrices en vez de escalares.

Para ilustar la relación de $\boldsymbol{\mu}_n$ con $\boldsymbol{\mu}$ y $\bar{\mathbf{y}}$, tomamos el caso de $p = 2$, y denotamos $\mathbf{A} = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + n\boldsymbol{\Gamma}\right)^{-1}$. Es claro que $\mathbf{A}$ es una matriz simétrica y definida positiva, lo denotaremos con $\mathbf{A}_1 = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$, donde $a_{11} > 0$ y $a_{22} > 0$. De esta forma

$$\begin{aligned}
\boldsymbol{\mu}_n &= \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} 1 - a_{11} & -a_{12} \\ -a_{12} & 1 - a_{22} \end{pmatrix}\begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \\
&= \begin{pmatrix} a_{11}\mu_1 + (1 - a_{11})\bar{y}_1 + a_{12}(\mu_2 - \bar{y}_2) \\ a_{22}\mu_2 + (1 - a_{22})\bar{y}_2 + a_{12}(\mu_1 - \bar{y}_1) \end{pmatrix}
\end{aligned}$$

Al observar la primera entrada de $\boldsymbol{\mu}_n$, podemos ver que este se compone de una combinación convexa entre μ_1 y $\bar{y}_1$ (pues $a_{11} > 0$) y una parte que depende de la diferencia $\mu_2 - \bar{y}_2$; un comportamiento silimar se observa en la segunda entrada de $\boldsymbol{\mu}_n$. Esta observación es interesante, pues ilustra que cada componente de la media posterior $\boldsymbol{\mu}_n$ no siempre será un promedio ponderado del componentes corrspondiente de la media previa y la estimación clásica.

Ilustramos los resultados encontrados suponiendo las dos siguientes situaciones

1. Supong que se quiere estimar el vector de medias $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ con una matriz de varianzas y covarianzas conocida de $\boldsymbol{\Sigma} = \begin{pmatrix} 20 & 8 \\ 8 & 30 \end{pmatrix}$. Para eso tenemos 10 datos que corresponden a vectores bivariadas con $\bar{\mathbf{y}} = (150, 230)'$. Como información previa, suponga que $\boldsymbol{\mu} = (100, 200)'$ y $\boldsymbol{\Gamma} = \begin{pmatrix} 5 & 3 \\ 3 & 10 \end{pmatrix}$. Cálculos arrojan que $\boldsymbol{\Gamma}_n = \begin{pmatrix} 1.42 & 0.64 \\ 0.64 & 2.31 \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} 0.3 & -0.026 \\ -0.026 & 0.235 \end{pmatrix}$ y $\boldsymbol{\mu}_n = (136, 224)$, podemos ver que en este caso, cada componente de $\boldsymbol{\mu}_n$ se encuentra entre los componentes correspondientes de $\boldsymbol{\mu}$ y $\bar{\mathbf{y}}$. Los códigos computacionales se muestran a continuación.

```
n <- 10
mu <- matrix(c(100,200)); y.bar <- matrix(c(150,230))
Gamma <- matrix(c(5,3,3,10),2,2); Sigma <- matrix(c(20,8,8,30),2,2)
Gamma.n <- solve(solve(Gamma) + n*solve(Sigma))
A <- Sigma %*% solve(Sigma + n*Gamma)
mu.n <- Gamma.n %*% (solve(Gamma)%*%mu + n*solve(Sigma)%*%y.bar)
```

2. Tomamos los mismo datos del caso anterior, pero suponga que $\bar{\mathbf{y}} = (150, 2300)'$, las matrices $\boldsymbol{\Gamma}_n$ y $\mathbf{A}$ no cambian de valor, pero la media de la distribución posterior está dada por $\boldsymbol{\mu}_n = (190, 1808)$. Podemos ver que el primer componente de $\boldsymbol{\mu}_n$ no está entre 100 y 150 que corresponden a las estimaciones previa y clásica, respectivamente, esto se debe a que la diferencia entre μ_2 y $\bar{y}_2$ es muy grande.

Distribución previa no informativa

Al tener en cuenta que la distribución previa del parámetro $\boldsymbol{\theta}$ es la distribución normal multivariada, y al observar la forma de la función de densidad, se puede afirmar que cuando $|\boldsymbol{\Gamma}^{-1}|$ es muy pequeño, los parámetros previas $\boldsymbol{\mu}$ y $\boldsymbol{\Gamma}$ pierden peso en los cálculos de $\boldsymbol{\mu}_n$ y $\boldsymbol{\Gamma}_n$. En este caso se puede ver que

$$\boldsymbol{\Gamma}_n \approx n^{-1}\boldsymbol{\Sigma}$$
$$\boldsymbol{\mu}_n \approx \bar{\mathbf{y}}$$

De donde podemos concluir que la estimación bayesiana será muy cercana a la estimación clásica $\bar{y}$, más aún, el intervalo de credibilidad también será muys similar al intervalo de confianza del enfoque clásico.

Ejemplo 3.2.1. Student (1908) introdujo un conjunto de datos clásicos sobre el incremento en horas de sueño producido con 2 medicamentos soporíferos diferentes comparados con grupo control en 10 pacientes. Estos datos se pueden encontrar en `R` con nombre `sleep`. Estos datos se pueden ver como realizaciones de vectores aleatorios con distribución normal bivariada. Supongamos que la matriz de varianzas y covarianzas de la distribución es conocida e igual a $\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}$, y el parámetro de interés es el vector de medias $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. Para la distribución previa, suponemos que $\boldsymbol{\mu} = (0, 1)'$, es decir que el primer medicamento no tiene ningún efecto soporífero, mientras que el segundo medicamento tiene un efecto promedio de aumentar 1 hora de sueño, también asumimos que $\Gamma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. Los siguientes códigos de `JAGS` ilustra el procedimiento de estimación del parámetro de interés.

```
set.seed(123456)
n <- 10
mu<- as.vector(c(0,1)); Gamma <- matrix(c(2,0,0,2),2,2)
Sigma  <- matrix(c(1,0.6,0.6,2),2,2); Tau <- solve(Sigma)

NormMult1.model <- function(){
  for(i in 1 : n)
  {
    y[i, 1:2] ~ dmnorm(theta[], Tau[,])
  }
  theta[1:2] ~ dmnorm(mu[], Gamma[,])
}

y <- structure(.Data = sleep[,1], .Dim=c(10,2))

NormMult1.data <- list("y","n","mu","Gamma","Tau")
NormMult1.param <- c("theta")
NormMult1.inits <- function(){
  list("theta"=c(0,0))
}

NormMult1.fit <- jags(data=NormMult1.data, inits=NormMult1.inits, NormMult1.param,
                      n.iter=10000, n.burnin=1000, model.file=NormMult1.model)

print(NormMult1.fit)
```

De donde podemos que la estimación bayesiana para el aumento de sueño es de 0.54 y 1.91 para los dos medicamentos, respectivament; mientras que los intervalos de crediblidad del 95 % corresponden a (0.004, 1.079) y (1.164, 2.629).

A continuación, mostramos los códigos de `R` para llevar a cabo los cálculos directamente.

```
n <- 10
y <- structure(.Data = sleep[,1], .Dim=c(10,2))
Sigma  <- matrix(c(1,0.6,0.6,2),2,2)
mu<- as.vector(c(0,1))
Gamma <- matrix(c(2,0,0,2),2,2)
y.bar <- colMeans(y)
Gamma.n <- solve(solve(Gamma) + n*solve(Sigma))
mu.n <- Gamma.n%*%(solve(Gamma)%*%mu + n*solve(Sigma)%*%y.bar)
mu.n

##      [,1]
## [1,] 0.68
## [2,] 2.19

Gamma.n

##       [,1]  [,2]
## [1,] 0.094 0.052
## [2,] 0.052 0.180
```

De los resultados arrojados, vemos que la distribución posterior del parámetro está dada por

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N_2\left(\begin{pmatrix}0.68\\ 2.19\end{pmatrix}, \begin{pmatrix}0.094 & 0.052\\ 0.052 & 0.180\end{pmatrix}\right)$$

De esta forma, la estimación bayesiana obtenida para los efectos promedios corresponde a 0.68 horas y 2.19 horas, respectivamente, que son similares a los obtenidos por `JAGS`. En cuanto a los intervalos de credibilidad del 95 %, estas son dados por los percentiles 2.5 % y 97.5 % de las dos distribuciones posteriores marginales de θ_1 y θ_2. Estos intervalos se pueden obtener así:

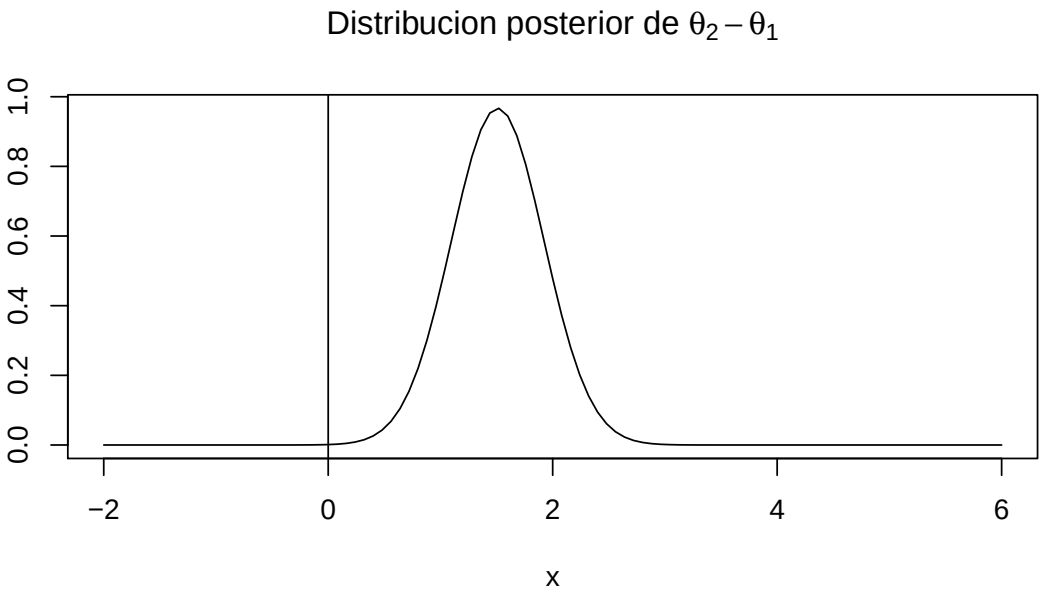
```
qnorm(c(0.025,0.975),mu.n[1],sqrt(Gamma.n[1,1]))

## [1] 0.08 1.28

qnorm(c(0.025,0.975),mu.n[2],sqrt(Gamma.n[2,2]))

## [1] 1.4 3.0
```

Ahora, suponga que el objetivo es comparar los medicmantos para concluir si el segundo medicamento es más efectivo que el primero, podemos encontrar la distribución posterior de la diferencia $\theta_2 - \theta_1$, utilizando propiedades de la distribución normal multivariante, podemos encontrar la distribución posterior de $\theta_2 - \theta_1$, calcular un intervalo de credibildiad para $\theta_2 - \theta_1$ e indagar cuál es la probabilidad de que θ_2 sea mayor a θ_1. Estos cálculos se pueden llevar a cabo de la siguiente forma

```
vec <- matrix(c(-1,1),1,2)
media <- vec %*% mu.n
varianza <- vec %*% Gamma.n %*% t(vec)
qnorm(c(0.025,0.975),media,sqrt(varianza))

## [1] 0.7 2.3

1-pnorm(0, media, sqrt(varianza))

## [1] 1
```

Observando los anteriores resultados, vemos que el intervalo de credibildad para $\theta_2 - \theta_1$ está dado por $(0.7, 2.3)$, el cual no contiene el valor 0, indicando que el segundo medicamento tiene un efecto mayor que el segundo. Adicionalmente vemos que con probabilidad 1, el segundo medicamento tiene efecto mayor al primero. Finalmente, podemos visulizar la distribución posterior con los siguientes comandos:

```
curve(dnorm(x, media, sqrt(varianza)), -2, 6, main=expression
      (paste("Distribucion posterior de ", theta[2] - theta[1])),ylab="")
abline(v=0)
```



De esta forma, podemos concluir que el segundo medicamento tiene un desempeño superior al primero.

Finalmente, ilustramos los resultados obtenidos al usar una distribución previa no informativa, para eso, usaremos $\mathbf{\Gamma} = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$, con $|\mathbf{\Gamma}^{-1}| = 0.0001$, representando una distribución previa no informativa. Los resultados de estimación arroja la siguiente distribución posterior para el vectro de parámetros

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0.75 \\ 2.33 \end{pmatrix}, \begin{pmatrix} 0.10 & 0.06 \\ 0.06 & 0.20 \end{pmatrix} \right)$$

Los intervalos de credibilidad del 95 % para los parámetros θ_1 y θ_2 están dados por $(0.129, 1.368)$ y $(1.451, 3.202)$, respectivamente. Observamos que estos intervalos de credibilidad son muy similares a los intervalos de confianza del 95 % del enfoque clásico[2] dados por $(0.130, 1.370)$ y $(1.453, 3.207)$. En cuanto a la comparación entre los dos medicamentos, lo dejamos como ejercicio para los lectores.

Finalmente, recordamos los dos siguientes resultados relacionados con la distribución normal multivariante que pueden resultar útiles en otros análisis.

Resultado 3.2.2. *La distribución posterior marginal de un subconjunto de parámetros, digamos $\boldsymbol{\theta}^{(1)}$ es también normal multivariante con media igual a la del subvector de medias apropiado, $\boldsymbol{\mu}_n^{(1)}$ y similar matriz de varianzas $\mathbf{\Gamma}_n^{(11)}$.*

Resultado 3.2.3. *La distribución posterior condicional de un subconjunto de parámetros, digamos $\boldsymbol{\theta}^{(1)}$, dado $\boldsymbol{\theta}^{(2)}$ es también normal multivariante dada por*

$$\boldsymbol{\theta}^{(1)} \mid \boldsymbol{\theta}^{(2)} \sim N_p \left(\boldsymbol{\mu}_n^{(1)} + \mathbf{\Gamma}_n^{(12)} \left(\mathbf{\Gamma}_n^{(22)} \right)^{-1} \left(\theta^{(2)} - \mu_n^{(2)} \right), \mathbf{\Gamma}_n^{(1|2)} \right).$$

En donde

$$\mathbf{\Gamma}_n^{(1|2)} = \mathbf{\Gamma}_n^{(11)} - \mathbf{\Gamma}_n^{(12)} \left(\mathbf{\Gamma}_n^{(22)} \right)^{-1} \mathbf{\Gamma}_n^{(21)} \tag{3.2.5}$$

[2]Estos intervalos de confianza fueron calculados con la expresión $\bar{y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$

$\boldsymbol{\mu}^{(1)}$ y $\boldsymbol{\mu}^{(2)}$ corresponden al vector de medias y $\boldsymbol{\Gamma}_n^{(11)}$, $\boldsymbol{\Gamma}_n^{(22)}$ denotan la matriz de varianzas y covarianzas de $\boldsymbol{\theta}^{(1)}$ y $\boldsymbol{\theta}^{(2)}$, respectivamente. $\boldsymbol{\Gamma}_n^{(12)}$ es la matriz de covarianzas entre $\boldsymbol{\theta}^{(1)}$ y $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\Gamma}_n^{(21)}$ es la matriz de covarianzas entre $\boldsymbol{\theta}^{(2)}$ y $\boldsymbol{\theta}^{(1)}$.

La prueba de los dos resultados anteriores se sigue inmediatamente de las propiedades de la distribución normal multivariante.

3.3 Normal multivariante con media y varianza desconocida

Al igual que en la distribución normal univariada, cuando se desconoce tanto el vector de medias como la matriz de varianzas y covarianzas de la distribución, es necesario plantear diversos enfoques y situarse en el más conveniente.[3] Suponiendo que el número de observaciones en la muestra aleatoria sea suficiente, existe otra situación que se debe surtir y es la asignación de las distribuciones previas para $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$. En estos términos, es posible

- Suponer que la distribución previa $p(\boldsymbol{\theta})$ es independiente de la distribución previa $p(\boldsymbol{\Sigma})$ y que ambas distribuciones son informativas. Luego, utilizar un análisis de simulación condicional conjunta para extraer muestras provenientes de las respectivas distribuciones posterior.
- Suponer que la distribución previa para $\boldsymbol{\theta}$ depende de $\boldsymbol{\Sigma}$ y escribirla como $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$, mientras que la distribución previa de $\boldsymbol{\Sigma}$ no depende de $\boldsymbol{\theta}$ y se puede escribir como $p(\boldsymbol{\Sigma})$. El análisis posterior de este enfoque encuentra la distribución posterior de $\boldsymbol{\Sigma} \mid \mathbf{Y}$ y con esta se encuentra la distribución posterior de $\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}$.
- Suponer que la distribución previa para $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ es una distribución no informativas.

3.3.1 Parámetros independientes con distribuciones previas informativas

En este enfoque se supone que las distribuciones previas para los parámetros de interés son independientes e informativas. La función de verosimilitud está dada en la expresión 3.2. Hacemos siguiente observación para lograr que las resultantes distribuciones posterior sean conjugadas.

$$\begin{aligned}
\sum_{i=1}^{n}(\mathbf{Y}_i-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\theta) &= traza\left(\sum_{i=1}^{n}(\mathbf{Y}_i-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\theta)\right)\\
&= \sum_{i=1}^{n} traza\left((\mathbf{Y}_i-\theta)'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\theta)\right)\\
&= \sum_{i=1}^{n} traza\left(\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i-\theta)(\mathbf{Y}_i-\theta)'\right)\\
&= traza\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i-\theta)(\mathbf{Y}_i-\theta)'\right)\\
&= traza\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}}\right)
\end{aligned}$$

Donde $\mathbf{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n}(\mathbf{Y}_i-\boldsymbol{\theta})(\mathbf{Y}_i-\boldsymbol{\theta})'$. En cuanto a la asignación de las distribuciones previas, para el vector de medias $\boldsymbol{\theta}$ es posible usar la distribución normal, esto es,

$$\boldsymbol{\theta} \sim Normal_p(\boldsymbol{\mu}, \boldsymbol{\Gamma})$$

[3] Nótese que en términos de parámetros, existen p parámetros correspondientes al vector de medias $\boldsymbol{\theta}$ y $\binom{p}{2} = \frac{p(p+1)}{2}$ parámetros correspondientes a la matriz de varianzas $\boldsymbol{\Sigma}$. Pensando en la gran cantidad de parámetros que se deben modelar, es necesario tener en cuenta que el número de datos en la muestra aleatoria sea lo suficientemente grande.

Por otro lado, la distribución para la matriz de varianzas $\boldsymbol{\Sigma}$ es

$$\boldsymbol{\Sigma} \sim Inversa - Wishart(\boldsymbol{\Lambda}, v)$$

donde v denota el grado de libertad y $\boldsymbol{\Lambda}$ la matriz de escala. Esto es, la función de densidad está dada por

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{v+p+1}{2}} \exp\left\{-\frac{1}{2} traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1})\right\}$$

Asumiendo independencia previa, la distribución previa conjunta resulta estar dada por

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) &= p(\boldsymbol{\theta}) p(\boldsymbol{\Sigma}) \\ &\propto | \boldsymbol{\Sigma} |^{-(v+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}) + (\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\} \end{aligned} \tag{3.3.1}$$

Una vez que se conoce la forma estructural de la distribución previa conjunta, es posible establecer la distribución posterior conjunta teniendo en cuenta la forma de la función de verosimilitud $p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})$ y la expresión equivalente para $\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})$ mostrada al inicio de esta sección. Adicionalmente, acudiendo a la simetría de las matrices $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}$ y $\mathbf{S}_{\boldsymbol{\theta}}$, se tiene que

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) &\propto p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \\ &\propto | \boldsymbol{\Sigma} |^{-(v+n+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\} \\ &\propto | \boldsymbol{\Sigma} |^{-(v+n+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})) + (\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\} \end{aligned} \tag{3.3.2}$$

Dado que la distribución posterior conjunta no tiene una forma estructural conocida, no es posible utilizar el método de integración analítica. Sin embargo, es posible obtener las distribuciones condicionales de cada uno de los parámetros suponiendo fijos los restantes y teniendo en cuenta que

$$p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}) \propto p(\boldsymbol{\theta}, \underbrace{\boldsymbol{\Sigma}}_{fijo} \mid \mathbf{Y}) \qquad \text{y} \qquad p(\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y}) \propto p(\underbrace{\boldsymbol{\theta}}_{fijo}, \boldsymbol{\Sigma} \mid \mathbf{Y})$$

Resultado 3.3.1. *La distribución posterior de la matriz de parámetros $\boldsymbol{\Sigma}$ condicional a $\boldsymbol{\theta}, \mathbf{Y}$ es*

$$\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y} \sim Inversa - Wishart_{v+n}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})$$

Prueba. La prueba es inmediata notando que

$$\begin{aligned} \boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \mathbf{Y} &\propto | \boldsymbol{\Sigma} |^{-(v+n+p+1)/2} \\ &\times \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})) + (\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Wishart_{v+n}(\boldsymbol{\Lambda} + \mathbf{S}_{\boldsymbol{\theta}})$. ■

Resultado 3.3.2. *La distribución posterior del vector de parámetros $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}, \mathbf{Y}$ es*

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y} \sim Normal_p(\boldsymbol{\mu}_n, \boldsymbol{\Gamma}_n) \tag{3.3.3}$$

donde $\boldsymbol{\mu}_n$ y $\boldsymbol{\Gamma}_n$ están dadas por las expresiones (3.2.2) y (3.2.3), respectivamente.

Prueba. La prueba de este resultado es inmediata pues corresponde a la misma situación de estimar $\boldsymbol{\theta}$ cuando $\boldsymbol{\Sigma}$ es conocida. ■

Una vez encontradas las distribuciones posteriores condicionales de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$, se puede obtener la estimación de estos parámetros vía el muestreador de Gibbs, que en este caso se resume en los siguientes pasos:

(1) Fijar un valor inicial para $\boldsymbol{\theta}$, lo denotamos por $\boldsymbol{\theta}_{(1)}$

(2) Simular un valor de la distribución de $\boldsymbol{\Sigma}|\boldsymbol{\theta}, \mathbf{Y}$ en 3.3.1 donde el parámetro $\mathbf{S}_{\boldsymbol{\theta}}$ que depende de $\boldsymbol{\theta}$, debe ser reemplazado por $\boldsymbol{\theta}_{(1)}$ del paso anterior. Este valor simulado se denotará por $\boldsymbol{\Sigma}_{(1)}$

(3) Simlar un valor de la distribución de $\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathbf{Y}$ en 3.3.3 donde en $\mathbf{mu}_n$ y $\boldsymbol{\Gamma}_n$ se debe reemplazar $\boldsymbol{\Sigma}$ por $\boldsymbol{\Sigma}$. Este valor simulado se denota por $\boldsymbol{\theta}$.

(4) Se repite los pasos (2) y (3) hasta completar un número de iteraciones suficientes para alcanzar la convergencia en ambos parámetros

Una vez tengamos los valores muestreados, se debe garantizar la convergencia y la correlación nula entre estos valores, con el fin de calcular las estimaciones. En el siguiente ejemplo ilustramos la implementación de este muestreador de Gibbs en `JAGS` y `R`.

Ejemplo 3.3.1. Retomamos los datos del efecto de dos medicamentos soporíferos introducidos por Student (1908) que fueron estudiados en el ejemplo 3.2.1 asumiendo que la matriz de varianzas y covarianzas es conocida. El vector de medias muestrales de estos datos están dados por $\bar{y} = (0.75, 2.33)'$, y la matriz de varianzas y covarianzas muestrales está dada por $\mathbf{S} = \begin{pmatrix} 3.20 & 2.85 \\ 2.85 & 4.01 \end{pmatrix}$.

Ahora supongamos que tanto el vector de medias como la matriz de varianzas y covarianzas y desconocidos. Para el vectro de medias, asumimos la distribución previa del ejemplo 3.2.1, es decir, $\boldsymbol{\mu} = (0, 1)'$ y $\boldsymbol{\Gamma} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. Para la matriz de varianzas y covarianzas asumimos la distribución inversa Wishart con matriz de escala igual a $\boldsymbol{\Lambda} = \begin{pmatrix} 20 & 8 \\ 8 & 20 \end{pmatrix}$ y grado de libertad $v = 10$, de esta forma, la estimación previa de $\boldsymbol{\Sigma}$ viene dada por $\frac{1}{v-2-1}\boldsymbol{\Lambda} = \begin{pmatrix} 2.86 & 1.14 \\ 1.14 & 2.86 \end{pmatrix}$.

Ilustramos los códigos de `JAGS` a continuación.

```
set.seed(123456)
n <- 10
mu<- as.vector(c(0,1)); Gamma <- matrix(c(2,0,0,2),2,2)
v <- 10; Lambda <- matrix(c(20,8,8,20),2,2)

NormMult2.model <- function(){
for(i in 1 : n)
{
  y[i, 1:2] ~ dmnorm(theta[], Tau[,])
}
theta[1:2] ~ dmnorm(mu[], Gamma[,])
Tau[1:2,1:2] ~ dwish(Lambda[,] , v)
Sigma[1:2,1:2] <- inverse(Tau[,])
}
```

```
y <- structure(.Data = sleep[,1], .Dim=c(10,2))

NormMult2.data <- list("y","n","mu","Gamma", "Lambda","v")
NormMult2.param <- c("theta", "Sigma")
NormMult2.inits <- function(){
list("theta"=c(0,0),"Tau"=diag(rep(1,2)))
}

NormMult2.fit <- jags(data=NormMult2.data, inits=NormMult2.inits, NormMult2.param,
                      n.iter=10000, n.burnin=1000, model.file=NormMult2.model)

print(NormMult2.fit)
```

Con base en los resultados anteriores, podemos ver que la estimación bayesiana para el número de horas de sueño producidas por los dos medicamentos son 0.29 y 1.74, respectivamente. En cuanto a la estimación de la matriz de varianzas y covarianzas, ésta está dada por $\hat{\mathbf{\Sigma}} = \begin{pmatrix} 3.08 & 2.2 \\ 2.2 & 3.63 \end{pmatrix}$.

A continuación se muestran los códigos para implementar el muestreador de Gibbs de forma maual en R.

código R

```
library(MCMCpack)
library(mvtnorm)
y <- as.matrix(data.frame(M1=sleep[1:10,1], M2=sleep[-(1:10),1]))
y.bar <- colMeans(y)
n <- nrow(y)

#parametros previos de theta
mu<- as.vector(c(0,1)); Gamma <- matrix(c(2,0,0,2),2,2)
#parametros previos de Sigma
v <- 10
Lambda <- matrix(c(20,8,8,20),2,2); Lambda.inv <- solve(Lambda)

nsim <- 10000
theta.pos <- matrix(NA,nsim,2)
Sigma.pos <- array(NA,c(nsim,2,2))

# Valor inicial de theta
theta.pos[1,] <- c(0,1)

#parametros posteriores de Sigma
v.pos <- v + n
matrix.theta <- kronecker(matrix(rep(1,n)),t(theta.pos[1,]))
S.theta <- t(y-matrix.theta) %*% (y-matrix.theta)
Lambda.pos <- Lambda + S.theta
#simulacion de la distribucion posterior condicional de Sigma
Sigma.pos[1,,] <- riwish(v.pos, Lambda.pos)

########################
# Muestreador de Gibbs #
########################
```

```
for(i in 2:nsim){
  #parametros posteriores de theta
  Gamma.n <- solve(solve(Gamma) + n*solve(Sigma.pos[i-1,,]))
  mu.n <- Gamma.n%*%(solve(Gamma)%*%mu + n*solve(Sigma.pos[i-1,,])%*%y.bar)
  #simulacion de la distribucion posterior condicional de theta
  theta.pos[i,] <- rmvnorm(1, mu.n, Gamma.n)
  #parametros posteriores de Sigma
  v.pos <- v + n
  matrix.theta <- kronecker(matrix(rep(1,n)),t(theta.pos[1,]))
  S.theta <- t(y-matrix.theta)%*%(y-matrix.theta)
  Lambda.pos <- Lambda + S.theta
  #simulacion de la distribucion posterior condicional de Sigma
  Sigma.pos[i,,] <- riwish(v.pos, Lambda.pos)
}
```

Una vez finalizada la ejecución del muestreador de Gibbs, debemos examinar la calidad de los valores muestreados para asegurar que las estimaciones bayesianas sean obtenidas de una muestra de valores que hayan convergido, y en segundo lugar sean aproximadamente incorrelacionados. Para eso a continuación observamos la gráfica de los valores muestreados para algunos parámetros (en particular, consideramos los parámetros θ_1, θ_2, σ_1^2 y σ_{12}), así como la gráfica de las autocorrelaciones muestrales. Con estas gráficas, observamos que los valores muestreados han alcanzado la convergencia, además estos tienen correlaciones cercanas a cero. De esta forma, podemos usar los valores muestreados para calcular las estimaciones y los intervalos de credibilidad.

```
theta.Bayes <- colMeans(theta.pos)
Sigma.Bayes <- matrix(c(mean(Sigma.pos[,1,1]), mean(Sigma.pos[,2,1]),
                        mean(Sigma.pos[,1,2]), mean(Sigma.pos[,2,2])),2,2)
theta.Bayes

## [1] 0.54 2.04

Sigma.Bayes

##      [,1] [,2]
## [1,]  3.2  2.6
## [2,]  2.6  4.3
```

El procedimiento inferencial sobre la comparación entre los efectos de los dos medicamentos se puede realizar de la misma manera como ilustró el ejemplo 3.2.1.

3.3.2 Parámetros dependientes

Al igual que en el caso univariado, la inferencia posterior de los parámetros de interés debe ser llevada a cabo en dos etapas: En la primera, se debe establecer la distribución previa conjunta para ambos parámetros mediante

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = p(\boldsymbol{\Sigma})p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$$

Luego, en la segunda etapa es posible analizar posterior propiamente cada uno de los parámetros de interés puesto que

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma})p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

Al igual que en el caso univariado, la anterior formulación conlleva a asignar una distribución previa para $\boldsymbol{\theta}$ dependiente de la matriz $\boldsymbol{\Sigma}$. Esto quiere decir que en la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$ el valor de $\boldsymbol{\Sigma}$ se considera una constante fija y conocida. Siguiendo los lineamientos de la Sección 3.2, una distribución previa para $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}$ es

$$p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}) \sim Normal_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/c_0)$$

Donde c_0 es una constante. Por otro lado, y siguiendo los argumentos de la sección anterior, una posible opción para la distribución previa de $\boldsymbol{\Sigma}$, corresponde a

$$p(\boldsymbol{\Sigma}) \sim Inversa - Wishart_{v_0}(\boldsymbol{\Lambda})$$

Resultado 3.3.3. *La distribución previa conjunta de los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ está dada por*

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-(v_0+p)/2-1} \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}) + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\}$$

Prueba. La prueba es inmediata al multiplicar las densidades y asignar los términos que no dependen de los parámetros de interés a la constante de proporcionalidad. ■

Para encontrar las distribuciones posterior de cada uno de los parámetros de interés se utilizan argumentos similares a los de la sección 3.1.2.

Resultado 3.3.4. *La distribución posterior de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}, \mathbf{Y}$ está dada por*

$$\theta \mid \boldsymbol{\Sigma}, \mathbf{Y} \sim Normal_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/(n+c_0))$$

donde

$$\mu_n = \frac{n\bar{\mathbf{Y}} + c_0\boldsymbol{\mu}}{n+c_0}$$

Prueba. Utilizando propiedades de la distribución condicional, tenemos que

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathbf{Y}) &\propto p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y}) \\ &\propto \mid \boldsymbol{\Sigma} \mid^{-(v_0+p)/2-1} \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1}) + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right\} \\ &\quad \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta})\right\} \\ &\propto \exp\left\{-\frac{1}{2}c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta})\right\} \end{aligned}$$

La anterior expresión es la misma de $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y})$ del resulado 3.2.1 donde $\boldsymbol{\Sigma}/c_0$ toma el valor de $\boldsymbol{\Gamma}$. así, teniendo en cuenta las ecuaciones (3.2.2) y (3.2.3), podemos afirmar que el vector de medias y la matriz de varianzas y covarianzas posterior están dadas por

$$\boldsymbol{\Gamma}_n = ((\boldsymbol{\Sigma}/c_0)^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1} = \frac{\boldsymbol{\Sigma}}{n+c_0} \tag{3.3.4}$$

$$\boldsymbol{\mu}_n = \frac{\boldsymbol{\Sigma}}{n+c_0}((\boldsymbol{\Sigma}/c_0)^{-1}\boldsymbol{\mu} + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}) = \frac{n\bar{\mathbf{Y}} + c_0\boldsymbol{\mu}}{n+c_0} \tag{3.3.5}$$

■

En cuanto a la distribución de $\boldsymbol{\Sigma}$, se tiene el siguiente resultado:

Resultado 3.3.5. *La distribución marginal posterior de la matriz de parámetros* $\boldsymbol{\Sigma}$ *es*

$$\boldsymbol{\Sigma} \mid \mathbf{Y} \sim Inversa - Whishart_{n+v_0}(\boldsymbol{\Lambda}_n)$$

Donde

$$\boldsymbol{\Lambda}_n = \boldsymbol{\Lambda} + (n-1)\mathbf{S} + \frac{c_0 n}{c_0+n}(\boldsymbol{\mu}-\bar{\mathbf{y}})(\boldsymbol{\mu}-\bar{\mathbf{y}})' \tag{3.3.6}$$

con S *la matriz de varianzas y covarianzas muestrales.*

Prueba.

$$\begin{aligned}
& p(\boldsymbol{\Sigma} \mid \mathbf{Y}) \\
&= \int_{R^p} p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) d\boldsymbol{\theta} \\
&\propto \int_{R^p} \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n)/2-1} \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}) + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) + \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta})\right]\right\} d\boldsymbol{\theta} \\
&\propto \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n)/2-1} \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1})\right]\right\} \\
&\qquad \int_{R^p} \exp\left\{-\frac{1}{2}\left[c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) + \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta})\right]\right\} d\boldsymbol{\theta} \\
&\propto \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n)/2-1} \exp\left\{-\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}) + c_0\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\bar{\mathbf{y}}) + n\bar{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}}\right]\right\} \\
&\qquad \int_{R^p} \exp\left\{-\frac{1}{2}\left[c_0\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2c_0\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2n\bar{\mathbf{y}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + n\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right]\right\} d\boldsymbol{\theta} \\
&\propto \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n)/2-1} \exp\left\{-\frac{1}{2}\left[traza\left((\boldsymbol{\Lambda} + c_0\boldsymbol{\mu}\boldsymbol{\mu}' + \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})' + n\bar{\mathbf{y}}\bar{\mathbf{y}}')\boldsymbol{\Sigma}^{-1}\right)\right]\right\} \\
&\quad \mid \frac{\boldsymbol{\Sigma}}{c_0+n} \mid^{1/2} \exp\left\{\frac{1}{2}\frac{c_0\boldsymbol{\mu}'+n\bar{\mathbf{y}}'}{c_0+n}\left(\frac{\boldsymbol{\Sigma}}{c_0+n}\right)^{-1}\frac{c_0\boldsymbol{\mu}+n\bar{\mathbf{y}}}{c_0+n}\right\} \\
&\quad \underbrace{\int_{R^p} \mid \frac{\boldsymbol{\Sigma}}{c_0+n} \mid^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\frac{c_0\boldsymbol{\mu}+n\bar{\mathbf{y}}}{c_0+n}\right)'\left(\frac{\boldsymbol{\Sigma}}{c_0+n}\right)^{-1}\left(\boldsymbol{\theta}-\frac{c_0\boldsymbol{\mu}+n\bar{\mathbf{y}}}{c_0+n}\right)\right\} d\boldsymbol{\theta}}_{\text{Igual a 1}}
\end{aligned}$$

Por otro lado,

$$\begin{aligned}
& \frac{c_0\boldsymbol{\mu}'+n\bar{\mathbf{y}}'}{c_0+n}\left(\frac{\boldsymbol{\Sigma}}{c_0+n}\right)^{-1}\frac{c_0\boldsymbol{\mu}+n\bar{\mathbf{y}}}{c_0+n} \\
&= \frac{1}{c_0+n}(c_0\boldsymbol{\mu}'+n\bar{\mathbf{y}}')\boldsymbol{\Sigma}^{-1}(c_0\boldsymbol{\mu}+n\bar{\mathbf{y}}) \\
&= traza\left(\frac{1}{c_0+n}(c_0\boldsymbol{\mu}+n\bar{\mathbf{y}})(c_0\boldsymbol{\mu}'+n\bar{\mathbf{y}}')\boldsymbol{\Sigma}^{-1}\right) \\
&= traza\left(\left(\frac{c_0^2\boldsymbol{\mu}\boldsymbol{\mu}'}{c_0+n} + \frac{2c_0 n\bar{\mathbf{y}}\boldsymbol{\mu}'}{c_0+n} + \frac{n^2\bar{\mathbf{y}}\bar{\mathbf{y}}'}{c_0+n}\right)\boldsymbol{\Sigma}^{-1}\right)
\end{aligned}$$

Reemplazando la anterior expresión en $p(\boldsymbol{\Sigma} \mid \mathbf{Y})$, se tiene que

$$\begin{aligned} & p(\boldsymbol{\Sigma} \mid \mathbf{Y}) \\ \propto & \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n+1)/2} \exp\left\{-\frac{1}{2} traza\left[\left(\boldsymbol{\Lambda} + (n-1)\mathbf{S} + \frac{c_0 n}{c_0+n}(\boldsymbol{\mu}-\bar{\mathbf{y}})(\boldsymbol{\mu}-\bar{\mathbf{y}})'\right)\boldsymbol{\Sigma}^{-1}\right]\right\} \end{aligned}$$

la cual corresponde a la distribución deseada. ■

En términos de simulación de densidades, para obtener las estimaciones bayesians de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ se debe primero simular valores de $\boldsymbol{\Sigma}$ de la distribución $p(\boldsymbol{\Sigma} \mid \mathbf{Y})$ y luego, se debe utilizar estos valores para simular valores de $\boldsymbol{\theta}$ de la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y})$.

Una forma equivalente de obtener las estimaciones es calcular directamente la esperanza teórica de las distribuciones posteriores marginales de $\boldsymbol{\theta}$ y de $\boldsymbol{\Sigma}$.

Del resultado 3.3.5, podemos concluir que la estimación bayesiana de la matriz de varianza y covarianzas $\boldsymbol{\Sigma}$ está dada por

$$\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Lambda} + (n-1)\mathbf{S} + \frac{c_0 n}{c_0+n}(\boldsymbol{\mu}-\bar{\mathbf{y}})(\boldsymbol{\mu}-\bar{\mathbf{y}})'}{n + v_0 - p - 1}$$

Teniendo en cuenta que la estimación previa de $\boldsymbol{\Sigma}$ viene dada por $\hat{\boldsymbol{\Sigma}}_{pre} = \frac{\boldsymbol{\Lambda}}{v_0-p-1}$, podemos ver que la estimación bayesiana de $\boldsymbol{\Sigma}$ está conformada por tres componentes: la estimación previa $\hat{\boldsymbol{\Sigma}}_{pre}$, la estimación clásica $\mathbf{S}$ y una medida de discrepancia entre la estimación previa y la clásica de $\boldsymbol{\theta}$. Para encontrar correctas formas de escoger los parámetros previas de $\boldsymbol{\Sigma}$, por ahora ignoramos el último componente, y vemos que la estimación previa $\hat{\boldsymbol{\Sigma}}_{pre}$ y la estimación clásica $\mathbf{S}$ entran al cómputo de la estimación bayesiana con los pesos de $v_0 - p - 1$ y $n - 1$, de esta forma, podemos escoger v_0 tal que $v_0 - p$ represente el número de la información previa, y el valor de $\boldsymbol{\Lambda}$ se puede calcular a partir de v_0 y $\hat{\boldsymbol{\Sigma}}_{pre}$.

El siguiente resultado muestra la distribución posterior marginal de $\boldsymbol{\theta}$.

Resultado 3.3.6. *La distribución marginal posterior del parámetro $\boldsymbol{\theta}$ es la distribución t de Student multivariante tal que*

$$\boldsymbol{\theta} \mid \mathbf{Y} \sim t_{n+v_0-p+1}\left(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Lambda}_n}{(c_0+n)(n+v_0-p+1)}\right)$$

con $\boldsymbol{\mu}_n = \frac{c_0\boldsymbol{\mu}+n\bar{\mathbf{y}}}{c_0+n}$ y $\boldsymbol{\Lambda}_n$ dado en la ecuación (3.3.6).

Prueba.

$$
\begin{aligned}
& p(\boldsymbol{\theta} \mid \mathbf{Y}) \\
&= \int_{R^p \times R^p} p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) d\boldsymbol{\Sigma} \\
&= \int_{R^p \times R^p} \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n)/2-1} \exp\left\{ -\frac{1}{2}\left[traza(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}) + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) + \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\theta}) \right] \right\} d\boldsymbol{\Sigma} \\
&= \int_{R^p \times R^p} \mid \boldsymbol{\Sigma} \mid^{-(v_0+p+n+2)/2} \exp\left\{ -\frac{1}{2} traza\left[\boldsymbol{\Lambda} + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})' + \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})(\mathbf{y}_i-\boldsymbol{\theta})' \right] \boldsymbol{\Sigma}^{-1} \right\} d\boldsymbol{\Sigma} \\
&\propto \left| \boldsymbol{\Lambda} + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})' + \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})(\mathbf{y}_i-\boldsymbol{\theta})' \right|^{-\frac{v_0+n+1}{2}} \\
&= \left| \boldsymbol{\Lambda} + c_0(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})' + \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})' + n(\bar{\mathbf{y}}-\boldsymbol{\theta})(\bar{\mathbf{y}}-\boldsymbol{\theta})' \right|^{-\frac{v_0+n+1}{2}} \\
&= \left| \boldsymbol{\Lambda} + (n-1)\mathbf{S} + \frac{c_0 n}{c_0+n}(\boldsymbol{\mu}-\bar{\mathbf{y}})(\boldsymbol{\mu}-\bar{\mathbf{y}})' + (c_0+n)(\boldsymbol{\theta}-\boldsymbol{\mu}_n)(\boldsymbol{\theta}-\boldsymbol{\mu}_n)' \right|^{-\frac{v_0+n+1}{2}} \\
&= \left| \boldsymbol{\Lambda}_n + (c_0+n)(\boldsymbol{\theta}-\boldsymbol{\mu}_n)(\boldsymbol{\theta}-\boldsymbol{\mu}_n)' \right|^{-\frac{v_0+n+1}{2}} \\
&\propto \left| \mathbf{I}_p + (c_0+n)\boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)(\boldsymbol{\theta}-\boldsymbol{\mu}_n)' \right|^{-\frac{v_0+n+1}{2}} \\
&= \left| 1 + (c_0+n)(\boldsymbol{\theta}-\boldsymbol{\mu}_n)'\boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_n) \right|^{-\frac{v_0+n+1}{2}} \\
&= \left| 1 + \frac{1}{n+v_0-p+1}(\boldsymbol{\theta}-\boldsymbol{\mu}_n)' \left(\frac{\boldsymbol{\Lambda}_n}{(c_0+n)(n+v_0-p+1)} \right)^{-1} (\boldsymbol{\theta}-\boldsymbol{\mu}_n) \right|^{-\frac{v_0+n+1}{2}}
\end{aligned}
$$

Esta expresión obtenida corresponde a la forma de la distribución t de Student multivariado. En el desarrollo se utilizó la propiedad $|\mathbf{I}+\mathbf{AB}| = |\mathbf{I}+\mathbf{BA}|$ para matrices $\mathbf{A}$ y $\mathbf{B}$ de tamaños compatibles para las multiplicaciones. ■

El anterior resultado indica que la estimación bayesiana del parámetro $\boldsymbol{\theta}$ está dada por

$$\hat{\boldsymbol{\theta}} = \mu_n = \frac{n\hat{\mathbf{Y}} + c_0\mu}{n+c_0} = \frac{n}{n+c_0}\hat{\mathbf{Y}} + \frac{c_0}{n+c_0}\mu$$

donde se puede observar que $\hat{\boldsymbol{\theta}}$ se acercará a la estimación clásica $\hat{\mathbf{y}}$ cuando n es grande comparado a c_0, de lo contrario se acercará a la estimación previa μ. La varianza posterior para el i-ésimo componente de $\boldsymbol{\theta}$ está dada por

$$var(\theta_i|\mathbf{Y}) = \frac{\lambda_{ii}}{(c_0+n)(n+v_0-p+1)} \frac{n+c_0-p+1}{n+c_0-p-1} \approx \frac{\lambda_{ii}}{(c_0+n)(n+v_0-p+1)}$$

donde λ_{ii} denota el i-ésimo elemento en la diagonal de la matriz $\boldsymbol{\Lambda}_n$.

Ejemplo 3.3.2. Peña (2002) reporta las mediciones de 6 variables indicadoras de desarrollo en 91 paises en los años noventa. Para este ejemplo, utilizamos tres variables: tasa de natalidad, tasa de mortalidad y mortalidad infantil en algunos paises de Suramérica y Asia mostrados en la tabla 3.1. específicamente, usaresmos los datos de los paises de Suramérica como datos muestrales y los de Asia para extraer la información previa.

País	Tasa Nat.	Tasa Mort.	Mort. Inf
Argentina	20.7	8.4	25.7
Bolivia	46.6	18	111
Brasil	28.6	7.9	63
Chile	23.4	5.8	17.1
Colombia	27.4	6.1	40
Ecuador	32.9	7.4	63
México	29	23.2	43
Paraguay	34.8	6.6	42
Perú	32.9	8.3	109.9
Uruguay	18	9.6	21.9
Venezuela	27.5	4.4	23.3
China	21.2	6.7	32
India	30.5	10.2	91
Indonesia	28.6	9.4	75
Malasia	31.6	5.6	24
Mongolia	36.1	8.8	68
Nepal	39.6	14.8	128
Singapur	17.8	5.2	7.5

Tabla 3.1: Tasa de natalidad, tasa de mortalidad, mortalidad infantil en algunos países

```
# Datos muestrales
y.sam <- data.frame(Nat=c(20.7,46.6, 28.6,23.4,27.4,32.9,29,34.8,32.9,18,27.5),
             Mort=c(8.4,18,7.9,5.8,6.1,7.4,23.2,6.6,8.3,9.6,4.4),
                Infa=c(25.7,111,63,17.1,40,63,43,42,109.9,21.9,23.3))
# Datos de la informacion previa
y.pre <- data.frame(Nat=c(21.2,30.5,28.6,31.6,36.1,39.6,17.8),
                         Mort=c(6.7,10.2,9.4,5.6,8.8,14.8,5.2),
                            Infa=c(32,91,75,24,68,128,7.5))
p <- ncol(y.pre)
# Estimacion clasica de los parametros
y.bar <- colMeans(y.sam); S <- var(y.sam); n <- nrow(y.sam)
# Estimacion previa de los parametros
mu <- colMeans(y.pre); c0 <- nrow(y.pre)
v0 <- p + nrow(y.pre); Lambda <- var(y.pre)*(v0-p-1)
# parametro de las distribuciones posteriores marginales
mu.n <- (n*y.bar + c0*mu)/(n+c0)
Lambda.n <- Lambda + (n-1)*S + matrix(mu-y.bar)%*%t(matrix(mu-y.bar))*c0*n
var.theta <- Lambda.n/((c0+n)*(n+v0-p+1))
mu.n

##  Nat Mort Infa
## 29.3  9.2 54.7

var.theta

##        Nat Mort Infa
## Nat   2.80 0.79 10.7
## Mort  0.79 1.35  2.2
## Infa 10.68 2.20 85.5
```

```
Lambda.n

##       Nat Mort  Infa
## Nat   957  270  3651
## Mort  270  463   751
## Infa 3651  751 29257
```

De los anteriores cálculos, se puede ver que la distribución posterior de $\boldsymbol{\theta}$ está dada por

$$\boldsymbol{\theta} \mid \mathbf{Y} \sim t_{19}\left(\begin{pmatrix}29.3\\9.2\\54.7\end{pmatrix}, \begin{pmatrix}2.80 & 0.79 & 10.68\\0.79 & 1.35 & 2.20\\10.68 & 2.20 & 85.55\end{pmatrix}\right)$$

Usando propiedades de la distribución multivariante t de Student, tenemos que $\theta_1 \sim t_{19}(29.29, 2.80)$, $\theta_2 \sim t_{19}(9.24, 1.35)$ y $\theta_3 \sim t_{19}(0.62, 85.55)$, de allí se puede encontrar fácilmente los intervalos de credibilidad para cada uno de estos tres parámetros.

En cuanto a la distribución posterior de $\boldsymbol{\Sigma}$, ésta está dada por

$$\boldsymbol{\Sigma} \mid \mathbf{Y} \sim Inversa - Wishart_{21}\left(\begin{pmatrix}957 & 270 & 3651\\270 & 463 & 751\\3651 & 751 & 29257\end{pmatrix}\right)$$

y la estimación bayesiana de $\boldsymbol{\Sigma}$ viene dada por $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix}56.3 & 15.9 & 214.8\\15.9 & 27.2 & 44.2\\214.8 & 44.2 & 1721.0\end{pmatrix}$. Por propiedad de la distribución inversa-Wishart, se puede concluir que los elementos diagonales de $\boldsymbol{\Sigma}$ tienen distribución inversa-Gamma. Por ejemplo, se tiene que $\sigma_1^2 \sim Inversa-Gamma(21/2, 56.3/2)$ y cualquier inferencia que se desear realizar sobre σ_1^2 es posible a partir de esta distribución.

Aparte de los análisis anteriores, también podemos realizar ejercicios de comparación y verificar posible independencia entre parejas de variables. Por ejemplo, queremos verificar la hipóstesis de que la tasa de natalidad es dos veces la tasa de mortalidad, esto es $\theta_1 = 2\theta_2$. Una forma de confirmar o refutar esta hipótesis es hallar el intervalo de credibilidad del $\theta_1 - 2\theta_2$ que se puede expresar como $(1, -2, 0)\begin{pmatrix}\theta_1\\\theta_2\\\theta_3\end{pmatrix}$. Por propiedad de la distribución t de Student multivariante, tenemos que $\theta_1 - 2\theta_2$ tiene distribución t de Student univariada con el mismo grado de libertad que $\boldsymbol{\theta}$, la esperanza está dada por $(1, -2, 0)\begin{pmatrix}29.3\\9.2\\54.7\end{pmatrix} = 10.8$ y la escala está dada por $(1, -2, 0)\begin{pmatrix}2.80 & 0.79 & 10.68\\0.79 & 1.35 & 2.20\\10.68 & 2.20 & 85.55\end{pmatrix}\begin{pmatrix}1\\-2\\0\end{pmatrix} = 5.054$, esto es,

$$\theta_1 - 2\theta_2 \mid \mathbf{Y} \sim t_{19}(10.8, 5.054)$$

De esta forma, un intervalo de credibilidad para $\theta_1 - 2\theta_2$ viene dado por los percentiles 2.5 % y 97.5 % de la anterior distribución, que a la vez son iguales a los percentiles 2.5 % y 97.5 % de la distribución t de Student estandarizada multiplicado por $\sqrt{5.054}$ y sumado 10.8. Este invervalo es igual a $(6.095, 15.505)$. Al observar que este intervalo no contiene el valor 0, podemos concluir que no es válido afirmar que la tasa de natalidad sea dos veces la tasa de mortalidad.

En el anterior análisis, vemos que el intervalo de crediblidad para $\theta_1 - 2\theta_2$ contiene solo valores positivos, lo cual es un indicio de que la variable $\theta_1 - 2\theta_2$ tenga la mayor parte de la función de densidad ubicado en ele eje positivo. He hecho podemos indagar $Pr(\theta_1 - 2\theta_2 > 0)$, esto lo podemos calcular de la distribución $t_{19}(10.8, 5.054)$ encontrada anteriormente. Esta probabilidad se puede calcular usando

`1-pt((0-10.8)/sqrt(5.054),19)` dando como resultado 0.9999383, de donde muestra una fuerte evidencia de que la tasa de natalidad es superior a dos veces la tasa de mortalidad.

Los anteriores resultados fueron obtenidos directamente de las distribuciones posteriores marginales $p(\boldsymbol{\theta}|\mathbf{Y})$ y $p(\boldsymbol{\Sigma}|\mathbf{Y})$. De forma equivalente también se puede usar las técnicas de simulación con base en las distribuciones $p(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{Y})$ y $p(\boldsymbol{\Sigma}|\mathbf{Y})$. A continuación se muestra los códigos:

```
library(MCMCpack)
library(mvtnorm)
# Datos muestrales
y.sam <- data.frame(Nat=c(20.7,46.6, 28.6,23.4,27.4,32.9,29,34.8,32.9,18,27.5),
             Mort=c(8.4,18,7.9,5.8,6.1,7.4,23.2,6.6,8.3,9.6,4.4),
                Infa=c(25.7,111,63,17.1,40,63,43,42,109.9,21.9,23.3))
# Datos de la informaci\'on previa
y.pre <- data.frame(Nat=c(21.2,30.5,28.6,31.6,36.1,39.6,17.8),
                        Mort=c(6.7,10.2,9.4,5.6,8.8,14.8,5.2),
                           Infa=c(32,91,75,24,68,128,7.5))
p <- ncol(y.pre)
# Estimacion clasica de los parametros
y.bar <- colMeans(y.sam); S <- var(y.sam); n <- nrow(y.sam)
# Estimacion previa de los parametros
mu <- colMeans(y.pre); c0 <- nrow(y.pre)
v0 <- p + nrow(y.pre); Lambda <- var(y.pre)*(v0-p-1)
# parametros de las posteriores
Lambda.n <- Lambda + (n-1)*S + matrix(mu-y.bar)%*%t(matrix(mu-y.bar))*c0*n
mu.n <- (n*y.bar + c0*mu)/(n+c0)

nsim <- 10000
theta.pos <- matrix(NA, nsim, p)
Sigma.pos <- array(NA, c(nsim,p,p))

for(i in 1:nsim){
  Sigma.pos[i,,] <- riwish(n+v0, Lambda.n)
  theta.pos[i,] <- rmvnorm(1, mu.n, Sigma.pos[i,,]/(n+c0))
}
# Estimaciones finales
theta.final <- colMeans(theta.pos)
Sigma.final <- matrix(c(mean(Sigma.pos[,1,1]),mean(Sigma.pos[,1,2]),
                        mean(Sigma.pos[,1,3]), mean(Sigma.pos[,2,1]),
                        mean(Sigma.pos[,2,2]), mean(Sigma.pos[,2,3]),
                        mean(Sigma.pos[,3,1]),mean(Sigma.pos[,3,2]),
                        mean(Sigma.pos[,3,3])), 3, 3)
theta.final

## [1] 29.3  9.2 54.8

Sigma.final

##      [,1] [,2] [,3]
## [1,]   56   16  213
## [2,]   16   27   44
## [3,]  213   44 1715
```

Podemos ver que los resultados obtenidos con los dos métodos. En cuanto al intervalo de credibilidad del $\theta_1 - 2\theta_2$, este se puede calcular con

```
quantile(theta.pos[,1]-2*theta.pos[,2], c(0.025, 0.975))

## 2.5%  98%
##  6.1 15.5
```

también podemos calcular $Pr(\theta_1 - 2\theta_2 > 0)$ como

```
sum(theta.pos[,1] > 2*theta.pos[,2])/nsim

## [1] 1
```

Podemos ver que estos resultados son muy simialres a los obtenidos usando $p(\boldsymbol{\theta}|\mathbf{Y})$.

3.3.3 parámetros no informativos

Gelman, Carlin, Stern & Rubin (2003) afirma que la distribución previa no informativa de Jeffreys conjunta para $\boldsymbol{\theta}, \boldsymbol{\Sigma}$, en este caso está dada por la siguiente expresión

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-(p+1)/2}$$

Nótese que para el caso de la distribución normal univariada, $p = 1$ y la anterior distribución previa se convierte en $p(\theta, \sigma^2) \propto \sigma^{-2}$, la cual coincide con la distribución previa no informativa de la ecuación 3.1.9

La distribución posterior conjunta para $\boldsymbol{\theta}, \boldsymbol{\Sigma}$ está dada por

$$p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) \propto \mid \boldsymbol{\Sigma} \mid^{-(p+n+1)/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right\}$$

De la anterior distribución, podemos encontrar la distribución condicional posterior de $\boldsymbol{\theta}$ dada en el siguiente resultado.

Resultado 3.3.7. *La distribución posterior del vector de parámetros $\boldsymbol{\theta}$ condicional a $\boldsymbol{\Sigma}, \mathbf{Y}$ es*

$$\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y} \sim Normal_p(\bar{\mathbf{y}}, \boldsymbol{\Sigma}/n)$$

Prueba. Algunas simples operaciones algebráicas muestran que:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}, \mathbf{Y}) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right\} \\ &\propto \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \bar{\mathbf{Y}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{\mathbf{Y}})\right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal_p(\bar{y}, \boldsymbol{\Sigma}/n)$. ■

En cuanto a la estimación de $\boldsymbol{\Sigma}$, en el siguiente resultado encontramos su distribución posterior.

Resultado 3.3.8. *La distribución marginal posterior de la matriz de parámetros* $\mathbf{\Sigma}$ *es*

$$\mathbf{\Sigma} \mid \mathbf{Y} \sim Inversa - Whishart_{n-1}(\mathbf{S})$$

donde $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$

Prueba. En primer lugar recordamos la expresión

$$\mathbf{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})' = \mathbf{S} + n(\boldsymbol{\theta} - \bar{\mathbf{y}})(\boldsymbol{\theta} - \bar{\mathbf{y}})'$$

Por otro lado, recurriendo a las propiedades del operador $traza$, e integrando la distribución posterior conjunta con respecto a $\boldsymbol{\theta}$, se tiene que

$$\begin{aligned}
p(\mathbf{\Sigma} \mid \mathbf{Y}) &= \int p(\boldsymbol{\theta}, \mathbf{\Sigma} \mid \mathbf{Y})\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n+1)/2} \int \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right\}\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n+1)/2} \int \exp\left\{-\frac{1}{2} traza(\mathbf{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}})\right\}\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n+1)/2} \int \exp\left\{-\frac{1}{2} traza(\mathbf{\Sigma}^{-1}(\mathbf{S} + n(\boldsymbol{\theta} - \bar{\mathbf{y}})(\boldsymbol{\theta} - \bar{\mathbf{y}})'))\right\}\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n)/2} \exp\left\{-\frac{1}{2} traza(\mathbf{\Sigma}^{-1}\mathbf{S})\right\} \\
&\qquad \times \int \mid \mathbf{\Sigma} \mid^{-1/2} \exp\left\{-\frac{n}{2} traza(\mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{\mathbf{y}})(\boldsymbol{\theta} - \bar{\mathbf{y}})')\right\}\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n)/2} \exp\left\{-\frac{1}{2} traza(\mathbf{\Sigma}^{-1}\mathbf{S})\right\} \\
&\qquad \times \int \mid \mathbf{\Sigma} \mid^{-1/2} \exp\left\{-\frac{n}{2} traza((\boldsymbol{\theta} - \bar{\mathbf{y}})'\mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{\mathbf{y}}))\right\}\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n)/2} \exp\left\{-\frac{1}{2} traza(\mathbf{\Sigma}^{-1}\mathbf{S})\right\} \\
&\qquad \times \int \underbrace{\mid \mathbf{\Sigma} \mid^{-1/2} \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \bar{\mathbf{y}})'\mathbf{\Sigma}^{-1}(\boldsymbol{\theta} - \bar{\mathbf{y}})\right\}}_{Normal_p(\bar{\mathbf{y}}, \mathbf{\Sigma}/n)}\, d\boldsymbol{\theta} \\
&= \mid \mathbf{\Sigma} \mid^{-(p+n)/2} \exp\left\{-\frac{1}{2} traza(\mathbf{\Sigma}^{-1}\mathbf{S})\right\}
\end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Whishart_{n-1}(\mathbf{S})$. ■

El anterior resultado indica que la estimación bayesiana de $\mathbf{\Sigma}$ cuando se utiliza una previa no informativa está dada por

$$\hat{\mathbf{\Sigma}} = \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'}{n - p - 2}$$

Esta expresión es muy similar a la estimación clásica de la matriz de varianzas y covarianzas dada por $\frac{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'}{n-1}$. Se puede observar que a medida que n se aumente, las dos expresiones darán

resultados muy similares, pero siempre la estimación bayesiana será mayor a la estimación clásica, especialmente en situaciones donde el tamaño muestral es pequeño.

Para obtener la estimación de $\boldsymbol{\Sigma}$ junto con la estimación de $\boldsymbol{\theta}$, podemos proceder de la siguiente forma para obetener valores simulados de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$ y así, obtener las estimaciones respectivas. Si el número de iteraciones se fija como G, entonces se procede a:

(1) Simular G valores de la distribución de $\boldsymbol{\Sigma}|\mathbf{Y}$, estos valores se denotan por $\boldsymbol{\Sigma}_{(1)}, \boldsymbol{\Sigma}_{(2)}, \cdots, \boldsymbol{\Sigma}_{(G)}$.

(2) Para cada valor de $\boldsymbol{\Sigma}_{(g)}$, con $g = 1, \cdots, G$, simlar un valor de la distribución de $\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathbf{Y}$, es decir, de la distribución $N_p(\bar{\mathbf{y}}, \boldsymbol{\Sigma}/n)$, donde $\boldsymbol{\Sigma}$ se reemplaza por $\boldsymbol{\Sigma}_{(g)}$. De sta forma, se obtiene los valores $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, \cdots, \boldsymbol{\theta}_{(G)}$.

El siguiente ejemplo ilustra la forma de obtener las estimaciones siguiendo el anterior procedimiento.

Ejemplo 3.3.3. Retomamos los datos del efecto de aumento en horas de sueño de dos medicamentos soporíferos utilizados en los ejemplos 3.2.1 y 3.3.1. Los siguientes códigos de `R` ilustra el procedimiento computacional para obtener valores de la distribución posterior conjunta de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$.

```
library(MCMCpack)
library(mvtnorm)
y <- as.matrix(data.frame(M1=sleep[1:10,1], M2=sleep[-(1:10),1]))
n <- nrow(y)
y.bar <- colMeans(y); S <- var(y)*(n-1)

nsim <- 10000
theta.pos <- matrix(NA, nsim, 2)
Sigma.pos <- array(NA, c(nsim,2,2))

for(i in 1:nsim){
  #simulacion de la distribucion posterior condicional de Sigma
  Sigma.pos[i,,] <- riwish(n-1, S)
  #simulacion de la distribucion posterior condicional de theta
  theta.pos[i,] <- rmvnorm(1, y.bar, Sigma.pos[i,,]/n)
}
```

Dado que en el cálculo no se hizo uso de valores iniciales y por la forma de las distribuciones posteriores de $p(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \mathbf{Y})$ y $\boldsymbol{\Sigma}|\mathbf{Y}$, los valores muestrados en la diferentes iteraciones no guardan relación entre sí, podemos usar directamente todos los valores muestreados para el cálculo de las estimaciones bayesianas.

```
theta.Bayes <- colMeans(theta.pos)
Sigma.Bayes <- matrix(c(mean(Sigma.pos[,1,1]),mean(Sigma.pos[,2,1]),
                        mean(Sigma.pos[,1,2]),mean(Sigma.pos[,2,2])), 2, 2)
theta.Bayes

## [1] 0.77 2.34

Sigma.Bayes

##      [,1] [,2]
## [1,]  4.9  4.3
## [2,]  4.3  6.0
```

Por otro lado, la estimación clásica de los parámetros está dada por

```
y.bar

##   M1   M2
## 0.75 2.33

var(y)

##     M1  M2
## M1 3.2 2.8
## M2 2.8 4.0
```

Podemos observar que en cuanto al parámetro $\boldsymbol{\theta}$ la estimación bayesiana es igual a la estimación clásica, mientras que la estimación bayesiana de $\boldsymbol{\Sigma}$ es mucho mayor que la estimación clásica, esto ocurre en situaciones cuando el tamaño muestral es pequeño.

En cuanto a la estimación por intervalo de los efectos promedios de los dos medicamentos, tenemos que:

```
quantile(theta.pos[,1], c(0.025,0.975))

## 2.5%  98%
## -0.6  2.2

t.test(y[,1])$conf.int

## [1] -0.53  2.03
## attr(,"conf.level")
## [1] 0.95

quantile(theta.pos[,2], c(0.025,0.975))

## 2.5%  98%
##  0.8  3.9

t.test(y[,2])$conf.int

## [1] 0.9 3.8
## attr(,"conf.level")
## [1] 0.95
```

Observamos que los resultados obtenidos con el enfoque bayesiano aunque no es exactamente igual a los obtenidos con el enfoque clásico, sí son muy similares.

En cuanto a la estimación por intervalo de confianza de las varianzas y covarianzas. Primero consideramos la varianza del primer medicamento denotada por σ_1^2. La distribución posterior de la matriz de varianzas y covarianzas está dada por

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \sim Inversa - Wishart_9(\mathbf{S})$$

con $\mathbf{S} = \sum_{i=1}^{10}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \begin{pmatrix} 28.81 & 25.64 \\ 25.64 & 36.08 \end{pmatrix}$. Usando propiedad de la distribución Inversa-Wishart, se puede concluir que la distribución marginal posterior de σ_1^2 está dada por $Inversa - Gamma(\alpha = \frac{9-1}{2}, \beta = \frac{28.81}{2})$, y su intervalo de credibilidad se puede calcular directamente de dicha distribución, o equivalentemente usando los percentiles muestrales de los valores de σ_1^2 muestrados. El intervalo obtenido por estos dos medios son muy similares como se puede ver a continuación.

```
library(pscl)
qigamma(0.025, alpha=8/2, beta=28.81/2)

## [1] 1.6

qigamma(0.975, alpha=8/2, beta=28.81/2)

## [1] 13

quantile(Sigma.pos[,1,1], c(0.025, 0.975))

## 2.5%  98%
##  1.6 13.4
```

El intervalo de confianza del 95 % se puede obtener con el siguiente código (consultar Zhang & Gutiérrez (2010, Sec.3.2.1) para mayor información)

```
c(9 * var(y[,1]) / qchisq(0.975,9), 9 * var(y[,2]) / qchisq(0.025,9))

## [1]  1.5 13.4
```

En comparación con el intervalo de credibilidad, el intervalo de confianza está ubicado levemente hacia la izquierda del eje real, esto se debe a que la estimación clásica de la varianza siempre será menor a la estimación bayesiana con una previa no informativa.

3.4 Multinomial

En esta sección discutimos el modelamiento bayesiano de datos provenientes de una distribución multinomial que corresponde a una extensión multivariada de la distribución binomial.

Suponga que $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ es un vector aleatorio con distribución multinomial, así, su distribución está parametrizada por el vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ y está dada por la siguiente expresión

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \binom{n}{y_1, \ldots, y_p} \prod_{i=1}^{p} \theta_i^{y_i} \qquad \theta_i > 0 \ , \ \sum_{i=1}^{p} y_i = n \ \text{ y } \ \sum_{i=1}^{p} \theta_i = 1 \tag{3.4.1}$$

donde

$$\binom{n}{y_1, \ldots, y_p} = \frac{n!}{y_1! \cdots y_p!}.$$

Como cada parámetro θ_i está restringido al espacio $\Theta = [0, 1]$, entonces es posible asignar a la distribución de Dirichlet como la distribución previa del vector de parámetros. Por lo tanto la distribución

previa del vector de parámetros $\boldsymbol{\theta}$, parametrizada por el vector de hiperparámetros $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$, está dada por

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} \prod_{i=1}^{p} \theta_i^{\alpha_i - 1} \qquad \alpha_i > 0 \text{ y } \sum_{i=1}^{p} \theta_i = 1 \tag{3.4.2}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 3.4.1. *La distribución posterior del parámetro $\boldsymbol{\theta}$ sigue una distribución Dirichlet$(y_1 + \alpha_1, \ldots, y_p + \alpha_p)$*

Prueba.

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \\ &= \binom{n}{y_1, \ldots, y_p} \prod_{i=1}^{p} \theta_i^{y_i} \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} \prod_{i=1}^{p} \theta_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{p} \theta_i^{y_i + \alpha_i - 1} \end{aligned}$$

Dado que $\sum_{i=1}^{p} \theta_i = 1$, entonces factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una vector aleatorio con distribución *Dirichelt*$(y_1 + \alpha_1, \ldots, y_p + \alpha_p)$. ■

Del anterior resultado, podemos ver que la estimación bayesiana de cada parámetro θ_i con $i = 1, \cdots, p$ está dada por

$$\hat{\theta}_i = \frac{y_i + \alpha_i}{\sum_{j=1}^{p} y_j + \sum_{j=1}^{p} \alpha_j}$$

Debido a que el valor de y_i normalmente denota el número de datos en la i-ésima categoría, y θ_i denota la probabilidad de que un dato está en esa categoría, la anterior expresión sugiere que podemos usar el número de datos en la i-ésima categoría como α_i. De esta forma, $\sum_{j=1}^{p} \alpha_j$ denota el número total de datos en la información previa, y la estimación de θ_i se puede ver como el proporción de datos en la i-ésima categoría combinando la información actual con la información previa.

En los dos siguientes resultados, examinamos la forma de la distribución predictiva previa y posterior para una nueva observación.

Resultado 3.4.2. *La distribución predictiva previa para una observación* $\mathbf{y}$ *está dada por*

$$p(\mathbf{Y}) = \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\sum_{i=1}^{p} \alpha_i)}{\prod_{i=1}^{p} \Gamma(\alpha_i)} \frac{\prod_{i=1}^{p} \Gamma(y_i + \alpha_i)}{\Gamma(\sum_{i=1}^{p} y_i + \sum_{i=1}^{p} \alpha_i)} \tag{3.4.3}$$

y define una auténtica función de densidad de probabilidad continua.

Prueba. De la definición de función de distribución predictiva se tiene que

$$\begin{aligned}
p(\mathbf{Y}) &= \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \, d\boldsymbol{\theta} \\
&= \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} \frac{\Gamma(y_1 + \alpha_1) \cdots \Gamma(y_p + \alpha_p)}{\Gamma(y_1 + \alpha_1 + \cdots + y_p + \alpha_p)} \\
&\times \int_0^1 \cdots \int_0^1 \frac{\Gamma(y_1 + \alpha_1 + \cdots + y_p + \alpha_p)}{\Gamma(y_1 + \alpha_1) \cdots \Gamma(y_p + \alpha_p)} \prod_{i=1}^{p} \theta_i^{y_i + \alpha_i - 1} \, d\theta_1 \cdots d\theta_p \\
&= \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} \frac{\Gamma(y_1 + \alpha_1) \cdots \Gamma(y_p + \alpha_p)}{\Gamma(y_1 + \alpha_1 + \cdots + y_p + \alpha_p)} \\
&= \binom{n}{y_1, \ldots, y_p} \frac{\Gamma(\sum_{i=1}^p \alpha_i)}{\prod_{i=1}^p \Gamma(\alpha_i)} \frac{\prod_{i=1}^p \Gamma(y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p y_i + \sum_{i=1}^p \alpha_i)}
\end{aligned}$$

■

Resultado 3.4.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva observación del vector aleatorio* $\tilde{\mathbf{y}}$ *de tamaño* p*, para* n^* *repeticiones del mismo experimento aleatorio, está dada por*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \binom{n^*}{\tilde{y}_1, \ldots, \tilde{y}_p} \frac{\Gamma(\sum_{i=1}^p (y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(y_i + \alpha_i)} \frac{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))} \tag{3.4.4}$$

Prueba. De la definición de función de distribución predictiva posterior se tiene que

$$\begin{aligned}
p(\tilde{\mathbf{y}} \mid \mathbf{Y}) &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \\
&= \binom{n^*}{\tilde{y}_1, \ldots, \tilde{y}_p} \frac{\Gamma(\sum_{i=1}^p (y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(y_i + \alpha_i)} \frac{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))} \\
&\times \int_0^1 \cdots \int_0^1 \frac{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)} \prod_{i=1}^{p} \theta_i^{\tilde{y}_i + y_i + \alpha_i - 1} \, d\theta_1 \cdots d\theta_p \\
&= \binom{n^*}{\tilde{y}_1, \ldots, \tilde{y}_p} \frac{\Gamma(\sum_{i=1}^p (y_i + \alpha_i))}{\prod_{i=1}^p \Gamma(y_i + \alpha_i)} \frac{\prod_{i=1}^p \Gamma(\tilde{y}_i + y_i + \alpha_i)}{\Gamma(\sum_{i=1}^p (\tilde{y}_i + y_i + \alpha_i))}
\end{aligned}$$

■

Ahora, suponga que no hay disponible ninguna fuente de información previa disponible, podemos usar la distribución previa no informativa de Jeffreys, teniendo en cuenta que en el caso de modelos multiparamétricos, este está dada por $p(\boldsymbol{\theta}) \propto \mid J(\boldsymbol{\theta}) \mid^{1/2}$, con

$$\begin{aligned}
J(\boldsymbol{\theta}) &= -E\left(\frac{\partial^2 \ln p(\mathbf{Y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \\
&= -E\left(\frac{\partial}{\partial \boldsymbol{\theta}}\left(\frac{y_1}{\theta_1}, \cdots, \frac{y_p}{\theta_p}\right)\right) \\
&= \begin{pmatrix} \frac{n}{\theta_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{n}{\theta_p} \end{pmatrix}
\end{aligned}$$

De donde podemos ver que la previa no informativa de Jeffreys para $\boldsymbol{\theta}$ está dada por

$$p(\boldsymbol{\theta}) \propto (\theta_1)^{-1/2} \cdots (\theta_p)^{-1/2}$$

la cual corresponde a una distribución $Dirichlet(1/2, \cdots, 1/2)$. El uso de esta distribución previa conduce a la distribución posterior $\boldsymbol{\theta} \mid \mathbf{Y} \sim Dirichlet(y_1 + 1/2, \cdots, y_p + 1/2)$, y la estimación posterior de cada θ_i viene dada por

$$\hat{\theta}_i = \frac{y_i + 1/2}{n + p/2}$$

la cual es muy similar a la estimación clásica de θ_i dada por y_i/n, especialmente cuando n es grande o p es pequeño.

Ejemplo 3.4.1. En este ejemplo se realiza un análisis bayesiano acerca de la intención de voto para una elección de la alcaldía de la ciudad de Bogotá del año 2011. El análisis electoral, en una primera instancia, se trata de conocer la probabilidad de éxito de un candidato, que aplicada a una población específica se traduce en la intención de voto hacia el candidato. Como hay varios candidatos en la disputa, entonces es conveniente suponer que el fenómeno puede ser descrito mediante el uso de una distribución multinomial. Como el parámetro en este caso es un vector de probabilidades, es adecuado suponer una distribución previa de tipo Dirichlet para este vector. Para este ejemplo, desarrollaremos un análisis básico con base en una primera encuesta realizada del 12 al 14 de agosto del 2011, en donde según el portal WEB de la revista Semana (`http://www.semana.com/nacion/articulo/petro-penalosa-empate-intencion-voto/245005-3`) se afirma que:

> Según la encuesta de Ipsos Napoleón Franco, hay un cabeza a cabeza (cada uno con el 22 %) entre los dos candidatos Peñalosa y Petro, Mockus es tercero, pero con notable diferencia: 12 %, seguido, muy cerca, por Gina Parody, con 9 %.

Con base en esta información, y teniendo en cuenta que hubo 604 respondientes, se afina la distribución previa que es Dirichlet con parámetros 133 (igual a 604*0.22), 133 (igual a 604*0.22), 72 (igual a 604*0.12) y 64 (igual a 604*0.09), para los candidatos Peñalosa, Petro, Mockus y Parody, respectivamente. Por otro lado, según la última encuesta electoral reportada por un medio de comunicación, correspondiente a la realizada por la firma Centro Nacional de Consultoría, entre el 30 de agosto y el primero de Septiembre, y publicada por el portal WEB de ElTiempo.com afirma que:

> (En 1000 respondientes) Peñalosa alcanza el 22 % de preferencia. Segundo aparece Gustavo Petro, con 17 %, en tercer lugar Antanas Mockus, con 12 %. El cuarto lugar es para la candidata Gina Parody, con 11 %.

Como se trata de la encuesta más reciente, supondremos que estos datos corresponden a la realización de una distribución multinomial. Es bien sabido que el análisis conjugado, señala que la distribución posterior del parámetro es de tipo Dirichlet, que en este ejercicio particular, tiene parámetros 353, 302, 192 y 164, para los candidatos Peñalosa, Petro, Mockus y Parody, respectivamente.

Otra pregunta de interés radica en comparar la intención de voto de los candidatos Peñalosa y Petro, pues son los que tienen mayor apoyo ciudadano. Los códigos `JAGS` para el análisis se presentan a continuación, donde se define un nuevo parámetro $\delta = \theta_1 - \theta_2$ (donde θ_1 y θ_2 denotan la intención de voto de Peñalosa y Petro, respectivamente), del cual se obtiene estimación e intervalo de credibilidad:

```
k=4; alpha=c(133,133,72,54)
y=c(220,170,120,110); n=sum(y)
MulNomial.model <- function(){
     y[1:k] ~ dmulti(theta[1:k],n)
    theta[1:k] ~ ddirch(alpha[1:k])
    delta <- theta[1]-theta[2]
}
```

```
MulNomial.data <- list("y","n","alpha","k")
MulNomial.param <- c("theta", "delta")
MulNomial.inits <- function(){
  list("theta"=c(0.3,0.3,0.2,0.2))
}

MulNomial.fit <- jags(data=MulNomial.data, inits=MulNomial.inits, MulNomial.param,
                      n.iter=10000, n.burnin=1000, model.file=MulNomial.model)

print(MulNomial.fit)
```

De los resultados obtenidos, vemos que la estimación bayesiana del vectro de inteciones de voto es $\hat{\boldsymbol{\theta}} = (34.9\,\%, 30\,\%, 19\,\%, 16.2\,\%)$, esto es, un resultado favorable para el candidato Peñalosa, con una ventaja de casi $5\,\%$ sobre el candidato Petro.

El mismo procedimiento se puede realizar en `R` usando los siguientes comandos:

```
k=4; alpha=c(133,133,72,54)
y=c(220,170,120,110); n=sum(y)
nsim <- 10000
theta.pos <- rdirichlet(nsim, y+alpha)
# Estimacion de intencion de voto para los candidatos
colMeans(theta.pos)

## [1] 0.35 0.30 0.19 0.16

# Ventaja de intencion de voto de Penalosa sobre Petro
mean(theta.pos[,1]-theta.pos[,2])

## [1] 0.049

# Intervalo de credibilidad para la ventaja de  Penalosa sobre Petro
quantile(theta.pos[,1]-theta.pos[,2], c(0.025,0.975))

##    2.5%     98%
## 0.00095 0.09878
```

Vemos que la estimación de $\boldsymbol{\theta}$ es similar a lo obtenido en `JAGS`. Ahora, para comparar la intención de voto de Peñalosa y Petro, se puede calcular la probabilidad $Pr(\theta_1 > \theta_2)$, tal como sigue:

```
# Probabilidad de que Penalosa obtenga mas votos que Petro
sum(theta.pos[,1]>theta.pos[,2])/nsim

## [1] 0.98
```

Observamos que la probabilidad de un triunfo de Peñalosa sobre Petro es muy cercana a 1.

3.5 Ejercicios

1. Realizar los desarrollos algebráicos necesarios para obtener la expresión (3.1.11).

2. En una muestra de variables con distribución $N(\theta, \sigma^2)$ donde θ y σ^2 son dependientes en la distribución previa, describa cómo es el procedimiento para el pronóstico para (1) una nueva variable aleatoria, y (2) la media de una nueva muestra $\bar{Y}^*$. Calcule el pronóstico y el intervalo de pronóstico para 20 nuevos pacientes en el ejemplo 3.1.2.

3. Para los datos de Student (1908) sobre el incremento de sueño producido por dos medicamentos soporíferos. Realizar un análisis bayesiano así como un análisis clásico para comparar los dos medicamentos. Utilizar $\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}$, $\boldsymbol{\mu} = (0, 1)'$ y $\Gamma = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}$. Compara las conclusiones obtenidas.

4. Para los datos de aumento de sueño utilizados en el ejemplo 3.3.1, realice los siguientes ejercicios

   (a) Manteniendo los valores para los demás parámetros, incrementa los valores de la matriz Γ y obtenga la estimación bayesiana de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$. qué conclusión puede obtener?

   (b) Manteniendo los valores para los demás parámetros, incrementa los valores del grado de libertad de la distribución previa inversa Wishart dejando invariante la estimación previa de $\boldsymbol{\Sigma}$ y obtenga la estimación bayesiana de $\boldsymbol{\theta}$ y $\boldsymbol{\Sigma}$. qué conclusión puede obtener?

5. En el ejemplo 3.4.1, utilice directamente la distribución posterior de $\boldsymbol{\theta}$ para calcular la estimación de los parámetros y un intervalo de credibilidad de cada uno de componentes de $\boldsymbol{\theta}$. Compare los resultados con los obtenidos con `JAGS`.

Capítulo 4

Modelos empíricos y jerárquicos

En las últimas décadas la formulación de modelos estadísticos ha evolucionado mucho En un principio, los modelos establecidos obedecían a reglas estándares que se suponían ciertas para toda la población. Sin embargo, el estado de la naturaleza de la mayoría de los problemas práctico no sigue una regla común para todos y cada uno de los elementos de una población aleatoria. De hecho el sentido común establece que para una misma población, pueden existir tendencias comunes entre diferentes miembros de la misma y la estructura de dispersión de los elementos puede obedecer comportamientos disímiles a través de éstos.

Lo anterior ha permitido que el investigador pueda proponer modelos que siguen comportamientos estructurales distintos y en algunos casos que se encuentran anidados en modelos más complejos. En el caso bayesiano, es claro que el momento de coyuntura en el cual el investigador no contempla un punto de retorno está dado en la formulación de la distribución previa para el vector de parámetros de interés $\boldsymbol{\theta}$. más aún, la influencia de la distribución previa en la resultante distribución posterior está dada por la asignación del vector de hiperparámetros $\boldsymbol{\eta}$ que parametriza la distribución previa. Cuando los valores exactos de los hiperparámetros se desconocen o cuando no se tiene plena certeza del comportamiento estructural de la distribución previa, entonces es necesario estimarlos pues de estos dependen los resultados en cualquier investigación de tipo causal. En otras palabras, una mala asignación de los valores de los hiperparámetros conduce a una distribución previa que no es acorde con la realidad y esto puede conllevar a su vez a que la distribución posterior no concuerde con la realidad, produciendo así resultados engañosos.

Siguiendo los fundamentos filosóficos de la estadística bayesiana, tener que estimar el vector de hiperparámetros envuelve al investigador en una paradoja cuya solución no siempre está dada por métodos bayesianos. En primer lugar, nótese la forma de la distribución previa del vector de parámetros de interés: $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. A simple vista se puede concluir que $\boldsymbol{\eta}$ hace parte de la distribución previa la cual, según la lógica de la filosofía bayesiana, involucra el conocimiento del investigador antes de la recolección de los datos. Por tanto la pregunta directa que surge es Por qué estimar algo que se debería suponer conocido?. En segundo lugar y si se concibe tal estimación, la otra pregunta natural es: Se deben utilizar los datos para estimar tales hiperparámetros?. Las posibles respuestas a las anteriores preguntas han creado toda una nueva corriente alterna a la bayesiana pura llamada ≪corriente bayesiana empírica≫[1] la cual utiliza los métodos de estimación puntual frecuentista para estimar estos hiperparámetros y por consiguiente definir la distribución previa del vector de parámetros de interés.

LADY TASTING TEA SOBRE EMPIRICAL Y BAYES

[1] Carlin & Louis (1996) menciona que el análisis empírico toma este nombre por dos razones: En primer lugar porque estima el vector de hiper-parámetros $\boldsymbol{\eta}$ con los datos observados, contradiciendo de alguna manera el espíritu y la filosofía de la corriente bayesiana radical. En segundo lugar, porque esta estimación se realiza con métodos frecuentistas ya sean paramétricos o no-paramétricos

Por supuesto, existe la contraparte teórica a la corriente empírica y es la llamada ≪corriente bayesiana jerárquica≫ la cual asume una posición totalmente bayesiana desde su concepción y establece un modelo posterior para los hiperparámetros.

Suponga entonces que la variable de interés sigue un modelo común a toda la población aunque parametrizado por parámetros que toman distintos valores para cada individuo y que está regido por la siguiente expresión

$$Y_i \sim p(Y_i \mid \theta_i)$$

4.1 análisis empírico

COLOCAR LOS TIPOS DE análisis: paramétricO Y NO paramétricO Y LOS PRINCIPALES RESULTADOS

Este enfoque, criticado por muchos bayesianos radicales, se centra en la escogencia de una estimación $\hat{\boldsymbol{\eta}}$ de $\boldsymbol{\eta}$ obtenida como el valor que hace maxima la verosimilitud marginal previa dada por

$$p(\mathbf{Y} \mid \boldsymbol{\eta}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \; d\boldsymbol{\theta} \tag{4.1.1}$$

Por lo tanto todo el andamiaje inferencial está supeditado a la distribución posterior estimada, $p(\boldsymbol{\theta} \mid Y, \hat{\boldsymbol{\eta}})$. Una vez que ésta está bien definida, el proceso de estimación puntual, estimación por intervalo y pruebas de hipótesis sigue su curso bayesiano idénticamente como en los capítulos anteriores.

En términos prácticos suponga que se tiene un modelo en dos etapas para cada una de las observaciones. Se asume que existen n observaciones que, si bien no conforman una muestra aleatoria, conservan la característica de intercambiebilidad y están definidas en los siguientes términos

$$Y_i \sim p(Y_i \mid \theta_i) \qquad i = 1, \ldots, n$$

La segunda etapa comienza con la asignación de una distribución[2] previa para los parámetros de interés θ_i.

$$\theta_i \sim p(\theta_i \mid \boldsymbol{\eta}) \qquad i = 1, \ldots, n$$

nótese que detrás de la asignación de la estructura probabilística para cada uno de los θ_i, se supone que éstos últimos determinan una muestra aleatoria de la distribución $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. El objetivo de este enfoque es encontrar estimadores que maximicen la verosimilitud marginal previa la cual, para este caso particular y considerando independencia marginal entre las observaciones y el vector de hiperparámetros, es

$$\begin{aligned} p(Y_i \mid \boldsymbol{\eta}) &= \int p(Y_i, \theta_i \mid \boldsymbol{\eta}) \; d\theta_i \\ &= \int p(Y_i \mid \theta_i, \boldsymbol{\eta}) p(\theta_i \mid \boldsymbol{\eta}) \; d\theta_i \\ &= \int p(Y_i \mid \theta_i) p(\theta_i \mid \boldsymbol{\eta}) \; d\theta_i \end{aligned} \tag{4.1.2}$$

De lo anterior, la verosimilitud marginal previa del vector de observaciones dada por la expresión

[2] En esta etapa la distribución previa no está completamente especificada puesto que se desconocen los hiperparámetros que la indexan.

(4.1.1) queda convertida en

$$p(Y \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} p(Y_i \mid \boldsymbol{\eta})$$
$$= \prod_{i=1}^{n} \int p(Y_i \mid \theta_i) p(\theta_i \mid \boldsymbol{\eta}) \, d\theta_i \tag{4.1.3}$$

A continuación, examninamos algunas distribuciones

4.1.1 Modelo Binomial-Beta

Suponga el siguiente modelo binomial (intercambiable) en una primera etapa

$$Y_i \mid \theta_i \sim Binomial(n_i, \theta_i) \qquad i = 1, \ldots, p$$

Para la segunda etapa, se supone una muestra aleatoria (independientes e idénticamente distribuidos) proveniente de una misma distribución tal que

$$\theta_i \sim Beta(\alpha, \beta) \qquad i = 1, \ldots, p$$

puesto que cada θ_i se encuentra en el intervalo $(0, 1)$ y es apropiado asignarle una distribución Beta.

análisis preliminar

Es bien sabido que la distribución posterior para cada uno de los parámetros de interés involucrados en el anterior contexto está dada por

$$\theta_i \mid Y_i \sim Beta(\alpha + Y_i, \beta + n_i - y_i)$$

para todo $i = 1, \cdots, p$. Sin embargo, como se desconoce totalmente el valor de los hiperparámetros α y β, entonces se debe encontrar una estimación de estos, $\hat{\alpha}$ y $\hat{\beta}$, respectivamente, para proseguir normalmente con la inferencia bayesiana, pero esta vez enfocados en la estimación de la distribución posterior dada por

$$\theta_i \mid Y_i \sim Beta(\hat{\alpha} + Y_i, \hat{\beta} + n_i - y_i)$$

Para tal fin, nótese que la esperanza y la varianza previa de θ_i están dadas por las siguientes expresiones

$$E(\theta_i) = \frac{\alpha}{\alpha + \beta} \tag{4.1.4}$$

$$Var(\theta_i) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{4.1.5}$$

De donde se tiene que

$$\alpha = E(\theta_i)(\alpha + \beta) \tag{4.1.6}$$

y también que

$$1 - E(\theta_i) = \frac{\beta}{\alpha + \beta} \tag{4.1.7}$$

por lo tanto

$$\beta = (1 - E(\theta_i))(\alpha + \beta) \tag{4.1.8}$$

y reemplazando (4.1.4) y (4.1.7) en (4.1.5) se concluye que

$$Var(\theta_i) = \frac{E(\theta_i)(1 - E(\theta_i))}{(\alpha + \beta + 1)}$$

por tanto

$$\alpha + \beta = \frac{E(\theta_i)(1 - E(\theta_i))}{Var(\theta_i)} - 1 \tag{4.1.9}$$

Con el anterior razonamiento, es posible encontrar los estimadores basados en el método frecuentista de los momentos los cuales corresponden a

$$\widehat{\alpha + \beta} = \frac{\bar{Y}(1 - \bar{Y})}{S^2} - 1 \tag{4.1.10}$$

Donde $\bar{Y}$ y S^2 es el promedio y la varianza de las cantidades $Y_1/n_1, Y_2/n_2, \ldots, Y_p/n_p$, respectivamente. Ahora, teniendo en cuenta que (4.1.6) y (4.1.8), se tiene que:

$$\hat{\alpha} = (\widehat{\alpha + \beta})\bar{Y} \tag{4.1.11}$$

$$\hat{\beta} = (\widehat{\alpha + \beta})(1 - \bar{Y}) \tag{4.1.12}$$

Con las anteriores estimaciones es posible ahora conectarlas a la distribución posterior de θ_i.

análisis legítimo

según Gelman, Carlin, Stern & Rubin (2003, p. 119), el anterior análisis no implica simplemente un punto de partida que da pie a la exploración de la idea de la estimación de los parámetros de la distribución posterior y, de ninguna manera, constituye un cálculo bayesiano puesto que no está basado en ningún modelo de probabilidad. Sin embargo, el análisis empírico de esta situación, hace uso de la esperanza y varianza condicional a la distribución beta de los parámetros θ_i $(i = 1, \ldots, p)$.

Para realizar este tipo de análisis, vamos a suponer que contamos con una variable Y, distribuida de forma binomial en n ensayos y con probabilidad de éxito θ. De esta manera, se tiene que el primer momento está dado por

$$\begin{aligned} E_{binom}\left(\frac{Y}{n}\right) &= E_{beta}\left(E_{binom}\left(\frac{Y}{n} \mid \theta\right)\right) \\ &= E_{beta}\left(\theta\right) \\ &= \frac{\alpha}{\alpha + \beta} \end{aligned} \tag{4.1.13}$$

Por otro lado, se tiene que la varianza, que es función del primer y segundo momento, está dada por

$$
\begin{aligned}
Var_{binom}\left(\frac{Y}{n}\right) &= E_{beta}\left(Var_{binom}\left(\frac{Y}{n} \mid \theta\right)\right) + Var_{beta}\left(E_{binom}\left(\frac{Y}{n} \mid \theta\right)\right) \\
&= E_{beta}\left(\frac{1}{n}\theta(1-\theta)\right) + Var_{beta}\left(\theta\right) \\
&= \frac{1}{n}E_{beta}(\theta) - \frac{1}{n}E_{beta}(\theta^2) + Var_{beta}\left(\theta\right) \\
&= \frac{1}{n}E_{beta}(\theta) - \frac{1}{n}Var_{beta}(\theta) - \frac{1}{n}(E_{beta}\theta)^2 + Var_{beta}\left(\theta\right) \\
&= \frac{n-1}{n}Var_{beta}(\theta) + \frac{1}{n}E_{beta}(\theta)(1 - E_{beta}(\theta)) \\
&= \frac{n-1}{n}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{1}{n}\frac{\alpha\beta}{(\alpha+\beta)^2} \\
&= \frac{1}{n}\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\left(\frac{n-1}{\alpha+\beta+1}+1\right) \\
&= \frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)\left(\frac{n-1}{\alpha+\beta+1}+1\right)
\end{aligned}
$$

De esta última expresión, y despejando $\alpha+\beta$, se tiene que

$$
\begin{aligned}
\alpha+\beta &= \frac{(n-1)E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)}{nVar_{binom}\left(\frac{Y}{n}\right) - E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)} - 1 \\
&= \frac{E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right) - Var_{binom}\left(\frac{Y}{n}\right)}{Var_{binom}\left(\frac{Y}{n}\right) - \frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)}
\end{aligned} \tag{4.1.14}
$$

Ahora, despejando α de la expresión (4.1.13) se tiene que

$$
\alpha = E_{binom}\left(\frac{Y}{n}\right)\frac{E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right) - Var_{binom}\left(\frac{Y}{n}\right)}{Var_{binom}\left(\frac{Y}{n}\right) - \frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)} \tag{4.1.15}
$$

además, también despejando β de (4.1.13) se tiene que

$$
\begin{aligned}
\beta &= \frac{\alpha\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)}{E_{binom}\left(\frac{Y}{n}\right)} \\
&= \frac{E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right) - Var_{binom}\left(\frac{Y}{n}\right)}{Var_{binom}\left(\frac{Y}{n}\right) - \frac{1}{n}E_{binom}\left(\frac{Y}{n}\right)\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)}\left(1 - E_{binom}\left(\frac{Y}{n}\right)\right)
\end{aligned} \tag{4.1.16}
$$

El anterior enfoque nos ha llevado a poder expresar los parámetros de interés en términos de $E_{binom}\left(\frac{Y}{n}\right)$, $Var_{binom}\left(\frac{Y}{n}\right)$ y n. Una vez que podamos estimar las anteriores cantidades, es posible realizar la inferencia bayesiana empírica de la manera correcta. Para lo anterior, es necesario observar al naturaleza de las observaciones que, aunque no representan una muestra aleatoria, sí son una sucesión de variables aleatorias intercambiables. Por lo anterior, y teniendo en cuenta que la inferencia se realiza con las

cantidades $Y_1/n_1, Y_2/n_2, \ldots, Y_p/n_p$, es posible proponer los siguientes estimadores

$$\hat{E}_{binom}\left(\frac{Y}{n}\right) = \bar{Y} \tag{4.1.17}$$

$$\hat{Var}_{binom}\left(\frac{Y}{n}\right) = S^2 \tag{4.1.18}$$

$$\hat{n} = \frac{1}{p}\sum_{i=1}^{p} n_i \tag{4.1.19}$$

Con base en lo anterior, unas estimaciones empíricas de los parámetros α y β son

$$\hat{\alpha} = \bar{Y}\left(\frac{\bar{Y}\left(1-\bar{Y}\right) - S^2}{S^2 - \frac{1}{\hat{n}}\bar{Y}\left(1-\bar{Y}\right)}\right) \tag{4.1.20}$$

y

$$\hat{\beta} = (1-\bar{Y})\frac{\bar{Y}\left(1-\bar{Y}\right) - S^2}{S^2 - \frac{1}{\hat{n}}\bar{Y}\left(1-\bar{Y}\right)} \tag{4.1.21}$$

respectivamente. Cuando la cantidad de ensayos n_i es diferente en cada experimento, existen otras formas de obtener estimaciones para los parámetros α y β (Carlin & Louis 1996, p. 81).

Ejemplo 4.1.1. En el ejemplo 2.2.1 se estudió datos que corresponden al porcentaje de bateo en 18 jugadores profesionales de beisbol.

```
library(pscl)
data(EfronMorris)
attach(EfronMorris)

## The following objects are masked _by_ .GlobalEnv:
##
##    n, p, y

y <- p # Porcentaje de bateo de los 18 jugadores
y.bar <- mean(y)
S2 <- var(y)
n.hat <- mean(n)
alfa <- y.bar*(y.bar*(1-y.bar)-S2)/(S2-y.bar*(1-y.bar)/n.hat)
beta <- (1-y.bar)*(y.bar*(1-y.bar)-S2)/(S2-y.bar*(1-y.bar)/n.hat)
alfa

## [1] NA

beta

## [1] NA
```

De donde podemos concluir que la distribución previa para cada θ_i es la distribución $Beta(57, 158)$, observamos que la esperanza de esta distribución coincide con el porcentaje de bateo promedio de los datos de los 18 jugadres. Ahora, podemos calcular los parámetros de la distribución para cada θ_i con $i = 1, \cdots, 18$ como sigue:

```
alfa.new <- alfa + p*n
beta.new <- beta + (1-p)*n
head(alfa.new)

## [1] NA

head(beta.new)

## [1] NA
```

Y así podemos realizar inferencias para cualquier θ_i. Por ejemplo, para primer jugador, Roberto Clemente, la distribución posterior para el porcentaje de bateo es $Beta(184, 398)$, por consiguiente la estimación para el porcentaje de bateo de este jugador es $184/(184 + 398) = 0.3162$ y un intervalo de credibilidad está dada por $(0.279, 0.354)$.

4.1.2 Modelo Poisson-Gamma

Suponga el siguiente modelo de Poisson intercambiable

$$Y_i \mid \theta_i \sim Poisson(\theta_i)$$

para $i = 1, \ldots, n$. Y considerando que cada θ_i debe ser estrictamente positivo, entonces la distribución del parámetro θ_i es

$$\theta_i \sim Gamma(\alpha, \beta)$$

Donde α y β son hiperparámetros desconocidos. Utilizando el resultado 2.4.1, se tiene que la distribución posterior de cada parámetro θ_i está dada por

$$\theta_i \mid \mathbf{Y} \sim Gamma\left(\sum_{i=1}^{n} Y_i + \alpha, \beta + n\right)$$

Por supuesto, la distribución anterior no es útil a no ser que los hiperparámetros puedan ser estimados. Para realizar esta estimación, el enfoque empírico sugiere utilizar el método de los momentos. Para esto, nótese que el primer momento está dado por

$$\begin{aligned} E_{Poisson}(Y_i) &= E_{Gamma}\left(E_{Poisson}(Y_i \mid \theta_i)\right) \\ &= E_{Gamma}\left(\theta_i\right) \\ &= \frac{\alpha}{\beta} \end{aligned} \tag{4.1.22}$$

Mientras que la varianza, función del primer y segundo momento, está dada por

$$\begin{aligned} Var_{Poisson}(Y_i) &= E_{Gamma}\left(Var_{Poisson}(Y_i \mid \theta_i)\right) + Var_{Gamma}\left(E_{Poisson}(Y_i \mid \theta_i)\right) \\ &= E_{Gamma}\left(\theta_i\right) + Var_{Gamma}\left(\theta_i\right) \\ &= \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} \\ &= \frac{\alpha}{\beta^2}(\beta + 1) \end{aligned} \tag{4.1.23}$$

Ahora, siguiendo el enfoque del método de los momentos, es claro que la expresión (4.1.22) puede ser estimada con la media muestral, $\bar{Y}$; mientras que la expresión (4.1.23) puede ser estimada con la varianza muestral, S^2. Por otro lado, al dividir estas expresiones se tiene que

$$\frac{\frac{\alpha}{\beta^2}(\beta+1)}{\frac{\alpha}{\beta}} = 1 + \frac{1}{\beta} \tag{4.1.24}$$

y, siguiendo un razonamiento similar, esta última expresión es estimada por $S^2/\bar{Y}$. Por tanto, un estimador del método de los momentos para β es

$$\hat{\beta} = \frac{1}{\frac{S^2}{\bar{Y}} - 1} = \frac{\bar{Y}}{S^2 - \bar{Y}} \tag{4.1.25}$$

De la expresión (4.1.12), se nota que $\alpha = \beta E_{Poisson}(Y_i)$. Por tanto, un estimador del método de los momentos para α es

$$\hat{\alpha} = \hat{\beta}\bar{Y} = \frac{\bar{Y}^2}{S^2 - \bar{Y}} \tag{4.1.26}$$

De lo anterior, se tiene que, siguiendo el enfoque bayesiano empírico, la distribución posterior para θ_i está dada por

$$\theta_i \mid \mathbf{Y} \sim Gamma\left(\sum_{i=1}^{n} Y_i + \hat{\alpha}, \hat{\beta} + n\right)$$

y la obtención de estimación de θ_i se sigue lo expuesto en capítulos anteriores.

Ejemplo 4.1.2. Retomamos los datos del ejemplo 2.4.1 sobre la ocurrencia de accidentes de tránsito relacionados con conductores en estado de embriaguez. En los siguientes códigos mostramos el procedmiento para calcular $\hat{\alpha}$ y $\hat{\beta}$ y las estimaciones obtenidas para θ que denota el número prommedio diario de accidentes.

```
Trans <- c(22, 9, 9, 20, 10, 14, 11, 14, 11, 11, 19, 12, 8, 9, 16, 8, 13, 8, 14, 12,
           14, 11, 14, 13, 11, 14, 13, 11, 7, 12 )
y.bar <- mean(Trans); S2 <- var(Trans); n <- length(Trans)
beta <- y.bar/(S2-y.bar)
alpha <- beta*y.bar
alpha/beta

## [1] 12

(sum(Trans)+alpha)/(beta+n)

## [1] 12

y.bar

## [1] 12

# Intervalo con enfoque jer\'arquico
qgamma(c(0.025,0.975), shape=sum(Trans)+alpha, rate=beta+n)
```

```
## [1] 12 13

# Intervalo con enfoque cl??sico
c(qchisq(alpha/2, df=2*y.bar*n)/(2*n), qchisq(1-alpha/2, df=2*(y.bar*n+1))/(2*n))

## Warning in qchisq(alpha/2, df = 2 * y.bar * n): NaNs produced
## Warning in qchisq(1 - alpha/2, df = 2 * (y.bar * n + 1)): NaNs produced

## [1] NaN NaN

# Intervalo con previa no informativa de Jeffreys
qgamma(c(0.025,0.975), shape=sum(Trans)+0.5, rate=n)

## [1] 11 14
```

En primer lugar, observamos que el no es posible calcular el intervalo de confianza para θ puesto que el grado de libertad de la distribución χ^2 es muy grande. Por otro lado, vemos que la estimación bayesina obtenida de esta forma es exactamente igual a la estimaciónclásica $\bar{y}$, con un intervalo de credibilidad de menor longitud que el intervalo de confianzaclásico. también se calculó el intervalo de credilidad para θ usando una previa no informativa de Jeffreys, el cual es más ancho que el obtenido con el enfoque empírico.

4.1.3 Modelo Normal-Normal

Uno de los modelos más utilizados en la aplicaciones prácticas se da cuando la distribución común de los datos es la distribución normal. Considere el siguiente modelo en dos etapas en donde cada una de las observaciones se supone intercambiable y para la primera etapa se tiene que

$$Y_i \mid \theta_i \sim Normal(\theta_i, \sigma^2) \qquad i = 1, \ldots, n$$

en donde el parámetro σ^2 se supone conocido. En la segunda etapa, la distribución previa para los parámetros de interés θ_i es

$$\theta_i \mid \mu \sim Normal(\mu, \tau^2) \qquad i = 1, \ldots, n$$

en donde el parámetro τ^2 se supone conocido.Para poder proseguir con el análisis empírico bayesiano, podemos calcular la esperanza de la variable Y_i similar a como se hizo en las secciones anteriores y encontrar que μ se puede calcular como $\bar{y}$. Otra forma de hallar el valor de μ es encontrar la estimación de máxima verosimilitud de μ escribiendo la densidad de Y_i como función de μ, el siguiente resultado nos da la expresión.

Resultado 4.1.1. *La verosimilitud marginal previa de una observación condicional al hiperparámetro μ es*

$$Y_i \mid \mu \sim Normal(\mu, \sigma^2 + \tau^2) \qquad i = 1, \ldots, n$$

Prueba. Desarrollando la expresión (4.1.2) se tiene que

$$
\begin{aligned}
p(Y_i \mid \mu) &= \int p(Y_i \mid \theta_i) p(\theta_i \mid \mu)\, d\theta_i \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\theta_i)^2}{\sigma^2}\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2}\frac{(\theta_i-\mu)^2}{\tau^2}\right\} d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{1}{2}\frac{\tau^2(\theta_i-y_i)^2+\sigma^2(\theta_i-\mu)^2}{\sigma^2\tau^2}\right\} d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left[\theta_i^2(\tau^2+\sigma^2)-2\theta_i(y_i\tau^2+\mu\sigma^2)+\tau^2y_i^2+\sigma^2\mu^2\right]\right\} d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[\theta_i^2-2\theta_i\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}+\frac{\tau^2y_i^2+\sigma^2\mu^2}{\tau^2+\sigma^2}\right]\right\} d\theta_i \\
&= \int \frac{1}{2\pi\sqrt{\sigma^2\tau^2}} \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left[(\theta_i-\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2})^2-\left(\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)^2+\frac{\tau^2y_i^2+\sigma^2\mu^2}{\tau^2+\sigma^2}\right]\right\} d\theta_i \\
&= \int \frac{1}{\sqrt{2\pi}\sqrt{\frac{\sigma^2\tau^2}{\tau^2+\sigma^2}}} \exp\left\{-\frac{1}{2\frac{\sigma^2\tau^2}{\tau^2+\sigma^2}}\left(\theta_i-\frac{y_i\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}\right)^2\right\} d\theta_i \\
&\times \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2\tau^2}}\sqrt{\frac{\sigma^2\tau^2}{\tau^2+\sigma^2}} \exp\left\{\frac{(y_i\tau^2+\mu\sigma^2)^2}{2\sigma^2\tau^2(\tau^2+\sigma^2)}-\frac{\tau^2y_i^2+\sigma^2\mu^2}{2\sigma^2\tau^2}\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[\frac{(\tau^2y_i^2+\sigma^2\mu^2)(\tau^2+\sigma^2)}{\sigma^2\tau^2}-\frac{(y_i\tau^2+\mu\sigma^2)^2}{\sigma^2\tau^2}\right]\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[\frac{\tau^2\sigma^2y_i^2+\sigma^2\tau^2\mu^2-2y_i\mu\tau^2\sigma^2}{\sigma^2\tau^2}\right]\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[y_i^2+\mu^2-2y_i\mu\right]\right\} \\
&= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\left[y_i-\mu\right]^2\right\}
\end{aligned}
$$

la cual corresponde a la función de distribución de una variable aleatoria con densidad $Normal(\mu, \sigma^2+\tau^2)$ ■

Del anterior resultado, y teniendo la independencia de las variables, se tiene que la verosimilitud marginal previa del vector de observaciones $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ condicionado al hiperparámetro μ es

$$
p(\mathbf{Y} \mid \mu) = \left(\frac{1}{2\pi(\tau^2+\sigma^2)}\right)^{n/2} \exp\left\{-\frac{1}{2(\tau^2+\sigma^2)}\sum_{i=1}^{n}(y_i-\mu)^2\right\}
$$

El objetivo del enfoque empírico bayesiano es encontrar una estadística que maximice la anterior expresión. No es difícil notar que un estimador de máxima verosimilitud para μ está dado por la media muestral $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Con este estimador para el hiperparámetro se considera que las distribuciones previa y posterior del parámetro de interés quedan totalmente definidas y es posible continuar con el análisis bayesiano común.

Y qué pasa con σ^2, si encontramos el valor de σ^2 que maximiza a $p(\mathbf{Y} \mid \mu)$, entonces me da que $\tau^2 = S^2 - \sigma^2$ (lo mismo me da si calculo $Var(Y_i)$), pero eso en la práctica puede dar negativo.

Ejemplo 4.1.3. Los datos del ejemplo 2.6.1 muestran el grosor de 12 láminas de vidrio, donde la media muestral es 3.18cm, la varianza teórica es $\sigma^2 = 0.1cm^2$, mientras que la varianza muestral es

de $s^2 = 0.1068cm^2$. Usando los resultados encontrados en esta sección, la distribución previa para el grosor promedio de las láminas está dada por $\theta \sim Normal(\mu = 3.18cm, \tau^2 = 0.1068cm^2)$.

Completar el ejemplo según los anteriores textos en rojo.

4.2 análisis jerárquico

En esta parte, consideramos el análisis jerárquico donde se asigma distribuciones de probabilidad también a los hiperparámetros. Consideramos una muestra aleatoria $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ parametrizada por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ cuya función de verosimilitud está dada por

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(Y_i \mid \theta_i) \tag{4.2.1}$$

Por otro lado, suponga que la distribución previa del parámetro de interés θ_i está parametrizada por un vector de hiperparámetros $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)$ tal que la distribución previa de cada θ_i queda denotada por $p(\theta_i \mid \boldsymbol{\eta})$.

De lo anterior, y suponiendo que existe intercambiabilidad entre cada uno de los parámetros de interés, la distribución previa del vector de parámetros $\boldsymbol{\theta}$, parametrizada por $\boldsymbol{\eta}$ está dada por

$$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} p(\theta_i \mid \boldsymbol{\eta}) \tag{4.2.2}$$

Por tanto, es posible formular una distribución previa conjunta para $\boldsymbol{\theta}, \boldsymbol{\eta}$ que al igual que en capítulos anteriores, teniendo en cuenta el espíritu jerárquico y dependiente, vendría dada por

$$p(\boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \tag{4.2.3}$$

Luego, la distribución marginal previa del vector de parámetros de interés viene dada por

$$\begin{aligned} p(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\eta}) \; d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \; d\boldsymbol{\theta} \\ &= \int \cdots \int \prod_{i=1}^{n} p(\theta_j \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \; d\theta_1 \cdots d\theta_n \end{aligned}$$

Con esta formulación, y suponiendo que las observaciones son condicionalmente independientes del vector de hiperparámetros $\boldsymbol{\eta}$[3],la distribución posterior conjunta para $\boldsymbol{\theta}, \boldsymbol{\eta}$ es

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ &= p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \\ &= p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \end{aligned} \tag{4.2.4}$$

nótese que tanto para la distribución previa como para la distribución posterior de los parámetros, se supone conocido la distribución marginal previa de $\boldsymbol{\eta}$, $p(\boldsymbol{\eta})$, y también la distribución previa de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\eta}$, $p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$. La anterior formulación es acorde con la filosofía jerárquica pues supone

[3]Esta suposición tiene como base que las observaciones sólo dependen de $\boldsymbol{\eta}$ a través del vector de parámetros de interés $\boldsymbol{\theta}$.

relaciones de dependencia en distintos niveles. Conociendo el comportamiento estructural de $\boldsymbol{\eta}$ se puede conocer el comportamiento estructural de $\boldsymbol{\theta}$. Gelman, Carlin, Stern & Rubin (2003) afirma que cuando no se tiene certeza acerca del comportamiento de $\boldsymbol{\eta}$ se debe utilizar una distribución previa no informativa aunque siempre se debe tener alguna sospecha acerca del espacio paramétrico al cual sea posible restringirlos.

En términos de estimación, los siguientes pasos son esenciales para realizar un análisis bayesiano propiamente dicho (Gelman, Carlin, Stern & Rubin 2003):

1. Escribir la distribución posterior de $\boldsymbol{\theta}, \boldsymbol{\eta}$ de forma no normalizada como en la expresión (4.2.4).
2. Determinar analíticamente la distribución posterior de $\boldsymbol{\theta}$ condicional a $\boldsymbol{\eta}, \mathbf{Y}$, utilizando la siguiente regla
$$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, \mathbf{Y}) \propto p(\boldsymbol{\theta}, \underbrace{\boldsymbol{\eta}}_{fijo} \mid \mathbf{Y})$$
Es decir, los términos que no dependan de $\boldsymbol{\theta}$ pueden ser introducidos en la constante de proporcionalidad.
3. Determinar la distribución posterior de $\boldsymbol{\eta}$, utilizando alguna de las siguientes expresiones (se debe escoger la más conveniente dependiendo del contexto del problema):
$$p(\boldsymbol{\eta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta}) \tag{4.2.5}$$
$$p(\boldsymbol{\eta} \mid \mathbf{Y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{Y}) d\boldsymbol{\theta} \tag{4.2.6}$$
$$p(\boldsymbol{\eta} \mid \mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{Y})}{p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, \mathbf{Y})} \tag{4.2.7}$$
4. Por medio de $p(\boldsymbol{\eta} \mid \mathbf{Y})$ encontrar una estimación para $\boldsymbol{\eta}$.
5. Recurriendo a $p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, \mathbf{Y})$ encontrar una estimación para $\boldsymbol{\theta}$.

A continuación, ilustramos la implmentación del anteior procedimiento en datos de distintas naturalezas.

4.2.1 Modelo Binomial

Suponga el mismo modelo binomial de la sección 4.1.1 dado por

$$Y_i \sim Binomial(n_i, \theta_i) \quad \text{con } i = 1, \cdots, n$$

en donde la distribución de los parámetros de interés es tal que

$$\theta_i \sim Beta(\alpha, \beta)$$

Para esta situación el análisis bayesiano propiamente dicho requiere el planteamiento de un modelo que contemple el comportamiento estructural tanto del vector de parámetros $\boldsymbol{\theta}$ como de los hiperparámetros α, β. Por lo tanto, suponiendo que existe total ignorancia acerca del comportamiento estructural de los hiperparámetros, la distribución previa marginal para los hiperparámetros es no informativa (Gelman, Carlin, Stern & Rubin 2003), ésta está dada por

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

Suponiendo que los parámetros de interés θ_i $(i = 1, \ldots, n)$ conforman una muestra aleatoria, entonces su distribución previa es

$$\begin{aligned} p(\boldsymbol{\theta} \mid \alpha, \beta) &= \prod_{i=1}^{n} p(\theta_i \mid \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \end{aligned}$$

Por último, teniendo en cuenta que la distribución de las observaciones es intercambiable, entonces es posible definir la verosimilitud de la muestra como una productoria tal que

$$\begin{aligned} p(\mathbf{Y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(Y_i \mid \theta_i) \\ &\propto \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \end{aligned}$$

De esta manera, siguiendo la expresión (4.2.4), la distribución posterior conjunta estaría dada por

$$p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{Y}) \propto (\alpha+\beta)^{-5/2} \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$$

Utilizando la regla del condicionamiento, la distribución posterior del vector de parámetros de interés condicionado a los hiperparámetros y a los datos observados es

$$\begin{aligned} p(\boldsymbol{\theta} \mid \alpha, \beta, \mathbf{Y}) &\propto p(\boldsymbol{\theta}, \underbrace{\alpha, \beta}_{fijos} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \\ &\propto \prod_{i=1}^{n} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1} \end{aligned}$$

De donde se concluye que la distribución posterior para el vector de parámetros de interés es

$$\theta_i \mid \alpha, \beta, Y_i \sim Beta(\alpha + Y_i, \beta + n_i - y_i)$$

Por supuesto, la anterior distribución no es útil frente al desconocimiento de los hiperparámetros que deben ser estimados posterior, en este caso particular, utilizando la expresión (4.2.7) la cual da como resultado

$$\begin{aligned} p(\alpha, \beta \mid \mathbf{Y}) &= \frac{p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{Y})}{p(\boldsymbol{\theta} \mid \alpha, \beta, \mathbf{Y})} \\ &\propto \frac{(\alpha+\beta)^{-5/2} \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}}{\prod_{i=1}^{n} \frac{\Gamma(\alpha+y_i+\beta+n_i-y_i)}{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}} \\ &= (\alpha+\beta)^{-5/2} \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n_i)} \end{aligned}$$

Aunque la anterior distribución no tiene una forma cerrada o conocida, es posible simular valores provenientes de ésta utilizando el método de la grilla. Una vez que se tienen las observaciones simuladas, entonces se encuentra un estimador para los hiperparámetros y con estos, la distribución posterior del vector de parámetros de interés queda correctamente definida.

En R, una función que calcula la probabilidad posterior para los hiperparámetros está dada por

```
post<-function(a,b,n,y){
   P1<- gamma(a+b)
   P2<- gamma(a)*gamma(b)
   P3<- gamma(a+y)*gamma(b+n-y)
   P4<- gamma(a+b+n)
   (a+b)^(-5/2)*prod((P1/P2)*(P3/P4))
 }
```

Para implementar el método de la grilla, se debe tener en cuenta que como la distribución es bivariada entonces la grilla debe estar contenida en R^2. En R, una función que devuelve una grilla bivariada está dada por el siguiente código

```
 grilla<-function(a,b){
   A<-seq(1:length(a))
   unoA <-rep(1,length(A))
   B<-seq(1:length(b))
   unoB <-rep(1,length(B))
   P1<-kronecker(A,unoB)
   P2<-kronecker(unoA,B)
   grid<-cbind(a[P1],b[P2])
   return(grid)
 }
```

Gelman, Carlin, Stern & Rubin (2003, p.118) presentan datos que corresponden a 70 grupos de ratones de laboratorio, cada grupo tiene entre 10 y 52 ratones y se registran el número de ratones por grupo que desarrolle un tipo específico de tumor. El parámetro de interés es la probabilidad de desarrollar dicho tumor. En primer lugar se creó una grilla bivariada contenida entre $\{1, 1.5 \ldots, 49.5, 50\}^2$ (el superíndice denota el producto cartesiano) y para cada punto se calculó la respectiva probabilidad dada por la distribución posterior.

```
n<-c(20,19,18,20,20,20,23,20,18,18,10,13,48,19,20,18,25,49,48,19,22,20,17,24,19,50,
     19,20,20,20,23,46,20,19,20,20,20,20,27,20,22,20, 20,20,20,17,20,46,52,19,20,20,
     49,20,49,47,19,19,20,47,20,20,46,19,19,20,20,20,20,24)
 y<-c(0,0,1,2,3,4,6,0,0,1,1,2,10,5,0,0,2,5,9,4,6,0,0,2,2,10,4,6,0,1,2,5,4,4,6,0,1,2,
      3,4,5,6,0,1,2,2,4,11,16,0,1,2,7,4,12,15,0,1,2,7,4,5,15,0,1,2,3,4,5,9)

 a.grid<-seq(1,50,by=0.5)
 b.grid<-seq(1,50,by=0.5)
 ab.grid<-grilla(a.grid,b.grid)
 N.grid<-dim(ab.grid)[1]

 p.ab <- rep(NA, N.grid)
 for(j in 1:N.grid){
   p.ab[j] <- post(ab.grid[j,1], ab.grid[j,2], n, y)
 }
```

Luego, se utilizó la función `sample` para generar una muestra aleatoria de tamaño $n = 1000$ proveniente de la distribución posterior normalizada de los hiperparámetros

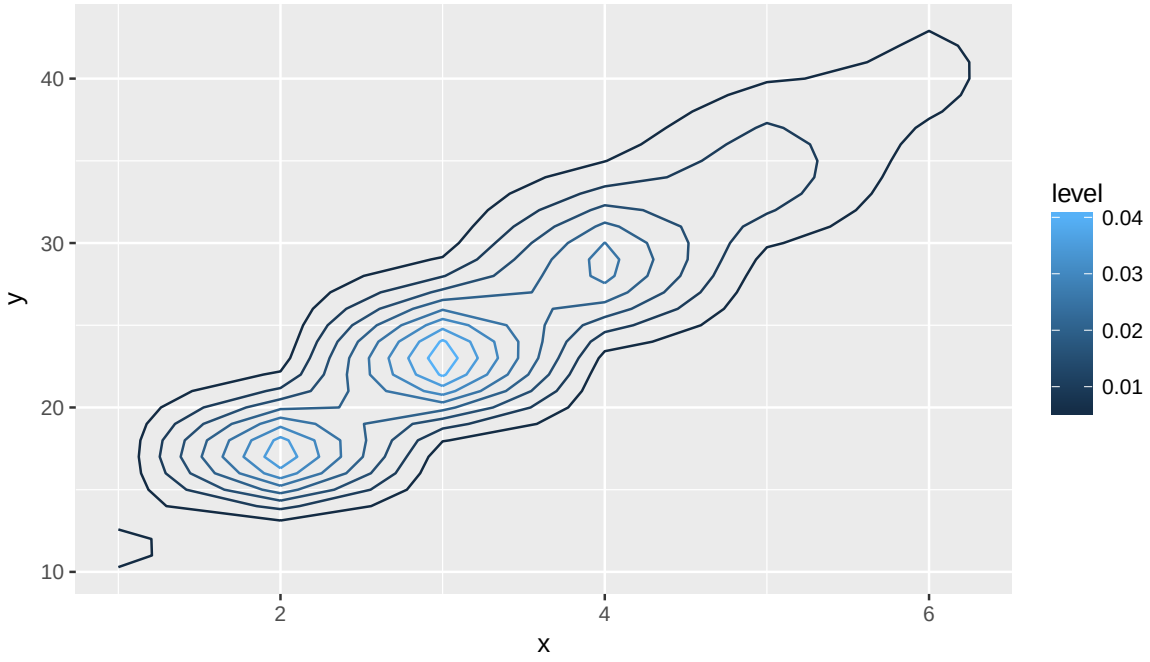
```
p.ab<-as.vector(p.ab/sum(p.ab))
sum(p.ab)

## [1] 1

r.post<-sample(N.grid,5000,prob=p.ab,replace=T)
rab.post<-ab.grid[r.post,]
ra.post<-rab.post[,1]
rb.post<-rab.post[,2]
```

El objeto `rab.post` es una matriz de dos columnnas y cinco mil filas. Cada fila contiene una observación simulada de la distribución posterior; por tanto, `rab.pos` contiene cinco mil duplas simuladas. A continuación, es posible obtener estimaciones puntuales posterior para el vector de hiperparámetros; teniendo en cuenta el criterio de mínima pérdida cuadrática, estas estimaciones son $(\hat{\alpha}, \hat{\beta})' = (2.3675, 14.2862)'$. De la misma manera, también es posible obtener intervalos de credibilidad al 95 %.

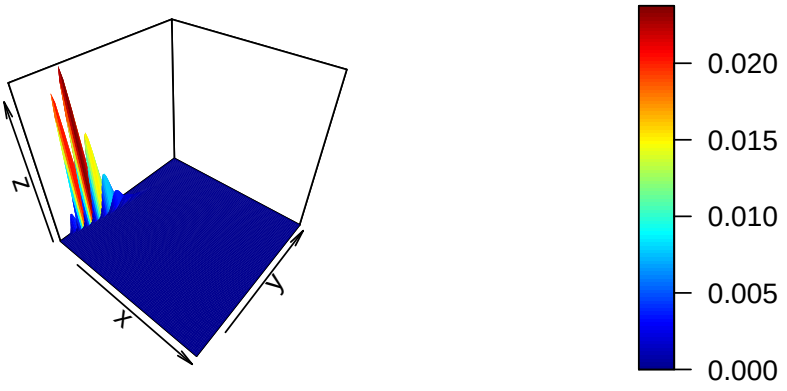
```
mean(ra.post)

## [1] 2.4

mean(rb.post)

## [1] 14

quantile(ra.post,c(0.025,0.975))

## 2.5%  98%
##  1.0  4.5

quantile(rb.post,c(0.025,0.975))

## 2.5%  98%
##    7   28
```

Aunque el objetivo primario del análisis jerárquico es obtener una estimación bayesiana de los hiperparámetros para conectarla directamente a la distribución posterior de cada uno de los parámetros de interés $\theta_i, \quad i = 1, \ldots, n$, es posible preguntarse acerca de la forma estructural de la distribución posterior de los hiperparámetros. De esta manera, un primer acercamiento gráfico se presenta cuando se genera el contorno bivariado para la distribución, se puede notar que la distribución posterior conjunta para $(\alpha, \beta)'$ no tiene una forma conocida y tiene varios picos, justo como se ve en la en la figura XXXX. La forma de la distribución en tercera dimensión, considerada la figura XXXX, comprueba que, en efecto, esta distribución debe ser tratada en conjunto. Por otra parte, se resalta la potencia del método de la grilla que permite simular observaciones bivariadas de una distribución compleja como la desarrollada acá. El código computacional para generar estas gráficas se presenta a continuación.

```
a<-a.grid
b<-b.grid

mat<-matrix(NA, nrow=length(a), ncol=length(b))
for(i in 1:length(a)){
for(j in 1:length(b)){
  mat[i,j]<-post(a[i],b[j],n,y)
}
}

mat<-mat/(sum(mat))
# gr\'afica de contorno
 mat3d <- melt(mat)
 names(mat3d) <- c("x", "y", "z")
 v <- ggplot(mat3d, aes(x, y, z = z))
 v + stat_contour(aes(colour = ..level..))
```




```
# gr\'afica de perspectiva
  persp3D(z=mat, x=a, y=b)
```



4.2.2 Modelo Poisson

Ahora, suponga el modelo Poisson para los datos, dado por

$$Y_i \mid \theta_i \sim Poisson(\theta_i)$$

donde los Y_i forman una sucesión de variables aleatorias intercambiables y con cada parámetro θ_i $(i = 1, \ldots, n)$ distribuido como

$$\theta_i \mid (\alpha, \beta) \sim Gamma(\alpha, \beta)$$

donde α y β son hiperparámetros desconocidos.Como estos hiperparámetros son positivos ambos, es razonable asignarles la distribución *Gamma* tales que

$$\alpha \sim Gamma(a, b)$$
$$\beta \sim Gamma(c, d)$$

Usualmente los parámetros a, b, c y d son conocidos y son tales que las distribuciones de α y β sean planas o no-informativas. De esta manera, el enfoque bayesiano jerárquico plantea que se debe realizar la inferencia conjunta para el vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ y para $(\alpha, \beta)'$. Con base en lo anterior, la distribución posterior de los parámetros de interés toma la siguiente forma

$$\begin{aligned} p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{Y}) &\propto \prod_{i=1}^{n} p(Y \mid \theta_i) p(\theta_i \mid \alpha, \beta) p(\alpha) p(\beta) \\ &\propto \prod_{i=1}^{n} \frac{e^{-\theta_i} \theta_i^{y_i}}{y_i!} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta_i^{\alpha-1} e^{-\beta\theta_i} e^{-\alpha b} \alpha^{a-1} e^{-\beta d} \beta^{c-1} \end{aligned}$$

Como es usual, y acudiendo a la anterior distribución posterior conjunta, se utilizará la técnica del condicionamiento sucesivo para encontrar las distribuciones posterior marginales de cada uno de los

parámetros de interés. En este orden de ideas, se tiene que para cada θ_i con $i = 1, \ldots, n$, la distribución posterior marginal está dada por

$$\begin{aligned} p(\theta_i \mid \alpha, \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_n, \mathbf{Y}) &\propto p(\theta_i, \underbrace{\alpha, \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_n}_{fijos} \mid \mathbf{Y}) \\ &\propto e^{-\theta_i} \theta_i^{y_i} \theta_i^{\alpha-1} e^{-\beta\theta_i} \\ &= \exp\{-\theta_i(\beta+1)\} \theta^{y_i+\alpha-1} \end{aligned}$$

Con base en lo anterior, se tiene que la distribución posterior para cada parámetro θ_i es

$$\theta_i \mid \alpha, \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \theta_n, \mathbf{Y} \sim Gamma(y_i + \alpha, \beta + 1)$$

Para el hiperparámetro α, se tiene que la distribución posterior marginal está dada por

$$\begin{aligned} p(\alpha \mid \beta, \boldsymbol{\theta}, \mathbf{Y}) &\propto p(\alpha, \underbrace{\beta, \boldsymbol{\theta}}_{fijos} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} \frac{\beta^\alpha \theta_i^{\alpha-1}}{\Gamma(\alpha)} \alpha^{a-1} \exp\{-\alpha b\} \end{aligned}$$

La anterior distribución no tiene una forma conocida y es necesario utilizar métodos numéricos para simular observaciones de provenientes de ésta. Para esto es posible utilizar el método de la grilla.

Por último, la distribución posterior marginal del hiperparámetro β se encuentra, similarmente mediante el condicionamiento sucesivo, de la siguiente forma

$$\begin{aligned} p(\beta \mid \alpha, \boldsymbol{\theta}, \mathbf{Y}) &\propto p(\beta, \underbrace{\alpha, \boldsymbol{\theta}}_{fijos} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} \beta^\alpha \exp\{-\beta\theta_i\} \beta^{c-1} \exp\{-\beta d\} \\ &= \beta^{n(\alpha+c-1)} \exp\left\{-\beta\left(nd + \sum_{i=1}^{n} \theta_i\right)\right\} \end{aligned}$$

Por lo tanto, se concluye que la distribución posterior para el hiperparámetro β es

$$\beta \mid \alpha, \boldsymbol{\theta}, \mathbf{Y} \sim Gamma\left(n(\alpha + c - 1) + 1, nd + \sum_{i=1}^{n} \theta_i\right)$$

Para realizar la inferencia bayesiana jerárquica para los parámetros de interés se deben fijar valores iniciales para cada parámetro y mediante simulación renovarlos hasta obtener convergencia. Por ejemplo, un posible camino para obtener convergencia en la simulación se describe a continuación:

- Fijar valores iniciales para α y β.
- Con los anteriores valores simular una observación para cada distribución posterior de los parámetros θ_i $(i = 1, \ldots, n)$.
- Con estos valores de θ_i y el valor inicial de β, simular una observación de la distribución posterior de α.
- Con los valores de θ_i y la anterior observación de α, simular un nuevo valor para β.

- Repetir el anterior proceso hasta lograr convergencia.

Dado que las distribuciones posterior de los parámetros θ_i $(i = 1, \ldots, n)$ y del hiperparámetro β están ligadas a la distribución Gamma, la simulación para estos parámetros es fácil. Sin eambrgo, como la distribución posterior marginal de α no tiene una forma cerrada, es necesario implementar un código propio en `R` que permita simular un valor proveniente de esta distribución. Es posible utilizar el método de la grilla que, en este caso, es univariado pues se trata de un sólo hiperparámetro. Con base en lo anterior, se tiene la siguiente función que reproduce esta distribución no conocida.

```
post <- function(theta, alpha, beta, a, b){
  P1 <- beta^alpha * (theta^(alpha-1)) / gamma(alpha)
  P2 <- alpha^(a-1)
  P3 <-  exp(-alpha*beta)
  res <- prod(P1)*P2*P3
  res
}
```

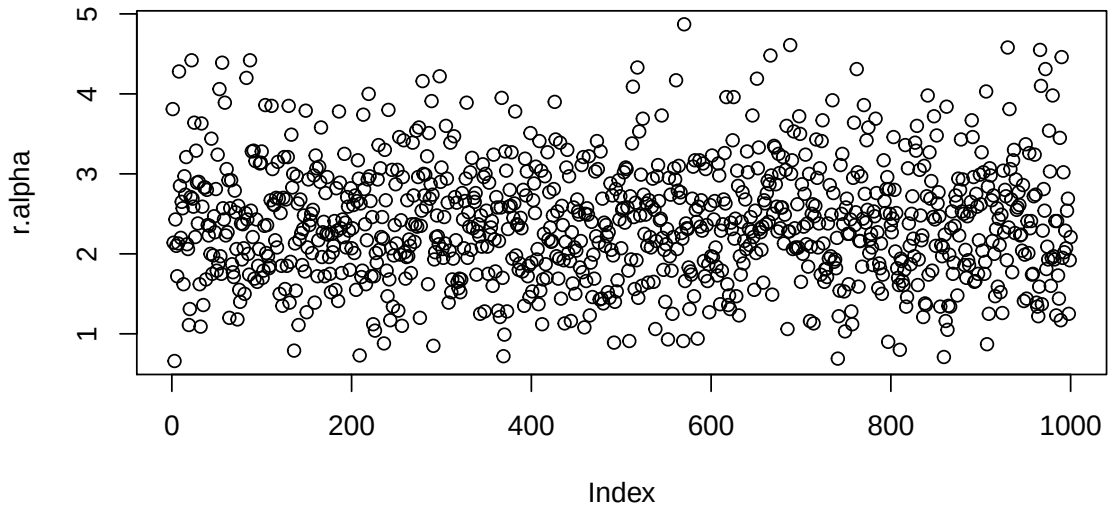
Por ejemplo, suponga que $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$ cuyas observaciones, para una iteración en particular, fueron $(2, 2, 3)$ y que la observación para β fue 0.9. De esta manera, se crea una grilla de los posibles valores que puede tomar α y mediante el uso de la función `sample` se simula un valor proveniente de esta distribución rara.

```
# creaci\'on de la grilla para alpha
alpha.grid <- seq(0.05, 20, by=0.01)
be <- 0.9
a1 <- 2
b2 <- 3
t <- c(2,2,3)

# probabilidad para cada valor en la grilla
post.alpha <- c()
for(k in 1:length(alpha.grid)){
 post.alpha[k] <- post(t,alpha.grid[k],be,a1,b1)
}
N.grid <- length(post.alpha)
post.alpha <- post.alpha/sum(post.alpha)
sum(post.alpha)

## [1] 1

# simulaci\'on de una sola observaci\'on
rpost <- sample(N.grid, 1, prob=post.alpha, replace=TRUE)
r.alpha <- alpha.grid[rpost]
r.alpha

## [1] 3.7
```

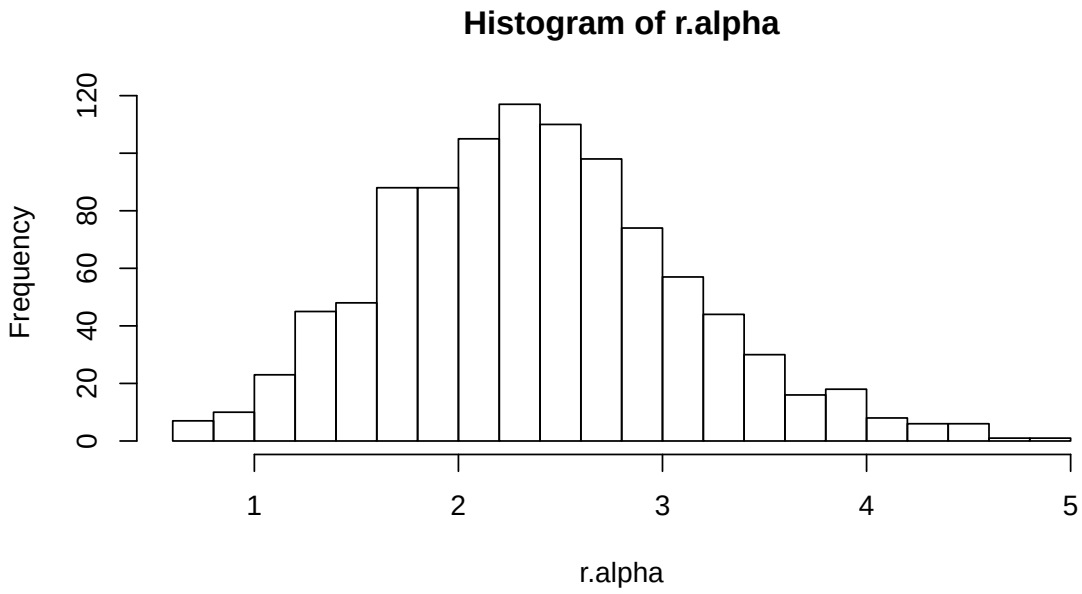
Por otro lado, en términos exploratorios, es posible simular varios valores de la distribución no conocida y determinar qué forma tiene. Para la anterior configuración, y utilizando el siguiente código, se simularon 100 valores de esta distribución. En general, es posible afirmar que su forma es parecida a la

de una distribución gamma, esto tiene sentido pues está en función de distribuciones gamma, sesgadas a la derecha y unimodales.

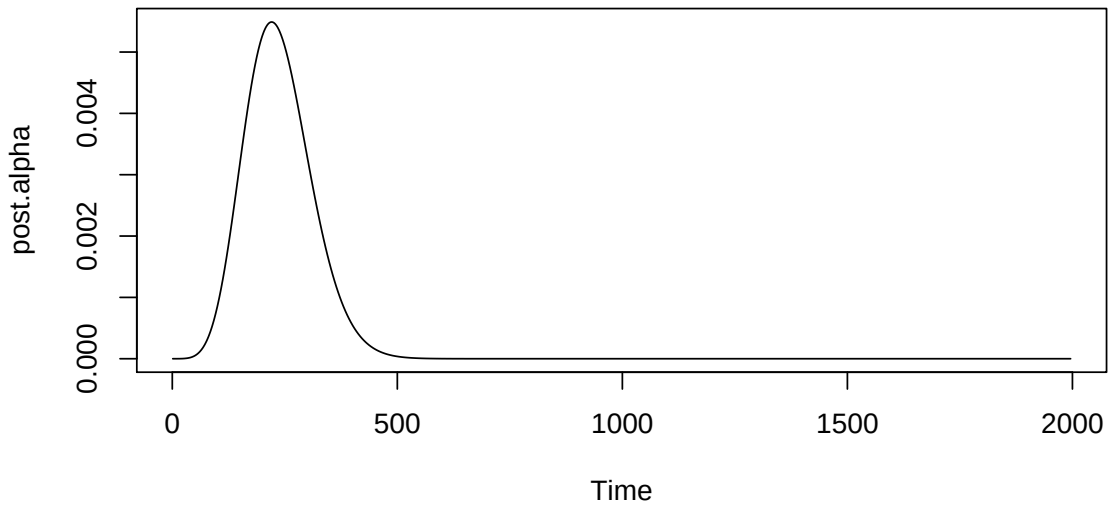
```
# corroborar la estructura de la cadena
N.sim <- 1000

rpost <- sample(N.grid, N.sim, prob=post.alpha, replace=TRUE)
r.alpha <- alpha.grid[rpost]
mean(r.alpha)

## [1] 2.4

plot(r.alpha)
```




```
hist(r.alpha, breaks=20)
```

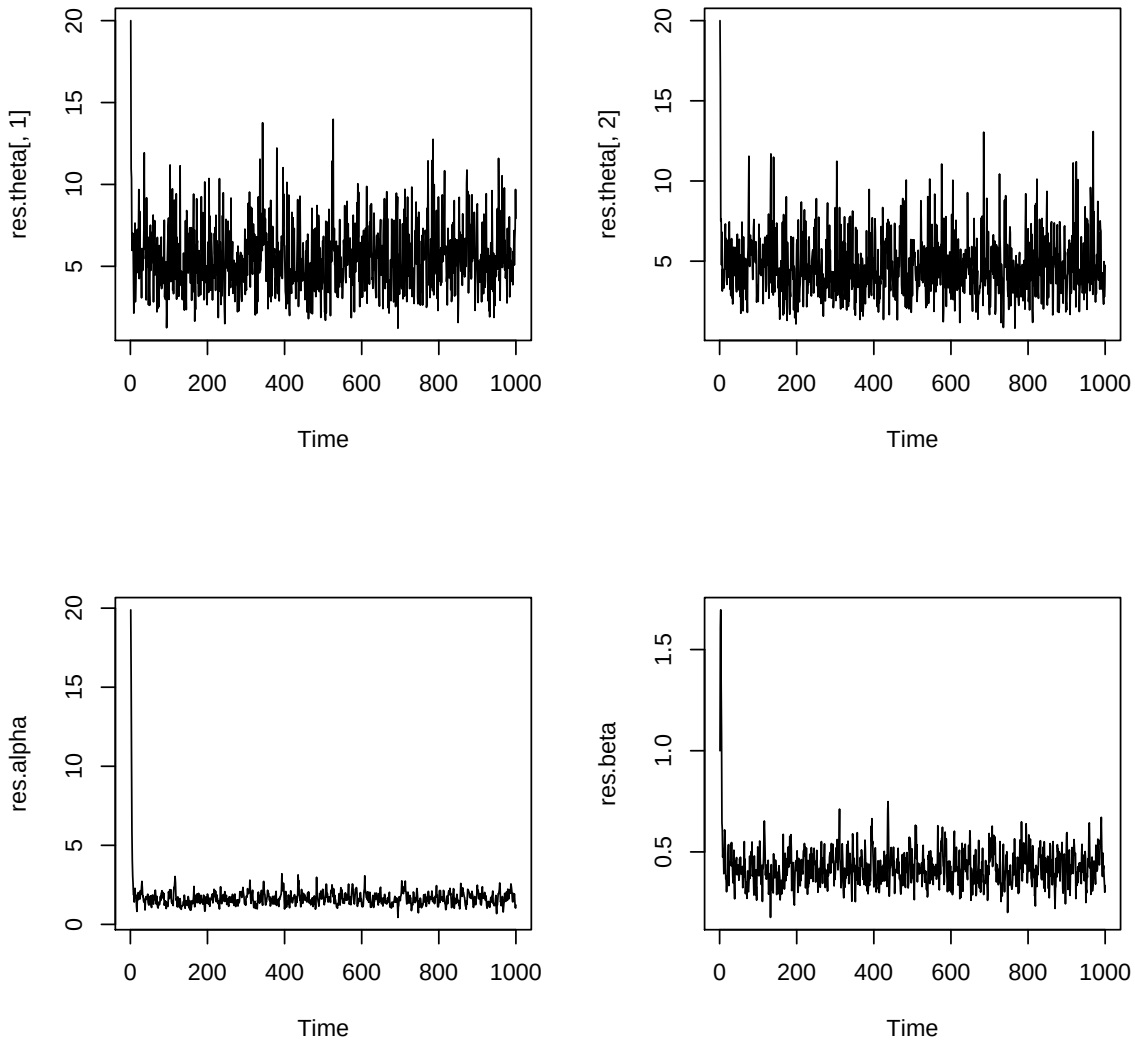



```
ts.plot(post.alpha)
```



Ahora ilustramos el procedimiento de obtener la estimación de α, β y los θ_i usando suponiendo que se quiere estimar el número de accidentes de tránsito relacionados con motociclistas en las veinte localidades de la ciudad de Bogotá, suponga que en un mismo día determinado los números de estos accidentes son: $6, 5, 9, 2, 3, 0, 4, 1, 1, 2, 1, 6, 7, 2, 4, 3, 0, 4, 3, 2$.

```
y <- c(6, 5, 9, 2, 3, 0, 4, 1, 1, 2, 1, 6, 7, 2, 4, 3, 0, 4, 3, 2)
n <- length(y)
n.sim <- 1000
res.theta <- matrix(NA,n.sim,n); res.beta <- c(); res.alpha <- c()
# Valor inicial para theta
res.theta[1,] <- 20
# Valor inicial para beta
res.beta[1] <- 1
# Simular un valor para alpha
 # creaci\'on de la grilla para alpha
 alpha.grid <- seq(0.05, 20, by=0.01)
 a <- c <- 2
 b <- d <- 3
 # probabilidad para cada valor en la grilla
 post.alpha <- c()
 for(k in 1:length(alpha.grid)){
  post.alpha[k] <- post(res.theta[1,], alpha.grid[k], res.beta[1],a,b)
 }
 N.grid <- length(post.alpha)
 post.alpha <- post.alpha/sum(post.alpha)
 res.alpha[1] <- alpha.grid[sample(length(alpha.grid), 1, prob=post.alpha, replace=TRUE)]
 # Aqu?? comienza a simular los valores de los par\'ametros
 for(i in 2:n.sim){
 # Simular un valor para theta
   for(j in 1:n){
     res.theta[i,j] <- rgamma(1, shape=y[j]+res.alpha[i-1], rate=res.beta[i-1]+1)
   }
 # Simular un valor para beta
   res.beta[i] <- rgamma(1, n*(res.alpha[i-1]+c-1)+1, rate=n*d+sum(res.theta[i,]))
 # Simular un valor para alpha
 post.alpha <- c()
 for(k in 1:length(alpha.grid)){
  post.alpha[k] <- post(res.theta[i,],alpha.grid[k],res.beta[i],a,b)
 }
 post.alpha <- post.alpha/sum(post.alpha)

 res.alpha[i] <- alpha.grid[sample(length(post.alpha), 1, prob=post.alpha, replace=TRUE)]
 }
 # Verificar la convergencia de algunos par??metros
 par(mfrow=c(2,2))
 ts.plot(res.theta[,1]); ts.plot(res.theta[,2])
 ts.plot(res.alpha); ts.plot(res.beta)
```




```
# Calcular la estimaci\'on de los par\'ametros tomando la segunda mitad de los valores simulados
colMeans(res.theta[-(1:(n.sim/2)),])
```

```
##  [1] 5.6 4.7 7.5 2.6 3.3 1.1 3.9 1.9 1.9 2.5 1.8 5.4 6.0 2.6 3.9 3.4 1.2
## [18] 3.8 3.2 2.6
```

```
mean(res.alpha[-(1:(n.sim/2))])
```

```
## [1] 1.7
```

```
mean(res.beta[-(1:(n.sim/2))])
```

```
## [1] 0.42
```

El anterior desarrollo teórico también se puede adaptar para el caso cuando se asume la misma media para todas las variables observadas, esto es, $Y_i \mid \theta$ para $i = 1, \cdots, n$. En este caso, se tiene que

$$p(\theta, \alpha, \beta \mid \mathbf{Y}) \propto e^{-(n+\beta)\theta} \theta^{\sum y_i + \alpha - 1} \beta^{\alpha + c - 1} e^{-\alpha b - \beta d} \alpha^{a-1} / \Gamma(\alpha)$$

de donde se puede concluir que

$$\theta \mid \alpha, \beta, \mathbf{Y} \sim Gamma(\sum y_i + \alpha, n + \beta) \tag{4.2.8}$$

$$\alpha \mid \theta, \beta, \mathbf{Y} \propto (\theta\beta)^{\alpha} e^{-\alpha b} \alpha^{a-1} / \Gamma(\alpha) \tag{4.2.9}$$

$$\beta \mid \theta, \alpha, \mathbf{Y} \sim Gamma(\alpha + c, \theta + d) \tag{4.2.10}$$

4.2.3 Modelo Normal

Considere una variación de la estructura jerárquica de la sección 4.1.3, en donde las observaciones siguen el siguiente modelo de probabilidad

$$Y_i \mid \theta_i \sim Normal(\theta_i, \sigma^2) \qquad i = 1, \ldots, n$$

y el parámetro σ^2 se supone conocido. Sin embargo, la distribución previa para los parámetros de interés θ_i es

$$\theta_i \mid \mu \sim Normal(\mu, \tau^2) \qquad i = 1, \ldots, n$$

en donde los parámetros μ y τ^2 son desconocidos. De esta forma, es necesario hallar una forma de estimar los valores de estos dos hiperparámetros, esto se puede llevar a cabo considerando diferente estructuras de dependencia entre mu y τ^2.

Hiperparámetros independientes

En primer lugar, supongamos que los hiperparámetros son independientes en la distribución previa, es decir que su función de densidad conjunta se puede factorizar como el producto de las distribuciones marginales de cada uno de los hiperparámetros. más aún, si se supone que las distribuciones previa marginales son no informativas y siguen una estructura probabilística uniforme, entonces se tiene que

$$p(\mu, \tau^2) = p(\mu)p(\tau^2) \propto k$$

Con esta formulación se deduce que la distribución posterior conjunta condicional a una sola observación está dada por

$$\begin{aligned} p(\theta_i, \mu, \tau^2 \mid Y_i) &\propto p(Y_i \mid \theta_i) p(\theta_i \mid \mu, \tau^2) p(\mu, \tau^2) \\ &\propto p(Y_i \mid \theta_i) p(\theta_i \mid \mu, \tau^2) \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \theta_i)^2 \right\} \frac{1}{\tau} \exp\left\{ -\frac{1}{2\tau^2}(\theta_i - \mu)^2 \right\} \end{aligned} \tag{4.2.11}$$

Y la distribución distribución posterior conjunta condicional a todas las observaciones y a todos los parámetros de interés es

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mu, \tau^2) \\ &\propto \prod_{i=1}^{n} p(Y_i \mid \theta_i) \prod_{i=1}^{n} p(\theta_i \mid \mu, \tau^2) \\ &\propto \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_i)^2 \right\} \frac{1}{\tau^n} \exp\left\{ \frac{1}{2\tau^2} \sum_{i=1}^{n} (\theta_i - \mu)^2 \right\} \end{aligned}$$

Utilizaremos la técnica del condicionamiento para encontrar la distribución condicional del vector de parámetros de interés $\boldsymbol{\theta}$ y de los hiperparámetros. Por lo tanto se tiene que

$$p(\boldsymbol{\theta} \mid \mu, \tau^2, \mathbf{Y}) \propto p(\boldsymbol{\theta}, \underbrace{\mu, \tau^2}_{fijos}, \mathbf{Y})$$
$$p(\mu \mid \boldsymbol{\theta}, \tau^2, \mathbf{Y}) \propto p(\mu, \underbrace{\boldsymbol{\theta}, \tau^2}_{fijos}, \mathbf{Y})$$
$$p(\tau^2 \mid \boldsymbol{\theta}, \mu, \mathbf{Y}) \propto p(\mu, \underbrace{\boldsymbol{\theta}, \tau^2}_{fijos}, \mathbf{Y})$$

Con la anterior formulación se tiene la siguiente serie de resultados que dan cuenta de las distribuciones apropiadas para cada uno de los parámetros.

Resultado 4.2.1. *La distribución posterior del parámetro de interés θ_i es*

$$\theta_i \sim Normal(\mu_i, \tau_1^2)$$

en donde

$$\mu_i = \frac{\frac{1}{\sigma^2} Y_i + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

Prueba. Utilizando la técnica del condicionamiento posterior se tiene que

$$p(\theta_i \mid \mu, \tau^2, Y_i) \propto p(\theta_i, \underbrace{\mu, \tau^2}_{fijos}, Y_i)$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta_i)^2 - \frac{1}{2\tau^2}(\theta_i - \mu)^2\right\}$$

y utilizando el mismo razonamiento que en la demostración del Resultado 2.6.1 se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\mu_i, \tau_1^2)$. ■

Resultado 4.2.2. *La distribución posterior del hiper-parámetro μ es*

$$\mu \sim Normal(\bar{\theta}, \tau^2/n)$$

en donde $\bar{\theta} = \dfrac{1}{n}\sum_{i=1}^{n} \theta_i$.

Prueba. Utilizando la técnica del condicionamiento posterior y teniendo en cuenta que

$$\sum_{i=1}^{n}(\theta_i - \mu)^2 = \sum_{i=1}^{n}(\theta_i - \bar{\theta})^2 + n(\mu - \bar{\theta})^2$$

entonces, se tiene que

$$p(\mu \mid \boldsymbol{\theta}, \tau^2, \mathbf{Y}) \propto p(\mu, \underbrace{\boldsymbol{\theta}, \tau^2}_{fijos}, \mathbf{Y})$$
$$\propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n}(\theta_i - \mu)^2\right\} \propto \exp\left\{-\frac{n}{2\tau^2}(\mu - \bar{\theta})^2\right\}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\bar{\theta}, \tau^2/n)$. ■

Resultado 4.2.3. *La distribución posterior del hiper-parámetro* τ^2 *es*

$$\tau^2 \sim Inversa - Gamma(n/2-1, nS_\mu^2/2)$$

en donde $nS_\mu^2 = \sum_{i=1}^n (\theta_i - \mu)^2$.

Prueba. Utilizando la técnica del condicionamiento posterior se tiene que

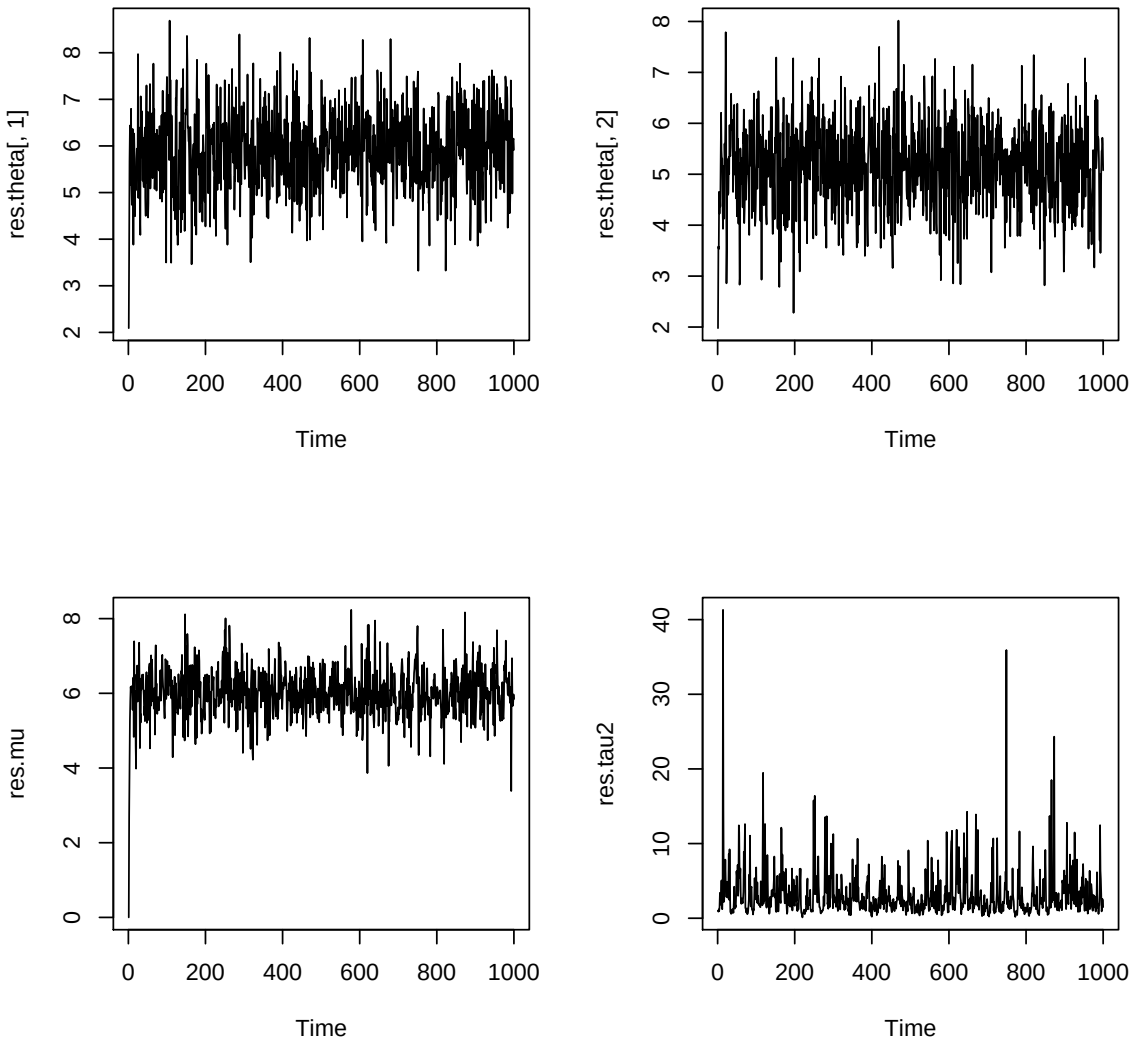
$$\begin{aligned} p(\tau^2 \mid \boldsymbol{\theta}, \mu, \mathbf{Y}) &\propto p(\tau^2, \underbrace{\boldsymbol{\theta}, \mu}_{fijos}, \mathbf{Y}) \\ &\propto \frac{1}{\tau^n} \exp\left\{ \frac{1}{2\tau^2} \sum_{i=1}^n (\theta_i - \mu)^2 \right\} \\ &\propto \left(\tau^2\right)^{-n/2} \exp\left\{ \frac{nS_\mu^2}{2\tau^2} \right\} \end{aligned}$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Inversa - Gamma(n/2-1, nS_\mu^2/2)$. ■

Utilizando un algoritmo que genere una cadena de Markov, y utilizando los anteriores resultados se realiza un análisis bayesiano propiamente dicho.

Ilustramos la implementación en `R` a continuación para datos de y de 5.8, 4.7, 7.0, 8.3, 3.7, 3.7, 5.5, 7.7, 6.7 y 6.7, usando $\sigma^2 = 1$.

```
library(pscl)
y <- c(5.8, 4.7, 7.0, 8.3, 3.7, 3.7, 5.5, 7.7, 6.7, 6.7)
n <- length(y); sigma2 <- 1
n.sim <- 1000
# Espacio para guardar los resultados simulados
res.mu <- rep(0, n.sim); res.tau2 <- rep(1, n.sim); res.theta<- matrix(NA, n.sim, n)
# Simular el primer valor para theta
tau2_1 <- (sigma2^-1 + res.tau2[1]^-1)^-1
mu_i <- (y/sigma2 + res.mu[1]/res.tau2[1])*tau2_1
for(j in 1:n){
  res.theta[1,j] <- rnorm(1, mu_i[j], sqrt(tau2_1))
}
# Aqu?? comienza a simular valores para todos los par\'ametros
for(i in 2:n.sim){
  res.mu[i] <- rnorm(1, mean(res.theta[i-1,]), sqrt(res.tau2[i-1]/n))
  res.tau2[i] <- rigamma(1, alpha=n/2-1, beta=sum((res.theta[i-1,]-res.mu[i])^2)/2)
  tau2_1 <- (sigma2^-1 + res.tau2[i]^-1)^-1
  mu_i <- (y/sigma2 + res.mu[i]/res.tau2[i])*tau2_1
  for(j in 1:n){
    res.theta[i,j] <- rnorm(1, mu_i[j], sqrt(tau2_1))
  }
}
 # Verificar la convergencia de algunos par??metros
 par(mfrow=c(2,2))
 ts.plot(res.theta[,1]); ts.plot(res.theta[,2])
 ts.plot(res.mu); ts.plot(res.tau2)
```




```
# Calcular la estimaci\'on de los par\'ametros tomando la segunda mitad de los valores simulados
colMeans(res.theta[-(1:(n.sim/2)),])
```

```
##  [1] 6.0 5.1 6.6 7.5 4.4 4.5 5.7 7.1 6.5 6.5
```

```
mean(res.mu[-(1:(n.sim/2))])
```

```
## [1] 6
```

```
mean(res.tau2[-(1:(n.sim/2))])
```

```
## [1] 3.1
```

Ahora consideramos el caso cuando la media de las variables observadas es común, esto es, $Y_i \mid \theta \sim Normal(\theta, \sigma^2)$ y $\theta \mid \mu \sim Normal(\mu, \tau^2)$ para $i = 1, \cdots, n$, asumimos la misma distribución no informativa para μ y τ^2. En este caso tenemos que

$$p(\theta, \mu, \tau^2 \mid \mathbf{Y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right\} \frac{1}{\tau} \exp\left\{-\frac{1}{2\tau^2}(\theta - \mu)^2\right\}$$

De la expresión se tiene que

$$\theta \mid \mu, \tau^2, \mathbf{Y} \sim Normal(\mu_n, \tau_n^2) \tag{4.2.12}$$

$$\mu \mid \theta, \tau^2, \mathbf{Y} \sim Normal(\theta, \tau^2) \tag{4.2.13}$$

$$p(\tau^2 \mid \theta, \mu, \mathbf{Y}) \propto (\tau^2)^{-1/2} \exp\left\{-\frac{(\theta - \mu)^2}{2\tau^2}\right\} \tag{4.2.14}$$

con $\tau_n^2 = (\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1}$ y $\mu_n = \tau_n^2(\frac{n\bar{y}}{\sigma^2} + \frac{\mu}{\tau^2})$. La expresión en (4.2.14) no corresponde a ninguna distribución con forma conocida, y debe hacer uso de métodos de simulación para muestrear valores de τ^2

Hiperparámetros dependientes

Siguiendo el algoritmo dado al comienzo de esta sección, en donde se dan los lineamentos generales para realizar un análisis jerárquico. En primer lugar se debe considerar la distribución posterior de los parámetros, que en este caso depende de la distribución previa de los hiperparámetros.

Suponga entonces, al igual que en capítulos anteriores, que los hiperparámetros son dependientes a una vía. Es decir, que μ depende de τ^2 pero que τ^2 no depende de μ. En estos términos, la distribución previa de los hiperparámetros está dada por

$$p(\mu, \tau^2) = p(\mu \mid \tau^2)p(\tau^2)$$

Luego, siguiendo la regla de bayes y suponiendo que los hiperparámetros son condicionalmente independientes de las observaciones dado el vector de parámetros de interés, la distribución posterior del vector de parámetros de interés $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ y de los hiperparámetros μ, τ^2 es

$$\begin{aligned} p(\boldsymbol{\theta}, \mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mathbf{Y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mu, \tau^2)p(\mu, \tau^2) \\ &\propto p(\mu, \tau^2) \prod_{i=1}^n p(Y_i \mid \theta_i) \prod_{i=1}^n p(\theta_i \mid \mu, \tau^2) \\ &\propto p(\mu, \tau^2) \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^n (y_i - \theta_i)^2\right\} \frac{1}{\tau^n} \exp\left\{\frac{-1}{2\tau^2}\sum_{i=1}^n (\theta_i - \mu)^2\right\} \end{aligned}$$

Con base en lo anterior, se tienen el siguiente resultado para el análisis bayesiano jerárquico de un sólo componente θ_i de $\boldsymbol{\theta}$.

Resultado 4.2.4. *La distribución posterior del componente θ_i perteneciente al vector de parámetros de interés $\boldsymbol{\theta}$ es*

$$\theta_i \sim Normal(\mu_i, \tau_1^2)$$

en donde

$$\mu_i = \frac{\frac{1}{\sigma^2}Y_i + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \qquad y \qquad \tau_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$$

Prueba. La prueba del resultado es inmediata al considerar la técnica del condicionamiento posterior como en la demostración del Resultado 4.2.1. puesto que

$$\begin{aligned} p(\theta_i \mid \mu, \tau^2, Y_i) &\propto p(\theta_i, \underbrace{\mu, \tau^2}_{fijos} \mid Y_i) \\ &\propto p(Y_i \mid \theta_i) p(\theta_i \mid \mu, \tau^2) p(\mu, \tau^2) \\ &\propto (Y_i \mid \theta_i) p(\theta_i \mid \mu, \tau^2) \end{aligned}$$

■

siguiendo con el algoritmo del análisis jerárquico, el siguiente paso corresponde a la determinación de la distribución posterior de los hiperparámetros μ, τ^2 la cual, suponiendo que la distribución previa conjunta para ambos hiperparámetros es uniforme y no informativa, está dada por el Resultado 4.1.1.

$$\begin{aligned} p(\mu, \tau^2 \mid \mathbf{Y}) &\propto p(\mu, \tau^2) p(\mathbf{Y} \mid \mu, \tau^2) \\ &\propto \prod_{i=1}^{n} p(Y_i \mid \mu, \tau^2)) \\ &\propto \prod_{i=1}^{n} Normal(\mu, \tau^2 + \sigma^2) \end{aligned}$$

Ahora, por otro lado, el análisis individual de los hiperparámetros está regido por la siguiente expresión

$$p(\mu, \tau^2 \mid \mathbf{Y}) = p(\mu \mid \tau^2, \mathbf{Y}) p(\tau^2 \mid \mathbf{Y})$$

En este orden de ideas, se tienen los siguientes resultados acerca de la distribución posterior para μ dada por $p(\mu \mid \tau^2, \mathbf{Y})$ y para τ^2 dada por $p(\tau^2 \mid \mathbf{Y})$

Resultado 4.2.5. *La distribución posterior del hiperparámetro mu condicionada a* $\tau^2, \mathbf{Y}$ *es*

$$\mu \mid \tau^2, \mathbf{Y} \sim Normal\left(\hat{\mu}, \hat{\tau}^2\right)$$

donde $\hat{\mu} = \bar{Y}$ *y* $n\hat{\tau}^2 = \sigma^2 + \tau^2$.

Prueba. Utilizando la técnica del condicionamiento posterior, nótese que la distribución posterior de *mu* toma la siguiente forma

$$\begin{aligned} p(\mu \mid \tau^2, \mathbf{Y}) &\propto p(\mu, \underbrace{\tau^2}_{fijo} \mid \mathbf{Y}) \\ &\propto \prod_{i=1}^{n} Normal(\mu, \tau^2 + \sigma^2) \end{aligned}$$

Partiendo de este hecho, es fácil confirmar que

$$
\begin{aligned}
p(\mu \mid \tau^2, \mathbf{Y}) &\propto \qquad\qquad\qquad\qquad\qquad\qquad \propto \exp\left\{\frac{1}{2(\sigma^2+\tau^2)}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\
&= \exp\left\{\frac{1}{2(\sigma^2+\tau^2)}\sum_{i=1}^{n}(y_i^2-2\mu Y_i+\mu^2)\right\} \\
&\propto \exp\left\{\frac{n}{2(\sigma^2+\tau^2)}(\mu^2-2\mu\bar{Y})\right\} \\
&\propto \exp\left\{\frac{n}{2(\sigma^2+\tau^2)}(\mu-\bar{Y})^2\right\}
\end{aligned}
$$

Por lo tanto, factorizando convenientemente, se encuentra una expresión idéntica a la función de distribución de una variable aleatoria con distribución $Normal(\hat{\mu}, \hat{\tau}^2)$. ■

Resultado 4.2.6. *La distribución posterior del hiperparámetro τ es*

$$
p(\tau^2 \mid \mathbf{Y}) \propto \sqrt{\hat{\tau}}\prod_{i=1}^{n}(\sigma^2+\tau^2)^{-1/2}\exp\left\{-\frac{1}{2(\sigma^2+\tau^2)}(y_i-\hat{\mu})^2\right\}
$$

Prueba. En primer lugar, nótese que

$$
\begin{aligned}
p(\tau \mid \mathbf{Y}) &= \frac{p(\mu, \tau^2 \mid \mathbf{Y})}{p(\mu \mid \tau^2, \mathbf{Y})} \qquad\qquad \forall\mu \\
&\propto \frac{\prod_{i=1}^{n} Normal(\mu, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \qquad\qquad \forall\mu
\end{aligned}
$$

La anterior igualdad debe mantenerse para cualquier valor de μ; en particular se debe mantener para $\mu = \hat{\mu}$ (Gelman, Carlin, Stern & Rubin 2003). Por tanto,

$$
\begin{aligned}
p(\tau \mid \mathbf{Y}) &\propto \frac{Normal(\hat{\mu}, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \\
&\propto \frac{\prod_{i=1}^{n} Normal(\hat{\mu}, \sigma^2+\tau^2)}{Normal(\hat{\mu}, \hat{\tau}^2)} \\
&\propto \sqrt{\hat{\tau}}\prod_{i=1}^{n}(\sigma^2+\tau^2)^{-1/2}\exp\left\{-\frac{1}{2(\sigma^2+\tau^2)}(y_i-\hat{\mu})^2\right\}\exp\left\{\frac{1}{2\hat{\tau}^2}(\hat{\mu}-\hat{\mu})^2\right\} \\
&\propto \sqrt{\hat{\tau}}\prod_{i=1}^{n}(\sigma^2+\tau^2)^{-1/2}\exp\left\{-\frac{1}{2(\sigma^2+\tau^2)}(y_i-\hat{\mu})^2\right\}
\end{aligned}
$$

■

En términos de simulación, los anteriores resultados garantizan una estructura formal que permita simular la distribución posterior del hiperparámetro τ^2, y mediante esta encontrar una estimación para reemplazarla en la distribución posterior del hiperparámetro μ y repetir el proceso anterior. Con estos valores bien definidos, entonces utilizar el Resultado 4.2.4 para proseguir con el análisis bayesianoclásico.

```
library(pscl)
y <- c(5.8, 4.7, 7.0, 8.3, 3.7, 3.7, 5.5, 7.7, 6.7, 6.7)
n <- length(y); sigma2 <- 1
```

```r
n.sim <- 1000
# Espacio para guardar los resultados simulados
res.mu <- rep(0, n.sim); res.theta <- matrix(NA, n.sim, n)
# Simular un valor para tau^2 con Grilla
grid.tau2 <- seq(0.001,5,by=0.001)
pos.tau2 <- grid.tau2^(1/4)*(sigma2+grid.tau2)^(-n/2)*exp(-(n-1)*var(y)/(2*(sigma2+grid.tau2)))
pos.tau2 <- pos.tau2/sum(pos.tau2)
res.tau2 <- sample(grid.tau2, n.sim, prob=pos.tau2, replace=TRUE)
for(i in 1:n.sim){
  # Simular el primer valor para mu
  res.mu[i] <- rnorm(1, mean(y), sqrt((sigma2+res.tau2[i])/n))
  # Simular el primer valor para theta
  tau2_1 <- (sigma2^-1 + res.tau2[i]^-1)^-1
  mu_i <- (y/sigma2 + res.mu[i]/res.tau2[i])*tau2_1
  for(j in 1:n){
    res.theta[i,j] <- rnorm(1, mu_i[j], sqrt(tau2_1))
  }
}
 # Calcular la estimaci\'on de los par\'ametros tomando la segunda mitad de los valores simulados
 colMeans(res.theta[-(1:(n.sim/2)),])

##  [1] 5.9 5.1 6.6 7.5 4.4 4.5 5.7 7.1 6.4 6.5

 mean(res.mu[-(1:(n.sim/2))])

## [1] 6

 mean(res.tau2[-(1:(n.sim/2))])

## [1] 2.4
```

4.3 Ejercicios

1. Para el modelo $Y_i \sim \theta_i Normal(\theta_i, \sigma^2)$ para $i = 1, \cdots, n$ del modelo Normal-Normal, desarrollando $E(Y_i)$, encuentra que μ se peude calcular como $\bar{y}$.

2. Demustre las ecuaciones 4.2.8, 4.2.9 y 4.2.10. Modifique los códigos del caso $Y_i \mid \theta_i \sim Poisson(\theta_i)$ para estimar α, β y θ, y aplíquelos a los datos del ejemplo 2.4.1 asumiendo (i) $\alpha = \beta = 0.1$ y (ii) $\alpha = \beta = 10$. Cómo afectan los valores de α y β sobre la estimación final de θ?

3. Para los datos del ejemplo 2.6.1, implementa las ecuaciones (4.2.12), (4.2.13) y (4.2.14) para estimar los valores de θ, μ y τ^2. Utilice $\sigma = 0.1cm$.

4. Encuentre la forma de estimar los parámetros μ y τ^2 en una muestra aleatoria $Y_i \sim Normal(\theta, \sigma^2)$ con σ^2 conocido, para $i = 1, \cdots, n$, asumiendo (i) independencia entre μ y τ^2, (2) μ depende de τ^2, pero τ^2 no depende de μ.

Bibliografía

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Apostol, T. M. (1957), *Mathematical Analysis*, McGraw - Hill.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2 edn, Springer.

Box, G. E. P. & Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis*, 1 edn, Wiley.

Carlin, B. P. & Louis, T. A. (1996), *Bayes and Empirical Bayes for Data Analysis*, 1 edn, Chapman and Hall/CRC.

Cavanaugh, J. E. (1997), 'Unifying the derivations of the Akaike and corrected Akaike information criteria', *Statistics & Probability Letters* **31**, 201–208.

Crowley, J. & Hu, M. (1977), 'Covariance analysis of heart transplant survival data', *Journal of the American Statistical Association* **72**, 27 – 36.

Dempster, A. P. (1974), The direct use of likelihood for significance testing, *in* 'Proceedings of Conference on Foundational Questions in Statistical Inference', Department of Theoretical Statistics: University of Aarhus., pp. 335 – 352.

Efron, B. (2010), *Large-Scale Inference. Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press.

Efron, B. & Morris, C. (1975), 'Data analysis using stein's estimator and its generalizations', *Journal of the American Statistical Association* **70**, 311 – 319.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995), *Bayesian Data Analysis*, 1 edn, Chapman and Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003), *Bayesian Data Analysis*, 2 edn, Chapman and Hall/CRC.

Jordan, M. I. (2004), The exponential family and generalized linear models.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011), 'MCMCpack: Markov chain monte carlo in R', *Journal of Statistical Software* **42**(9), 22.
URL: *http://www.jstatsoft.org/v42/i09/*

Migon, H. S. & Gamerman, D. (1999), *Statistical Inference: An Integrated Approach*, Arnold.

Peña, D. (2002), *Análisis de datos multivariantes*, McGraw-Hill.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461 – 464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & VanderLinde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society* **B 64**, 583 – 639.

Student (1908), 'The probable error of a mean', *Biometrika* **6**(1), 1 – 25.

Therneau, T. & Lumley, T. (2011), *survival: Survival analysis, including penalised likelihood.* R package version 2.36-5.

Wikipedia (2011*a*), 'Hit — Wikipedia, the free encyclopedia'.

Wikipedia (2011*b*), 'Porcentaje de bateo. Wikipedia'.

Yee, T. W. (2012), *VGAM: Vector Generalized Linear and Additive Models.*, URL http://CRAN.R-project.org/package=VGAM. R package version 0.9-0.

Zhang, H. & Gutiérrez, H. A. (2010), *Teoría estadística. Aplicación y métodos.*, Universidad Santo Tomás.

Índice de figuras

Índice de Tablas