Modelos Bayesianos con R y STAN

Andrés Gutiérrez - Hanwen Zhang

2021-06-05

Índice general

Prefacio

Antes de comenzar

Cuestionamientos sobre el enfoque bayesiano

Gelman (2008) presenta algunos de los cuestionamientos que algunos estadísticos anti-bayesianos han argumentado en contra de este paradigma que, sin lugar a dudas, ha proporcionado una valiosa herramienta de modelación en la ciencia contemporanea. Revisemos algunos de estos argumentos:

> La inferencia bayesiana es una teoría matemática coherente pero no brinda la suficiente confianza en usos científicos. Las distribuciones *previas* subjetivas no inspiran confianza porque ni siquiera existe algún principio objetivo para elegir una distribución previa no informativa. ¿De dónde vienen las distribuciones previas? No confío en ellas y no veo ninguna razón para recomendarlas a otra gente, apenas me siento cómodo acerca de su coherencia filosófica.

Este argumento es débil puesto que la teoría bayesiana es una teoría cinetífica apoyada en los axiomas matemáticos de la teoría de la medida y de probabilidad. De la mismaforma, nótese que tampoco existe un principio objetivo para escoger una verosimilitud. ¿De dónde vienen las regresiones logísticas? ¿quién dijo que los datos eran normales? Como toda ciencia, la estadística se basa en procedimientos subjetivos que inducen resultados que se pueden probar de una manera objetiva. Al decidir usar una determinada distribución previa, el investigador está haciendo uso de su conocimiento objetivo sobre el fenómeno de interés. Esto no dista mucho de la planificación de un estudio por muestreo o de un experimento, en donde se hace uso de la información auxiliar disponible para definir la mejor versión del estudio. Además, como se verá más adelante, sí existen principios objetivos que permiten decidir acerca de la elección de una distribución previa; por ejemplo, la invarianza de la distribución previa frente a transformaciones de los parámetros.

> La teoría bayesiana requiere un pensamiento mucho más profundo sobre la situación y recomendarle a los investigadores comunes el uso del teorema de Bayes es como darle al hijo del vecino la llave de un *F-16*. De veras que, yo comenzaría con algo de métodos probados y

> confiables, y entonces generalizaría la situación utilizando los principios estadísticos y la teoría del minimax, que no dependen de ninguna creencia subjetiva. Especialmente cuando las distribuciones previas que veo en la práctica toman formas conjugadas. ¡Qué coincidencia!

Como científicos e investigadores debemos tratar con el conocimiento objetivo y dejar a un lado las creencias subjetivas. Es por eso que las distribuciones previas que se manejan en la inferencia bayesiana son objetivas de la misma forma que lo son los métodos frecuentistas al asignar un modelo probabilístico a la verosimilitud de los datos. El resultado final sólo depende del modelo asumido y de los datos recolectados. A pesar de que algunos resultados de la inferencia bayesiana coinciden con el acercamiento frecuentista, esto no sucede en todos los casos. Si la distribución es conjugada, simplemente quiere decir que es posible utilizar un generador de números aleatorios conocido; sin embargo, en pleno siglo XXI, esto ya no constituye un problema.

> Dejando de lado las preocupaciones matemáticas, me gustan las estimaciones insesgadas, los intervalos de confianza con un nivel real de cobertura. Pienso que la manera correcta de inferir es acercarse al parámetro tanto como sea posible y desarrollar métodos robustos que trabajen con supuestos mínimos. El acercamiento bayesiano intenta aproximar el insesgamiento, mientras asume supuestos más y más fuertes. En los viejos tiempos, los métodos Bayesianos por lo menos tenían la virtud de estar matemáticamente limpios. Hoy en día, cualquier inferencia se realiza mediante el uso de las cadenas de Markov con métodos de Monte Carlo (MCMC). Lo anterior significa que, no sólo no se pueden evaluar las características estadísticas del método, sino que tampoco se puede asegurar su convergencia.

Los métodos bayesianos parecen moverse rápidamente hacia la computación elaborada. Para bien o para mal, la computación se está convirtiendo en una plataforma central para el desarrollo científico y estadístico. Por otro lado, estos mismos adelantos de computación científica permiten evaluar las características de los modelos bayesianos y la convergencia de las cadenas de la distribución posterior. Haciendo uso de la rigurosidad científica, el investigador debe conocer a profundidad el espíritu de los métodos MCMC y verificar que la distribución posterior conjunta sobre un vector de parámetros no sea impropia, y por supuesto verificar que las cadenas tienen propiedades estacionarias.

> La gente tiende a creer los resultados que apoyan sus preconceptos y descreen los resultados que los sorprenden, ésta es una forma errada y sesgada de pensar. Pues bien, los métodos bayesianos animan este modo indisciplinado de pensamiento. Estoy seguro que muchos estadísticos bayesianos están actuando de buena fe; sin embargo, al mismo tiempo, también están proporcionando estímulo a investigadores descuidados y poco éticos por todas partes, porque el investigador queda estancado al momento de escoger una distribución previa.

Si hay una seria diferenciación entre las creencias subjetivas y los resultados posteriores, debería ser un indicador de revaluar el modelo usado. Además, ante el desconocimiento del fenómeno, el investigador bayesiano puede utilizar una distribución previa débil y añadir más información si se necesita. Las verificaciones predictivas (previas y posteriores) son una parte esencial del método bayesiano que obliga a repensar las creencias del investigador con respecto al parámetro de interés. Este ejercicio redunda en el replanteamineto de la distribución previa mediante el estudio de las distribuciones predictivas, decantándose al final por el mejor modelo.

> Los cálculos de la teoría de la decisión guían a la idea de que el muestreo probabilístico y la asignación aleatoria de tratamientos son ineficaces, de que los mejores diseños y muestras son los determinísticos. No tengo ningún conflicto con estos cálculos matemáticos; el conflicto es más profundo, en los fundamentos filosóficos, en la idea de que el objetivo de la estadística consiste en tomar una decisión óptima. Un estimador bayesiano es un estimador estadístico que reduce al mínimo el riesgo promedio. Sin embargo, cuando hacemos estadística, no estamos intentando *reducir al mínimo el riesgo promedio*, estamos intentando hacer estimación y juzgamiento de hipótesis.

Un estimador bayesiano es un estimador estadístico que minimiza el riesgo promedio. Uno de los primeros tópicos que se presentan en este libro es el de la teoría de la decisión y funciones de perdida, como herramientas fundamentales del aprendizaje estadístico (Hastie et al., 2009). Además, como se verá más adelante, la asignación de las unidades experimentales al tratamiento o la inclusión de las unidades muestrales en un estudio probabilístico debe y puede ser tenido en cuenta en los modelos bayesianos, mediante la inclusión en el modelo de las variables que intervinieron en la selección de las unidades. De la misma forma, el juzgamiento de hipótesis es una práctica que se extiende en la modelación bayesiana.

> No puedo estar al tanto de lo que están haciendo todos esos Bayesianos hoy en día. Desafortunadamente, toda clase de personas están siendo seducidas por las promesas de la inferencia automática con la *magia del MCMC*. Desearía que todos paráramos de una vez y por todas y empezáramos, de nuevo, a hacer estadística de la forma en que debe ser hecha: volviendo a los viejos tiempos en que un p-valor era utilizado para algo, cuando un intervalo de confianza tenía significado, y el sesgo estadístico era algo que se quería eliminar y no algo que se debiera abrazar.

Los métodos Bayesianos algunas veces son presentados como un motor de inferencia automática. Sin embargo, la inferencia bayesiana tiene tres etapas: formulación del modelo, ajuste del modelo a los datos, evaluación del ajuste. Así que el procedimiento no es mágico ni automático. Además, una de las ventajas de la estadística bayesiana es que deja de lado las sofisticaciones de la inferencia clásica en donde, por ejemplo, la simple interpretación de un intervalo

de confianza se hace muy complicada a la luz del razonamiento lógico. De la misma forma los valores p constituyen un paradigma cada vez más revalorado en la investigación social.

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 et al. (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 usen f para las distribuciones de muestreo; π, para las distribuciones previas y L, para las 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 mismo 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, los autores asumen que el lector sigue la idea de la referencia de la distribución p en cuestión.

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, la 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 ($A(\mathbf{X})$ o simplemente A) del parámetro de interés. 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 se tome (toda estimación que se provea) puede traer consigo un grado de falla o riesgo. Y es necesario escoger la acción óptima que de alguna forma minimice ese riesgo. Formalizando esta idea intuitiva, se define la función de pérdida L que asocia a cada dupla conformada por 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. 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 sigue siendo una cantidad fija, pero la incertidumbre que tiene el investigador sobre la localización del parámetro se puede modelar mediante funciones de probabilidad. 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 cualquiera de los dos casos anteriores, se busca 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. 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 $A = \sum_{i=1}^{n} c_i X_i$. Por tanto, 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 impone la restricción de que $\sum_{i=1}^{n} c_i = 1$. De esta forma, el riesgo está dado por $R = \sum c_i^2 \sigma^2$. Dado que σ^2 es fijo, al minimizar $\sum c_i^2$ sujeto a la restricción, se tiene que la solución es $c_i = 1/n$ 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 las formas lineales con $\sum c_i = 1$ es la media muestral $\bar{X}$.

Ejemplo 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, 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 su tomada con respecto a la distribución posterior $p(\theta|\mathbf{X})$.
Ejemplo 1.3. En el mismo contexto del ejemplo anterior, si cambiamos la función de pérdida a la siguiente

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

El riesgo estará 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 se observa que, bajo un mismo contexto, cuando se utilizan diferentes funciones de pérdida, también se obtienen distintas estimaciones, y distintas acciones que optimizan el riesgo.

1.2. Algunos resultados de probabilidad

Antes de entrar en el repaso de estos conceptos fundamentales, se definen los conceptos de **parámetro** y **espacio paramétrico** asociados a una distribución de probabilidad.

1. Un parámetro es aquella cantidad que define la forma funcional de una distribución de probabilidad; es decir, cuando el parámetro cambia de valor, la función de densidad y la función de distribución cambian. Las distribuciones de probabilidad pueden tener más de un parámetro. Cuando una distribución tiene solo un parámetro, éste se denota usualmente por θ, cuando se presenta más de un parámetro, la notación se cambia a $\boldsymbol{\theta}$, representando el vector de parámetros.
2. El espacio paramétrico, Θ, es el conjunto que contiene todos los posibles valores que puede tomar el parámetro o el vector de parámetros. Para distribuciones con un solo parámetro, Θ será un subconjunto de $\mathbb{R}$, mientras que para distribuciones con dos o más parámetros, Θ será un subconjunto de $\mathbb{R} \times \mathbb{R}$.

Para entender los fundamentos de la modelación bayesiana, es necesario recordar algunas definiciones y resultados de la teoría de probabilidad que ayudarán a hacer más expedito este periplo por la estadística bayesiana. En términos de notación, se utilizará indistintamente la expresión de integral, $\int$, indicando la sumatoria, en el caso de las variables aleatorias discretas o la integral de Riemann-Stieltjes en el caso de las variables aleatorias continuas.

Definición 1.1. 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

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{Y}) \; d\mathbf{Y} \tag{1.1}$$

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

$$p(\mathbf{X} \mid \mathbf{Y}) = \frac{p(\mathbf{X}, \mathbf{Y})}{p(\mathbf{Y})} \tag{1.2}$$

Resultado 1.1. *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*

$$p(\mathbf{X} \mid \mathbf{Z}) = \int p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \; d\mathbf{Y} \tag{1.3}$$

y

$$p(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})}{p(\mathbf{Y} \mid \mathbf{Z})} \tag{1.4}$$

Demostración. 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. 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

$$p(\mathbf{X},\mathbf{Y}\mid\mathbf{Z})=p(\mathbf{X}\mid\mathbf{Z})p(\mathbf{Y}\mid\mathbf{Z}) \tag{1.5}$$

Resultado 1.2. *Si* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *con respecto a* $\mathbf{Z}$*, entonces se tiene que*

$$p(\mathbf{X}\mid\mathbf{Y},\mathbf{Z})=p(\mathbf{X}\mid\mathbf{Z}) \tag{1.6}$$

Demostración. 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 1.3. *Si* $\mathbf{X}$ *es independiente de* $\mathbf{Y}$*, entonces* $\mathbf{X}$ *es condicionalmente independiente de* $\mathbf{Y}$ *dado cualquier otro vector* $\mathbf{Z}$*.*

Demostración. 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 solo 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, sería 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.

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; $p(\mathbf{Y} \mid \boldsymbol{\theta})$ es la distribución de muestreo, verosimilitud o distribución de los datos. Por otro lado, la distribución del vector de parámetros condicionada a los datos observados está dada por

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\boldsymbol{\theta}, \mathbf{Y})}{p(\mathbf{Y})} = \frac{p(\boldsymbol{\theta}) p(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})} \tag{1.7}$$

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 (1.7) 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

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{1.8}$$

Gelman et al. (2003) menciona que esta expresión se conoce como la distribución *posterior no-normalizada* y encierra el núcleo técnico de la inferencia bayesiana. La constante $p(\mathbf{Y})$ faltante en la expresión (1.8) se da a continuación.

Resultado 1.4. *La expresión $p(\mathbf{Y})$ corresponde a una constante k tal que*

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

Demostración. 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 nombrado en honor de Thomas Bayes, es casi seguro que él mismo no sospechaba del gran impacto de su resultado. De hecho, aproximadamente una década más tarde, 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 la estadística bayesiana.

A continuación se presenta un ejemplo simple de este sencillo pero poderoso teorema.

Ejemplo 1.4. 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.

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

probabilística es necesario enfocar la 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: $\mathfrak{F} = \{\Omega, \phi, \{\text{Defectuoso}\}, \{\text{No Defectuoso}\}\}$
3. La función de probabilidad:

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

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 probabilidad condicional, se define

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

[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.

[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.

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.

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 1.5. El Grupo de Trabajo de Servicios Preventivos de los Estados Unidos (USPSTF) hizo unas nuevas y controversiales recomendaciones recomendaciones sobre la detección del cáncer de mama 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. Por otro lado, la USPSTF sí recomienda tal práctica de forma bienal en grupos de mujeres de entre 50 y 74 años de edad, puesto que no encontró suficiente evidencia de beneficio o daño adicional en realizar este examen en mujeres mayores a los 74 años. Además, también recomendó *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*. Como información adicional, se sabe 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	40 - 49	50 - 59 años
Cáncer	1/69=0.01449	1/38=0.02632
No cáncer	68/69=0.9855	37/38=0.97368
Positivo | No cáncer	0.0978	0.0866
Negativo | No cáncer	0.9022	0.9134
Positivo | Cáncer	0.999	0.9989
Negativo | Cáncer	0.001	0.0011

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.

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

Los anteriores resultados muestran cómo cambia la probabilidad de tener cáncer al condicionar 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.

Capítulo 2

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.

Inferencia previa

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

$$Pr(\boldsymbol{\theta} \in G) = \int_G p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{2.1}$$

En esta primera etapa también es posible calcular, con fines confirmatorios (Carlin and 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

$$(\#eq : est.prio)\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{2.2}$$

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

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C) = \int_C p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{2.3}$$

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

$$Pr(\boldsymbol{\theta} \in G \mid \mathbf{Y}) = \int_G p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{2.4}$$

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

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \mathbf{Y}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{2.5}$$

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

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C \mid \mathbf{Y}) = \int_C p(\boldsymbol{\theta} \mid \mathbf{Y})\ d\boldsymbol{\theta} \tag{2.6}$$

[1] 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.

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.

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 1.4, la distribución predictiva previa de $\mathbf{Y}$ está dada por

$$p(\mathbf{y}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \, d\boldsymbol{\theta} \tag{2.7}$$

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 (1.7). Por lo tanto, la distribución predictiva posterior de $\mathbf{Y}$ está dada por

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \mathbf{Y}) &= \int p(\tilde{\mathbf{y}}, \boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, d\boldsymbol{\theta} \end{aligned} \tag{2.8}$$

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 1.1 y la última línea se obtiene del resultado 1.2 de la independencia condicional.

2.1. La distribución 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, como podría ser 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 se debe tener en cuenta las implicaciones a la hora

de efectuar cálculos de la estimación puntual o del intervalo de crediblidad, procurando en la mayoría de casos, obtener una distribución posterior fácil de manejar. A continuación exponemos algunos aspectos generales relacionados con las distribuciones previas.

2.1.1. 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, pueden ser comparadas y así ver claramente cómo es la influencia de los datos observados sobre la creencia inicial 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 2.1. 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 and Gamerman (1999) describe los siguientes dos casos en donde esta definición es completamente inútil:

1. *Caso amplio*: sea $\mathcal{P} = \{\text{Todas las distribuciones de probabilidad}\}$ 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

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

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 previas 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.

2.1.2. 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

$$p(Y \mid \theta) = \exp\{d(\theta)T(y) - c(\theta)\}h(y) \tag{2.9}$$

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

$$p(Y \mid \boldsymbol{\theta}) = \exp\{\mathbf{d}(\boldsymbol{\theta})'\mathbf{T}(y) - c(\boldsymbol{\theta})\}h(y) \tag{2.10}$$

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 que a la vez conservan la propiedad de ser distribuciones conjugadas, tal como muestra el siguiente resultado:

Resultado 2.1. *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.*

Demostración. Observando la expresión (2.9), 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,

$$p(\theta \mid \alpha) \propto \exp\{w(\alpha)d(\theta) - \delta c(\theta)\}, \tag{2.11}$$

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

$$p(\theta \mid \alpha) = \frac{1}{k(\alpha, \delta)} \exp\{w(\alpha)d(\theta) - \delta c(\theta)\}, \tag{2.12}$$

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

$$p(\theta \mid \alpha) = \exp\{\underbrace{w(\alpha)}_{d(\alpha)}\underbrace{d(\theta)}_{T(\theta)} - \underbrace{\ln k(\alpha, \delta)}_{c(\alpha)}\}\underbrace{\exp\{-\delta c(\theta)\}}_{h(\theta)}. \tag{2.13}$$

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 puede ser extendedida para el caso en el que se cuenta con una muestra aleatoria de observaciones, tal como se expone a continuación:

Resultado 2.2. *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.*

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

$$p(\mathbf{Y} \mid \theta) = \exp\left\{ d(\theta) \sum_{i=1}^{n} T(y_i) - nc(\theta) \right\} \prod_{i=1}^{n} h(y_i) \tag{2.14}$$

la cual hace parte de la familia exponencial. □

Otra extensión del resultado 2.1 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 2.3. *Sea* 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.*

Demostración. En primer lugar, la distribución de probabilidad de Y perteneciente a la familia exponencial multiparamétrica está dada por (2.10). Siguiendo el mismo razonamiento de la demostración del Resultado 2.1, la distribución previa del parámetro de interés debe estar definida de la siguiente manera

$$p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \exp\left\{ \underbrace{w(\boldsymbol{\eta})'}_{\mathbf{d}(\boldsymbol{\eta})} \underbrace{\mathbf{d}(\boldsymbol{\theta})}_{\mathbf{T}(\boldsymbol{\theta})} - \underbrace{\ln k(\boldsymbol{\eta}, \delta)}_{c(\boldsymbol{\eta})} \right\} \underbrace{\exp\{-\delta c(\boldsymbol{\theta})\}}_{h(\boldsymbol{\theta})}, \tag{2.15}$$

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 2.4. *Sean* $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ *una muestra aleatoria con función de densidad conjunta o verosimilitud dada por* (2.10)*. Bajo este escenario la familia exponencial multi-paramétrica es conjugada con respecto a sí misma.*

Demostración. 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 2.5. *Sea* Y *una variable aleatoria con función de densidad perteneciente a la familia exponencial, dada por* (2.9)*. Sea* θ *el parámetro de interés con distribución previa en la familia exponencial biparamétrica. La distribución predictiva previa de* Y *está dada por*

$$
p(Y) = \frac{k(\alpha + T(y), \delta + 1)}{k(\alpha, \delta)} h(y) \tag{2.16}
$$

donde

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

Demostración.

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

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 2.6. *Sea* $\mathbf{Y} = \{Y_1 \ldots, Y_n\}$ *una muestra aleatoria con función de densidad conjunta perteneciente a la familia exponencial, dada por* (2.10)*. Sea* θ *el parámetro de interés con distribución previa exponencial multiparamétrica. La distribución predictiva previa de* $\mathbf{Y}$ *está dada por*

$$p(\mathbf{Y}) = \frac{k(\alpha + T(\mathbf{y}), \delta + n)}{k(\alpha,\beta)} h(\mathbf{y}) \tag{2.17}$$

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

Demostración. La prueba se tiene de inmediato siguiendo los lineamentos de la demostración del anterior resultado. □

Resultado 2.7. *En términos de la distribución predictiva posterior, se tiene que para una sola observación* $\tilde{y}$*, ésta está dada por*

$$p(\tilde{y} \mid Y) = \frac{k(\alpha + T(y) + T(\tilde{y}), \delta + 2)}{k(\alpha + T(y), \delta + 1)} h(\tilde{y}) \tag{2.18}$$

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*

$$p(\tilde{\mathbf{y}} \mid \mathbf{Y}) = \frac{k(\alpha + T(\mathbf{y}) + T(\tilde{\mathbf{y}}), \delta + n + n^*)}{k(\alpha + T(\mathbf{y}), \delta + n)} h(\tilde{\mathbf{y}}) \tag{2.19}$$

Demostración. De la definición de distribución predictiva posterior dada por la expresión (2.8) 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. □

2.1.3. Distribuciones previas 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 previas que sean no informativas. Es decir, 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. Por tanto la pregunta de interés que surge en este instante es: ¿cómo seleccionar distribuciones previas no informativas[2] sobre el parámetro de interés?

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

[2]Existen muchas denominaciones para las distribuciones uniformes que no son informativas. Por ejemplo, Box and Tiao (1992) 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.

conveniente puesto que sería una distribución impropia en el espacio de muestreo del parámetro de interés. Es decir

$$\text{Si } p(\theta) \propto k\ I_{\Theta}(\theta), \text{ entonces } \int_{\Theta} p(\theta)\ d(\theta) \longrightarrow \infty$$

donde Θ denota espacio de muestreo del parámetro θ e 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

$$p(\phi) = p(\theta) \mid \frac{d\theta}{d\phi} \mid = p(\theta) \mid h'(\theta) \mid^{-1} \tag{2.20}$$

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 2.1. 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 (2.20) 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 2.2. 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

$$p(\theta) \propto (I(\theta))^{1/2} \tag{2.21}$$

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

$$p(\theta) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2} \tag{2.22}$$

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 manifiesten su incorformidad 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 2.8. *La distribución previa no informativa de Jeffreys es invariante a transformaciones uno a uno. Es decir, si $\phi = h(\theta)$, entonces $p(\phi) \propto (I(\phi))^{1/2}$.*

Demostración. En primer lugar nótese que

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

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

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

Ahora, de la definición de función de distribución para una función y utilizando (2.20), 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 and Tiao (1992, p. 59) es posible encontrar un resumen exhaustivo de distribuciones previas no informativas para las distribuciones de verosimilitud más comunes. A continuación, se exponen algunos ejemplos que utilizan este enfoque.

Ejemplo 2.2. 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

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

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 2.3. 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 2.4. 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

$$\begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{2.23}$$

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

$$p(\theta, \sigma^2) \propto 1/\sigma^3 \tag{2.24}$$

2.2. 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 dos estadísticos, uno frecuentista clásico y el otro 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 la hipótesis nula.
- *Razonamiento del bayesiano*: 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 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 la hipótesis nula.

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 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. .

2.2.1. Factor de Bayes

El juzgamiento de hipótesis del enfoque frecuentista se puede efectuar en el ámbito Bayesiano por medio del contraste entre dos modelos. 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. Es posible demostrar que este factor es equivalente a la siguiente expresión:

$$FB = \frac{p(\mathbf{Y} \mid M1)}{p(\mathbf{Y} \mid M2)} = \frac{Pr(M1 \mid \mathbf{Y})/Pr(M2 \mid \mathbf{Y})}{Pr(M1)/Pr(M2)} \tag{2.25}$$

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 (2.25) se claro que valores grandes del factor muestran evidencia a favor del modelo $M1$; valores menores de 1, a favor del modelo $M2$; mientras que valores cercanos a 1 no muestran evidencias claras hacia ninguno de los dos modelos.

En Gelman et al. (1995) se presenta el siguiente ejemplo sencillo sobre la presencia o ausencia de la enfermedad de la hemofilia, una enfermedad genética especialmente grave en 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 $Pr(\theta = 1) = Pr(\theta = 0) = 0.5$. 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, el factor de Bayes se expresa como:

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

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

2.2.2. 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 establecido, para tomar una decisión acerca de la hipótesis nula. En el ámbito Bayesiano, el valor-p se define como la probabilidad de que la estadística de prueba T calculada 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 la siguiente expresión:

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

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 replicados si el modelo fuera verdadero.

2.3. Criterios de información

Los criterios de información constituyen una herramienta muy importante en el modelamiento estadístico, pues contribuyen 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.

2.3.1. Criterio DIC

El criterio de información de *devianza* (DIC, por sus 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

$$D(y, \boldsymbol{\theta}) = -2 * \log(p(y|\boldsymbol{\theta})) \tag{2.26}$$

cuya media posterior es una medida usual del ajuste del modelo. Dempster (1974) sugirió graficar la distribución posterior 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 mediposterior 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 mediposterior de la devianza menos la devianza de las medias posterior (Spiegelhalter et al., 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 el DIC, no es invariante a parametrizaciones del modelo. Se sugiere en la práctica usar parametrizaciones que conducen a la normalidad en la distribución posterior.

2.3.2. Criterios AIC y BIC

El criterio de información de Akaike (AIC) fue formalmente presentado por 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.

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.

Capítulo 3

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 programación en `R` y en `STAN` junto con ejemplos prácticos que permiten la familiarización e interiorización del ambiente computacional de este software que será indispensable en el desarrollo de capítulos posteriores.

3.1. Modelo Bernoulli

Suponga que Y es una variable aleatoria con distribución Bernoulli dada por:

$$p(Y \mid \theta) = \theta^y (1-\theta)^{1-y} I_{\{0,1\}}(y) \tag{3.1}$$

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. Puesto que la distribución uniforme es un caso particular de la distribución Beta, entonces se iniciará con ésta. Por lo tanto la distribución previa del parámetro θ estará dada por

$$p(\theta \mid \alpha, \beta) = \frac{1}{Beta(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} I_{[0,1]}(\theta). \tag{3.2}$$

Bajo este marco de referencia se tienen los siguientes resultados

Resultado 3.1. *La distribución posterior del parámetro θ sigue una distribución*

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

Demostración.

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

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 distribuciones Beta es conjugada con respecto a la familia de distribuciones Bernoulli. Ahora consideremos cuál sería la distribución previa no informativa de Jeffreys para el parámetro θ. De acuerdo a la definición 2.2, se tiene que

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

En donde $I(\theta)$ es la información de Fisher 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, la distribución previa no informativa de Jeffreys debe ser proporcional a $\theta^{-1/2}(1-\theta)^{-1/2}$, que asimismo corresponde a la distribución $Beta(1/2, 1/2)$, cuya función de densidad se muestra en la figura 3.1 la cual asigna iguales pesos a los valores extremos del parámetro de interés y su característica de ser no informativa se representa en la simetría de la función alrededor del valor 0.5.

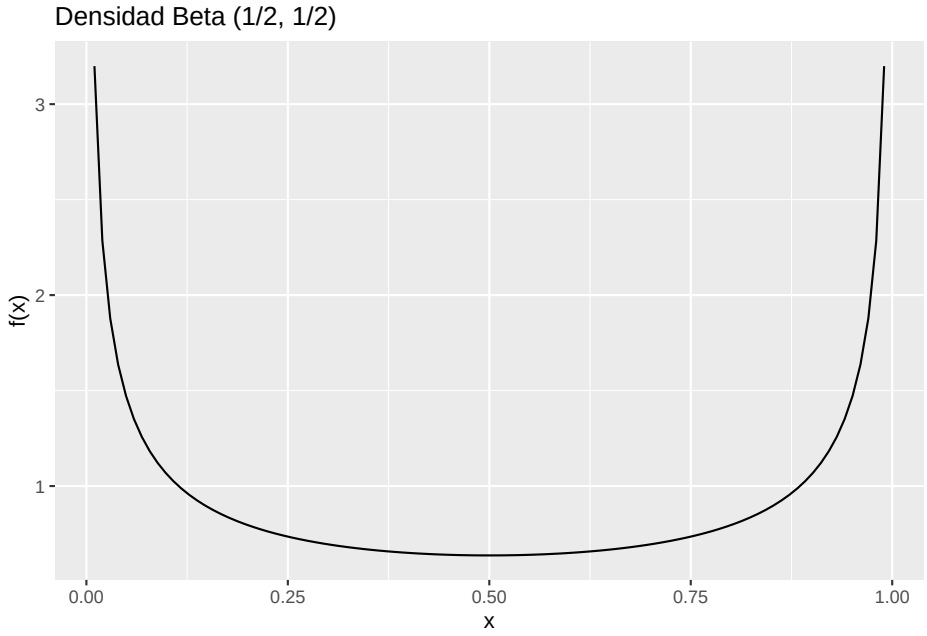


Figura 3.1: Distribución previa no informativa de Jeffreys para el parámetro de una distribución Bernoulli

Resultado 3.2. *La distribución predictiva previa para una observación y está dada por*

$$p(Y) = \frac{Beta(y+\alpha, \beta - y + 1)}{Beta(\alpha, \beta)} I_{\{0,1\}}(y) \tag{3.3}$$

La cual define una auténtica función de densidad de probabilidad continua.

Demostración. 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 expresión, 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 (3.3) 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 la inferencia en la distribución posterior del parámetro; esta distribución se da en el siguiente resultado.

Resultado 3.3. *Después de la recolección de los datos, la distribución predictiva posterior para una nueva observación $\tilde{y}$ está dada por*

$$p(\tilde{y} \mid Y) = \frac{Beta(\tilde{y} + y + \alpha, \beta - \tilde{y} - y + 2)}{Beta(y + \alpha, \beta - y + 1)} I_{\{0,1\}}(\tilde{y}), \tag{3.4}$$

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

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

□

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 3.4. *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*

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

Ejemplo 3.1. 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 el candidato presidencial A está interesado en conocer su intención de voto previa a las elecciones. Para esto, él contrata a una firma encuestadora para la realización de una encuesta 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 A?

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 %, 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 % 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 el electorado.

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 Beta, definiendo los parámetros de la distribución previa como α igual al número de votantes a favor y β igual al número de votantes en contra. Es decir, $Beta(\alpha = 350, \beta = 650)$. Por lo anterior, 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.

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 $\frac{6350.5}{12001} = 0.529$.

La figuras 3.2 muestra 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.

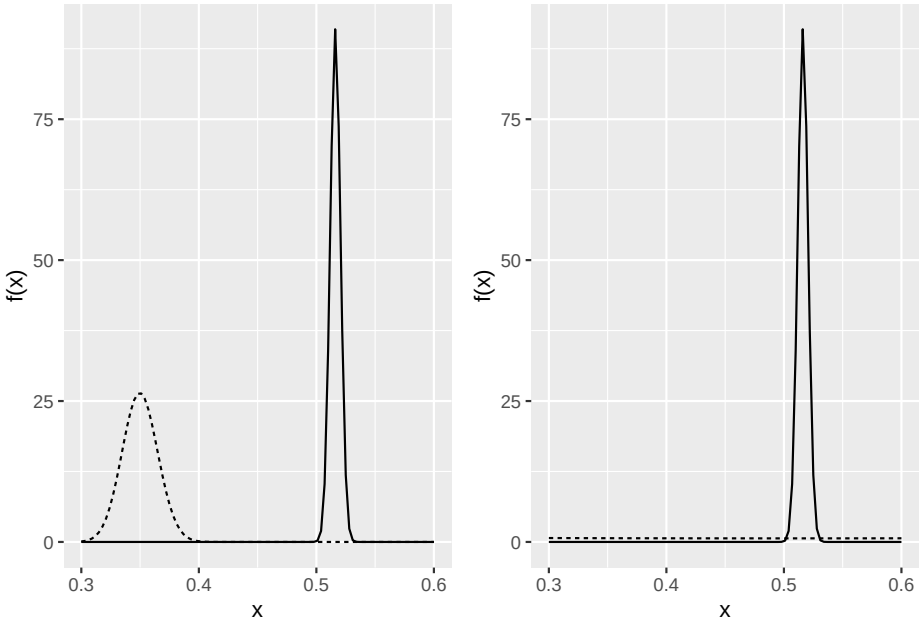


Figura 3.2: Distribuciones previas (línea punteada) y posteriores (línea sólida) para el ejemplo de las encuestas electorales.

Utilizando el siguiente código en R, es posible conocer los intervalos de credibilidad para las dos distribuciones posteriores. Además, 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] 0.5075614 0.5247415
```

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

```
## [1] 0.5206678 0.5385340
```

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

```
Bernoulli <- "
data {
  int<lower=0> n;
  int y[n];
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  y ~ bernoulli(theta);
  theta ~ beta(350, 650);
}

"

library(rstan)
options(mc.cores = parallel::detectCores())

n <- 12000
s <- 6350
y <- c(rep(1, s), rep(0, n - s))
sample_data <- list(n = n, y = y)

Berfit <- stan(model_code = Bernoulli,
               data = sample_data, verbose = FALSE)
```

La siguiente salida de STAN permite conocer la estimación bayesiana posterior y los límites del intervalo de credibilidad al 95 %.

```
print(Berfit, pars = "theta",
      digits = 4, probs = c(0.025, 0.975))
```

```
## Inference for Stan model: 62d7c91114fcc6227deb68f6059b2b09.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##         mean se_mean     sd   2.5% 97.5% n_eff   Rhat
## theta 0.5155   1e-04 0.0042 0.5071 0.524  1591 1.0009
##
```

```
## Samples were drawn using NUTS(diag_e) at Sat Jun  5 00:10:46 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

3.2. 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:

$$p(S \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s} I_{\{0,1,\ldots,n\}}(s), \tag{3.5}$$

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 (3.2). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 3.5. *La distribución posterior del parámetro θ sigue una distribución*

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

Demostración.

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

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 esperanza de la distribución posterior, dada por

$$\hat{\theta}_B = \frac{s+\alpha}{n+\alpha+\beta} \tag{3.6}$$

En la práctica, se acostumbra a escoger los hiperparámetros α y β de tal forma que correspondan respectivamente al número de éxitos y fracasos obtenidos en datos que pudieron ser recolectados previamente. 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 (3.6) de alguna forma combina el estimador clásico con el estimador previo. 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 figura 3.3 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[1] con máximo en $8/10$, entonces la distribución posterior estará centrada en $10/17$; es decir, la estimación bayesiana se encuentra situada entre la estimación previa y la estimación clásica.

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 parametrizada 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 3.4 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.

[1] 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.

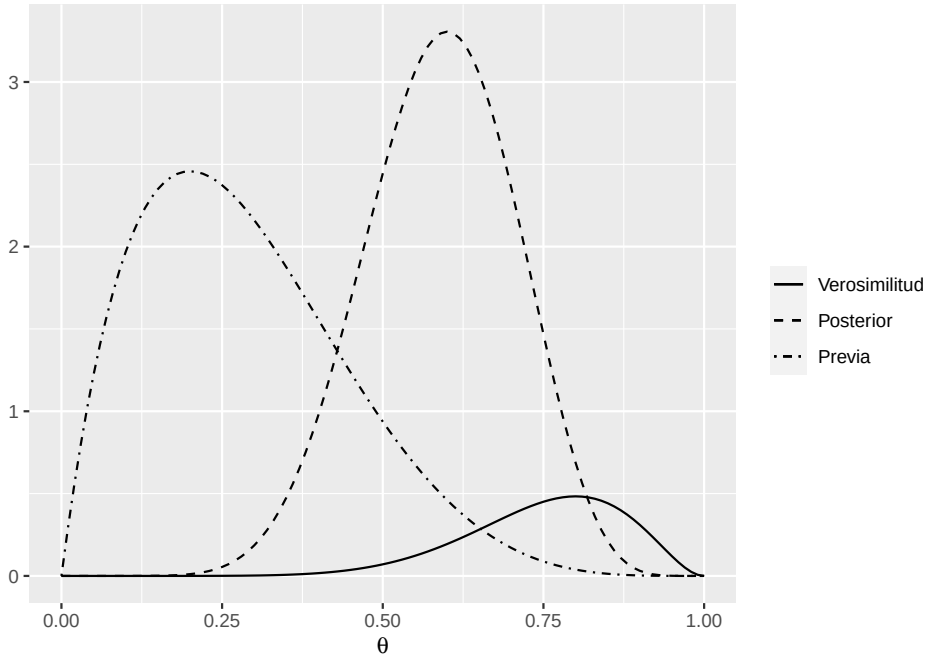


Figura 3.3: Funciones de verosimilitud, previa y posterior para $\alpha = 2$, $\beta = 5$, $s = 8$ y $n = 10$.

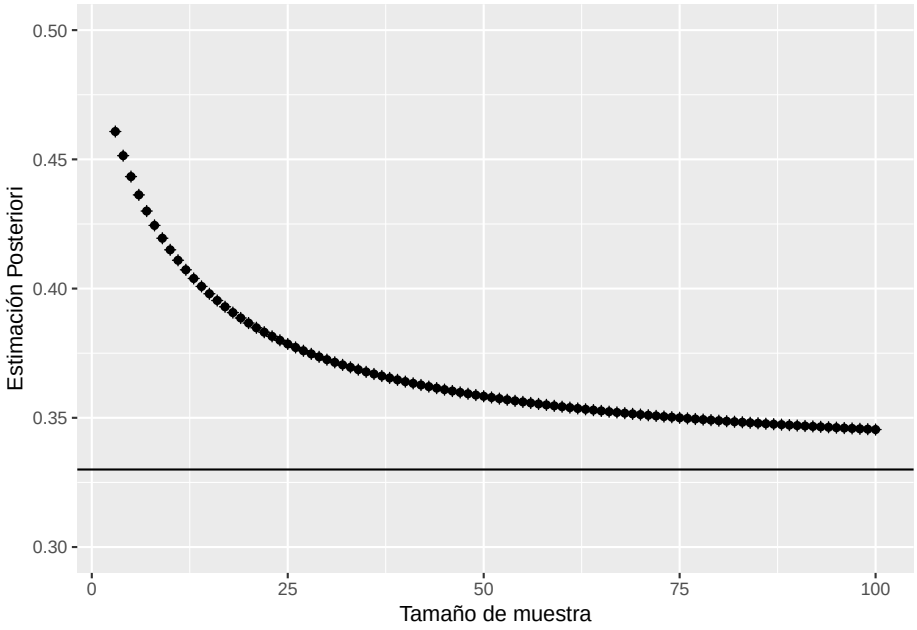


Figura 3.4: Estimación posterior de θ para diferentes valores de n y s con $\alpha = \beta = 5$.

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 previa 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

$$\hat{\theta}_B = \frac{s + 1/2}{n + 1} \tag{3.7}$$

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 (3.7) en situaciones donde el valor teórico de θ es muy pequeño, y como consecuencia $s = 0$ en la muestra. Por ejemplo, cuando θ representa el porcentaje de personas que están 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 $\hat{\theta}_B = \frac{0.5}{n+1}$ tiende a un porcentaje muy pequeño a medida que aumente el tamaño muestral n, pero nunca llega a ser nulo.

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

Resultado 3.6. *La distribución predictiva previa para la observación particular de la suma de variables aleatorias Bernoulli, s, está dada por una distribución Beta-Binomial*

$$p(S) = \binom{n}{s} \frac{Beta(s+\alpha, \beta - s + n)}{Beta(\alpha, \beta)} I_{\{0,1,\ldots,n\}}(s) \tag{3.8}$$

Demostración. 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 3.7. *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*

$$p(\tilde{s} \mid S) = \binom{\tilde{n}}{\tilde{s}} \frac{Beta(\tilde{s}+s+\alpha, \beta-\tilde{s}-s+n+\tilde{n})}{Beta(s+\alpha, \beta-s+n)} I_{\{0,1,\ldots,\tilde{n}\}}(\tilde{s}), \tag{3.9}$$

Demostración. 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, `R` puede presentar problemas numéricos al momento de calcular directamente estas funciones. 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

$$Pr(\tilde{s} = 500|S) = \binom{800}{500} \frac{Beta(1350, 950)}{Beta(850, 650)} \tag{3.10}$$

y desafortunadamente, en `R` se presenta error al intentar ejecutar `beta(1350,950)` o `beta(850,650)`.

```
beta(1350, 950)
```

```
## [1] 0
```

```
beta(850, 650)
```

```
## [1] 0
```

Por ende, es posible plantear la siguiente solución numérica cuando se quiere calcular la función predictiva (3.9) 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

$$
\begin{aligned}
&\frac{Beta(a,b)}{Beta(c,d)} \\
&= \frac{(a-1)!(b-1)!(c+d-1)!}{(c-1)!(d-1)!(a+b-1)!} \\
&= \frac{(a-1)(a-2)\cdots(a-(a-c))(b-1)(b-2)\cdots(b-(b-d))}{(a+b-1)(a+b-2)\cdots(a+b-(a+b-c-d))} \\
&= \frac{a^{a-c}(1-\frac{1}{a})(1-\frac{2}{a})\cdots(1-\frac{a-c}{a})b^{b-d}(1-\frac{1}{b})(1-\frac{2}{b})\cdots(1-\frac{b-d}{b})}{(a+b)^{a+b-c-d}(1-\frac{1}{a+b})(1-\frac{2}{a+b})\cdots(1-\frac{a+b-c-d}{a+b})} \\
&= \underbrace{\left(\frac{a}{a+b}\right)^{a-c}}_{t_1}\underbrace{\left(\frac{b}{a+b}\right)^{b-d}}_{t_2}\underbrace{(1-\frac{1}{a})(1-\frac{2}{a})\cdots(1-\frac{a-c}{a})}_{t_3} \\
&\quad \underbrace{(1-\frac{1}{b})(1-\frac{2}{b})\cdots(1-\frac{b-d}{b})}_{t_4}\underbrace{(1-\frac{1}{a+b})(1-\frac{2}{a+b})\cdots(1-\frac{a+b-c-d}{a+b})}_{t_5}
\end{aligned}
$$

Calculando separadamente los términos t_1, t_2, t_3, t_4 y t_5 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 (3.9) 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
  return(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, el objeto `res` contiene las 801 probabilidades asociadas a todos los posibles valores de $\tilde{s}$. Los resultados obtenidos con la anterior técnica son equivalentes a lo obtenido usando la función `lbeta` que computa el logaritmo natural de la función beta. Así, para calcular la probabilidad en (3.10), simplemente usamos el siguiente código

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

```
## [1] 0.0005969157
```

Nótese que esta probabilidad es la misma contenida en la posición 501 del objeto `res` igual a 5.969157×10^{-4}. Finalmente, se observa que la distribución predictiva (3.9) corresponde a una distribución Beta-binomial con parámetros $s + \alpha$ y $\beta - s + n$. El paquete `VGAM` (Yee, 2012) en `R` 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 observar que la posición 501 del objeto `res2` es igual a 5.969157×10^{-4}, el cual es idéntico a lo obtenido en `res`. Adicionalmente, al escribir la distribución predictiva de (3.9) 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 multiplicando todos los posibles valores de $\tilde{S}$ con su respectiva probabilidad, y sumando al final, como se muestra a continuación.

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

```
## [1] 453.3333
```

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

```
## [1] 453.3333
```

Retomando el ejemplo 3.1, 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 3.8. *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*

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

Demostración.

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

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 3.2. El siguiente conjunto de datos fue estudiado inicialmente por Efron and 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) establece que, en términos generales, este valor representa la razón entre la cantidad de *hits* y el número de turnos al bate[2]. 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.

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

name	team	league	r	y	n	p
Roberto Clemente	Pitts	NL	18	0.400	367	0.346
Frank Robinson	Balt	AL	17	0.378	426	0.298
Frank Howard	Wash	AL	16	0.356	521	0.276
Jay Johnstone	Cal	AL	15	0.333	275	0.222
Ken Berry	Chi	AL	14	0.311	418	0.273
Jim Spencer	Cal	AL	14	0.311	466	0.270
Don Kessinger	Chi	NL	13	0.289	586	0.263
Luis Alvarado	Bos	AL	12	0.267	138	0.210
Ron Santo	Chi	NL	11	0.244	510	0.269
Ron Swoboda	NY	NL	11	0.244	200	0.230
Del Unser	Wash	AL	10	0.222	277	0.264
Billy Williams	Chi	AL	10	0.222	270	0.256
George Scott	Bos	AL	10	0.222	435	0.303
Rico Petrocelli	Bos	AL	10	0.222	538	0.264
Ellie Rodriguez	KC	AL	10	0.222	186	0.226
Bert Campaneris	Oak	AL	9	0.200	558	0.285
Thurman Munson	NY	AL	8	0.178	408	0.316
Max Alvis	Mil	NL	7	0.156	70	0.200

En la primera columna se tiene el número del jugador, la segunda columna proporciona el nombre del jugador, la cuarta columna representan el número de *hits* en los primeros 45 turnos al bate. La sexta columna representa el número de turnos al bate al final de la temporada regular y la última columna representa el promedio de bateo en la temporada.

Suponga que, partiendo de la muestra de los 18 jugadores, el objetivo es estimar el porcentaje de bateo, notado como θ, en toda la liga 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 bateo es común a todos, puesto que pertenecen a la misma liga profesional. Por lo tanto, es posible establecer

[2]Un *hit* es 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.

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, constituida por la tercera y cuarta columna del archivo de datos. 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 3.8, 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.

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

```
Binomial <- 'data {
  int<lower=0> n;
  int<lower=0> m[n];
  int<lower=0> s[n];
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  for(i in 1:n) {
  s[i] ~ binomial(m[i], theta);
  }
  theta ~ beta(215, 595);
}
'

library(rstan)
options(mc.cores = parallel::detectCores())

s <- round(EfronMorris$n * EfronMorris$p)
sample_data <- list(s = s,
                    n = nrow(EfronMorris),
                    m = EfronMorris$n)
```

```
Binfit <- stan(model_code = Binomial,
               data = sample_data, verbose = FALSE)
```

La siguiente salida de STAN permite conocer la estimación bayesiana posterior y los límites del intervalo de credibilidad al 95 %.

```
print(Binfit, pars = "theta",
      digits = 4, probs = c(0.025, 0.975))
```

```
## Inference for Stan model: b5a37600d5b0f80332bf311eb740e4c6.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##         mean se_mean     sd  2.5%  97.5% n_eff  Rhat
## theta 0.2736   1e-04 0.0051 0.264 0.2838  1311 1.003
##
## Samples were drawn using NUTS(diag_e) at Sat Jun  5 00:11:22 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Por otro lado, el mismo intervalo de credibilidad del 95 % correspondiente se puede hallar mediante el siguiente código computacional de R.

```
qbeta(c(0.025, 0.975), 2040, 5419)
```

```
## [1] 0.2634379 0.2836674
```

La figura 3.5 muestra el comportamiento de las distribuciones previa y posterior para este ejemplo. Nótese que, con un análisis frecuentista, se hubiese llegado a una estimación cercana de $\frac{1825}{6649} = 0.274$.

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 3.3. 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 y no tuvo muchos turnos al bate. 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 3.7. Para este análisis, se define la caracterización estructural de la distribución previa del jugador que está dada por una $Beta(\alpha = 7, \beta = 38)$. La siguiente función en R permite obtener la distribución predictiva para este jugador, que se muestra en la figura 3.6.

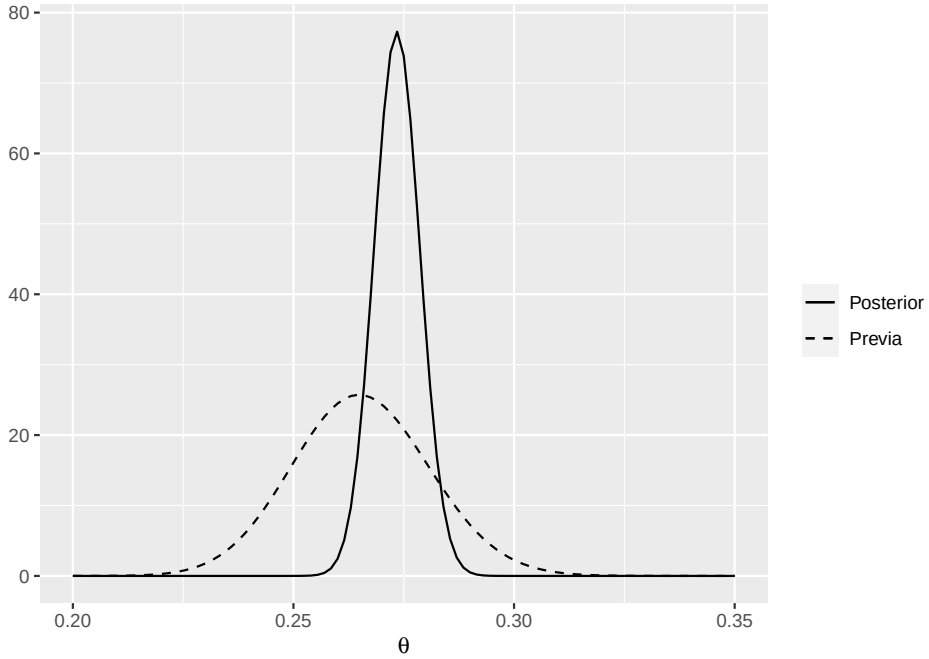


Figura 3.5: Función de densidad previa y función de densidad posterior para el ejemplo de bateo.

```
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)
}

sum(predictiva)
```

```
## [1] 1
```

```
which(predictiva==max(predictiva))
```

```
## [1] 71
```

La última línea del código computacional permite concluir 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.

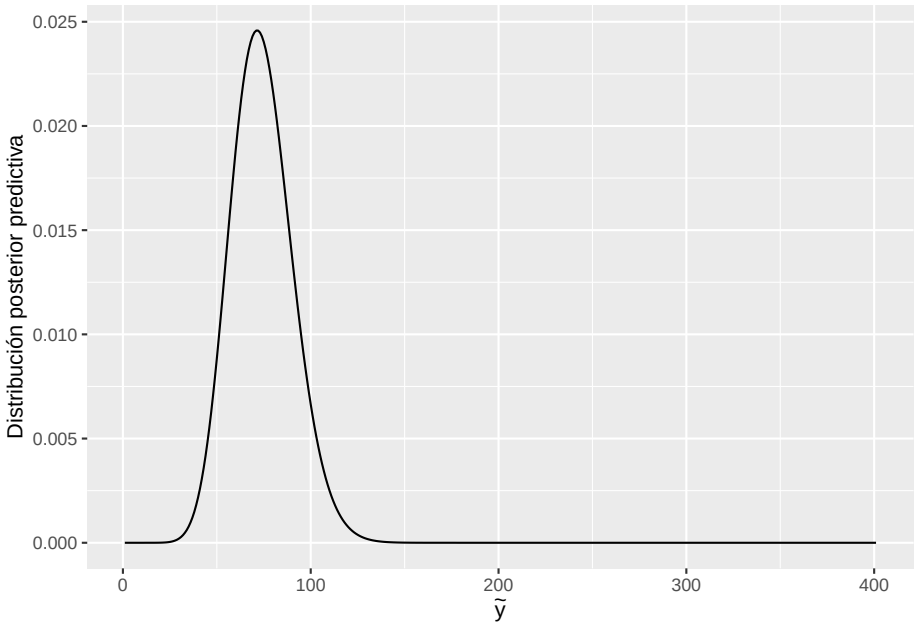


Figura 3.6: Función de densidad predictiva posterior para el jugador Max Alvis.

3.3. Modelo Binomial negativo

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. Esta distribución es particularmente útil cuando el parámetro θ que se quiere estimar es muy pequeño, como la proporción de una población que padece de alguna enfermedad rara. La razón por la que no se utiliza la distribución binomial es que al fijar el número de ensayos n, con un aprobabilidad θ muy pequeña, 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, y 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

$$p(Y \mid \theta) = \binom{y-1}{k-1}\theta^k(1-\theta)^{y-k}I_{\{k,k+1,\ldots,\}}(y), \tag{3.11}$$

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 (3.2). Bajo este marco de referencia se tienen los siguientes resultados

Resultado 3.9. *La distribución posterior del parámetro θ sigue una distribución*

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

Demostración.

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

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 3.10. *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*

$$\theta \mid Y_1, \ldots, Y_n \sim Beta(\alpha + \sum_{i=1}^{n} k_i, \beta + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} k_i) \tag{3.12}$$

Ejemplo 3.4. Una franquicia de investigación farmacéutica ha desarrollado un nuevo tratamiento farmacológico sobre pacientes diabéticos que padezcan, a su vez, de enfermedades cardíacas o cardiopatías (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 muestra 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 parámetros 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 `STAN`, haciendo la salvedad de que `STAN` define esta distribución en términos del número de fracasos $m = y - k$ necesarios para obtener k éxitos.

```
BinNegativa <- 'data {
  int<lower=0> k;
  int<lower=0> y;
}
transformed data {
  int<lower=0> m;
  m = y - k;
}
```

```
parameters {
  real<lower=0> beta;
}
transformed parameters {
  real<lower=0> theta;
  theta = beta/(beta + 1);
}
model {
  m ~ neg_binomial(k, beta);
  theta ~ beta(0.5, 0.5);
}
'

y <- 1106
k <- 5
sample_data <- list(k = k, y = y)

BNfit <- stan(model_code = BinNegativa,
              data = sample_data, verbose = FALSE)
```

La siguiente salida de `STAN` permite conocer la estimación bayesiana posterior y los límites del intervalo de credibilidad al 95 %.

```
print(BNfit, pars = "theta",
      digits = 4, probs = c(0.025, 0.975))
```

```
## Inference for Stan model: ca45da8b0e94b143378403f155105e09.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##         mean se_mean     sd   2.5% 97.5% n_eff   Rhat
## theta 0.0051   1e-04 0.0021 0.0019  0.01  1720 1.0009
##
## Samples were drawn using NUTS(diag_e) at Sat Jun  5 15:23:38 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Después de las iteraciones necesarias, la salida del anterior código muestra la estimación puntual dada por 0.00498 y un intervalo de credibilidad al 95 %, dado por $(0.00174, 0.01013)$.

Ejemplo 3.5. 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. 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[3] no informativa $Beta(0.5, 0.5)$, se llega a que la distribución posterior del parámetros 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 a una proporción de 0.51 % de personas con estas enfermedades. El siguiente código computacional muestra cómo se puede llegar a las mismas conclusiones con `STAN`

[3]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.

```
BinNegativa2 <- 'data {
  int<lower=0> n;
  int<lower=0> k[n];
  int<lower=0> y[n];
}
transformed data {
  int<lower=0> m[n];
  for(i in 1:n){
    m[i] = y[i] - k[i];
  }
}
parameters {
  real<lower=0> b;
}
transformed parameters {
  real<lower=0> theta;
  theta = b/(b + 1);
}
model {
  for(i in 1:n){
    m[i] ~ neg_binomial(k[i], b);
  }
  theta ~ beta(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)
sample_data <- list(k = k, y = y, n = length(y))

BNfit2 <- stan(model_code = BinNegativa2,
               data = sample_data, verbose = FALSE)
```

Después de cinco mil iteraciones, la salida del anterior código muestra la estimación puntual dada por 0.00515 y un intervalo de credibilidad al 95 %, dado por $(0.00439, 0.00603)$, mucho más estrecho que el intervalo de credibilidad del anterior ejemplo

```
print(BNfit2, pars = "theta",
      digits = 4, probs = c(0.025, 0.975))
```

```
## Inference for Stan model: 756b8df22716f40c50aa8045790299a4.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##         mean se_mean    sd   2.5% 97.5% n_eff   Rhat
## theta 0.0051       0 4e-04 0.0044 0.006  1325 1.0008
##
## Samples were drawn using NUTS(diag_e) at Sat Jun  5 15:24:08 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

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

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

Demostración.

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

□

Ejemplo 3.6. Siguiendo con los datos del ejemplo 3.5, suponga que se quiere recolectar información de tres pacientes con cardiopatía en cierta ciudad, y se quiere conocer acerca del número de entrevistas necesarias para . 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.

```
BNpred <- 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 <- sum(k)
n <- sum(y)
k <- 5

fun <- rep(NA)
for(y in 5:5000){
  fun[y - 4] <- BNpred(y, alfa, beta, s, n, k)
}

sum(fun)
```

```
## [1] 0.9999994
```

Se puede ver que el número de entrevistas que tiene mayor probabilidad asociadas es el valor 768, usando el comando `which(fun == max(fun))`. También, se puede observar que la probabilidad de que en menos de 500 entrevistas se encuentren los 5 pacientes es de solo el 0.1200985 usando el comando `sum(fun[1:(500 - 4)])`

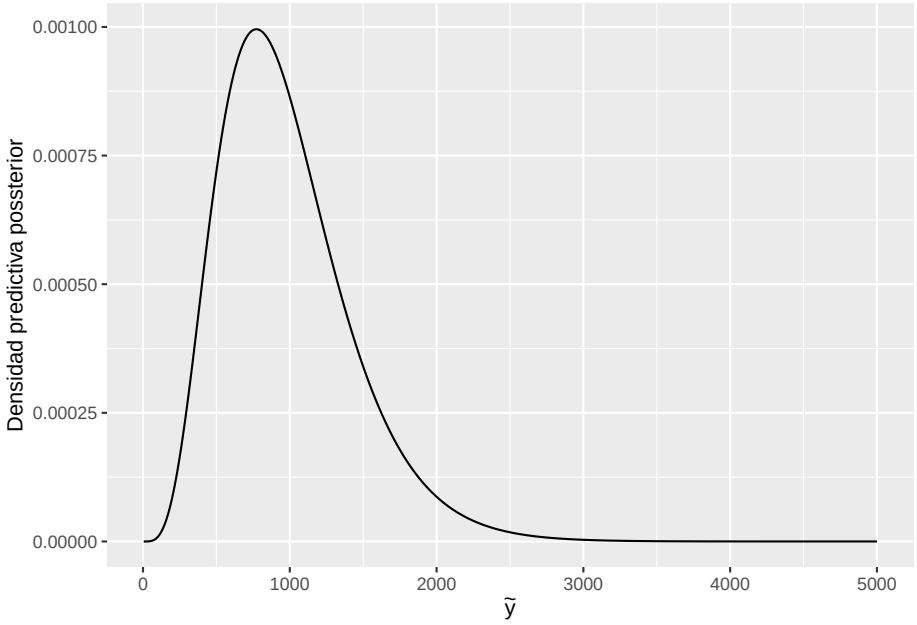


Figura 3.7: Distribución predictiva posterior para el número de entrevistas necesarias para encontrar 5 pacientes.

Apéndice A

Elementos de probabilidad

ss

A.1. Distribuciones discretas

A.1.1. Distribución uniforme discreta

Definición A.1. Una variable aleatoria Y tiene distribución uniforme discreta sobre el conjunto $\{1, 2, \cdots, N\}$ si su función de densidad está dada por:

$$f_Y(y) = \frac{1}{N} I_{\{1,2,\cdots,N\}}(y) \tag{A.1}$$

Esta distribución describe situaciones donde los resultados de un experimento aleatorio tienen la misma probabilidad de ocurrencia. Entre los ejemplos de la distribución uniforme discreta en la vida práctica están el lanzamiento de una moneda corriente, el lanzamiento de un dado corriente, la extracción de una urna que contiene bolas enumeradas de 1 a N.

Resultado A.1. *Si Y es una variable aleatoria con distribución uniforme discreta sobre el conjunto $\{1, 2, \cdots, N\}$, entonces:*

- $E(Y) = \frac{N+1}{2}$.
- $Var(Y) = \frac{N^2-1}{12}$.
- $m_Y(t) = \sum_{i=1}^{N} \frac{e^{ti}}{N}$.

A.1.2. Distribución hipergeométrica

Definición A.2. Una variable aleatoria Y tiene distribución hipergeométrica con parámetros n, R y N si su función de densidad está dada por:

$$f_Y(y) = \frac{\binom{R}{y}\binom{N-R}{n-y}}{\binom{N}{n}} I_{\{0,1,\cdots,n\}}(y), \tag{A.2}$$

y se nota como $Y \sim Hg(n, R, N)$.

Suponga que en una urna hay N bolas en total, donde R de ellas son del color negro y los $N - R$ son del color blanco, se extrae aleatoriamente n bolas de la urna ($n < N$), entonces la variable "número de bolas negras extraídas" tiene distribución hipergeométrica con parámetros n, R y N. Otro uso de la distribución hipergeométrica es el problema de captura-recaptura.

Resultado A.2. *Si Y es una variable aleatoria con distribución hipergeométrica con parámetros n, R y N, entonces:*

- $E(Y) = \frac{nR}{N}$.
- $Var(Y) = \frac{nR(N-R)(N-n)}{N^2(N-1)}$.

El anterior resultado no incluye la función generadora de momentos, pues éste no ha resultado ser útil en la teoría relacionada con la distribución hipergeométrica.

A.1.3. Distribución Bernoulli

La distribución Bernoulli debe su nombre al matemático suizo Jacob Bernoulli (1654-1705) que describe el éxito o fracaso de un evento.

Definición A.3. Una variable aleatoria Y tiene distribución Bernoulli con parámetro $p \in (0, 1)$ si su función de densidad está dada por:

$$f_Y(y) = p^y(1-p)^{1-y} I_{\{0,1\}}(y), \tag{A.3}$$

y se nota como $Y \sim Ber(p)$.

Resultado A.3. *Si Y es una variable aleatoria con distribución Bernoulli con parámetro p, entonces:*

- $E(Y) = p$.
- $Var(Y) = p(1-p)$.
- $m_Y(t) = pe^t + 1 - p$.

A.1.4. Distribución binomial

Definición A.4. Una variable aleatoria Y tiene distribución binomial con los parámetros $n \in \mathbb{N}$ y $p \in (0, 1)$ si su función de densidad está dada por:

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} I_{\{0,1,\cdots,n\}}(y), \tag{A.4}$$

y se nota como $Y \sim Bin(n, p)$.

Resultado A.4. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes e idénticamente distribuidas con distribución Bernoulli con parámetro p, entonces la variable $\sum_{i=1}^n Y_i$ tiene distribución $Bin(n, p)$. Por ende, la distribución Bernoulli es un caso particular de la distribución binomial cuando $n = 1$.*

Demostración. La demostración radica en el hecho de que la función generadora de momentos caracteriza la distribución probabilística, entonces basta demostrar que la función generadora de momentos de $\sum_{i=1}^n X_i$ es la de una distribución $Bin(n, p)$. Tenemos lo siguiente:

$$\begin{aligned} m_{\sum Y_i}(t) = E(e^{\sum tY_i}) &= E(\prod_{i=1}^n e^{tY_i}) \\ &= \prod_{i=1}^n E(e^{tY_i}) \quad \text{(por independencia)} \\ &= \prod_{i=1}^n (pe^t + 1 - p) \quad \text{(definición de } m_{Y_i}(t)) \\ &= (pe^t + 1 - p)^n \end{aligned}$$

□

Una aplicación de esta distribución es cuando tenemos un número n de repeticiones independientes de un experimento donde cada uno tiene dos posibles resultados que se podrían llamarse como éxito o fracaso y donde la probabilidad de éxito p es constante en cada una de las repeticiones. Por tanto, la variable número de éxitos obtenidos en las n repeticiones tiene distribución $Bin(n, p)$. La distribución binomial tiene dos parámetros, n y p; sin embargo, cuando n es conocido, la distribución dependerá sólo del valor p que sería el único parámetro con espacio paramétrico $\Theta = (0, 1)$.

Resultado A.5. *Si Y es una variable aleatoria con distribución binomial con parámetros n y p, entonces*

- $E(Y) = np$.
- $Var(Y) = np(1-p)$.
- $m_Y(t) = (pe^t + 1 - p)^n$.

A.1.5. Distribución Binomial negativa

Definición A.5. Una variable aleatoria Y tiene distribución Binomial negativa con parámetros (θ, r) si su función de densidad está dada por:

$$P(y \mid \theta, r) = \frac{\Gamma(r + y_i)}{y_i!\Gamma(r)}\theta^r(1-\theta)^{1-y_i}I_{(0,1,2,\ldots)}(y) \tag{A.5}$$

Esta distribución siempre ha tenido lugar al resolver el problema del número de ensayos necesarios para lograr un número específico de éxitos. Por supuesto, si r es el número de éxitos necesarios y se conoce que la probabilidad de éxito es θ, entonces la distribución binomial negativa corresponde a un modelo probabilístico, afianzado durante siglos, que permite la resolución de este tipo de situaciones.

Por otro lado, es posible asignar al parámetro r valores que sean reales; en este caso no hay ninguna interpretación práctica en el contexto del número de ensayos necesarios para determinados éxitos. Sin embargo, en términos de distribución, r es un parámetro más. Esto nos lleva a uno de los verdaderos usos prácticos de esta distribución: la sobredispersión. Dado que la forma funcional de arriba corresponde a una generalización de la función de distribución Poisson, entonces es posible suponer que los datos de conteo vienen de una distribución binomial negativa.

Lo anterior trae ventajas puesto que, si la media de los datos recolectados no corresponde con la varianza (característica esencial de la Poisson), entonces cualquier modelo que de allí surgiese sería altamente cuestionable. Si lo anterior se presenta es mejor acudir a la distribución binomial negativa dando valores reales al parámetro r.

Resultado A.6. *Si Y es una variable aleatoria con distribución binomial-negativa con parámetros (θ, r), entonces*

- $E(Y) = \frac{r\theta}{1-\theta}$.
- $Var(Y) = \frac{r\theta}{(1-\theta)^2}$.
- $m_Y(t) = \left(\frac{1-\theta}{1-\theta e^t}\right)^r$.

A.1.6. Distribución de Poisson

La distribución de Poisson debe su nombre al francés Siméon-Denis Poisson (1781-1840) quien descubrió esta distribución en el año 1838, cuando la usó para describir el número de ocurrencias de algún evento durante un intervalo de tiempo de longitud dada.

Definición A.6. Una variable aleatoria Y tiene distribución de Poisson con parámetros $\lambda > 0$ si su función de densidad está dada por:

$$f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!} I_{\{0,1,\cdots\}}(y) \tag{A.6}$$

y se nota como $Y \sim P(\lambda)$.

Nótese que la distribución Poisson tiene solo un parámetro $\theta = \lambda$, y el espacio paramétrico es $\Theta = (0, \infty)$.

Resultado A.7. *Si Y es una variable aleatoria con distribución Poisson con parámetro λ, entonces*

- $E(Y) = \lambda$.
- $Var(Y) = \lambda$.
- $m_Y(t) = \exp\{\lambda(e^t - 1)\}$.

Resultado A.8. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes con distribución $P(\lambda_i)$ para $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n X_i$ tiene distribución $P(\sum_{i=1}^n \lambda_i)$.*

A.2. Distribuciones continuas

A.2.1. Distribución Uniforme Continua

Definición A.7. Una variable aleatoria Y tiene distribución uniforme continua sobre el intervalo $[a, b]$ con $a < b$ si su función de densidad está dada por:

$$f_Y(y) = \frac{1}{b-a} I_{[a,b]}(y) \tag{A.7}$$

Resultado A.9. *Si Y es una variable aleatoria con distribución uniforme continua sobre $[a, b]$, entonces*

- $E(Y) = \frac{a+b}{2}$.
- $Var(Y) = \frac{(b-a)^2}{12}$.
- $m_Y(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$.

A.2.2. Distribución Weibull

Definición A.8. Una variable aleatoria Y tiene distribución uniforme continua sobre los reales positivos si su función de densidad está dada por:

$$p(Y \mid \theta, \gamma) = \frac{\theta}{\gamma^\theta} y^{\theta-1} exp\left\{-\frac{y^\theta}{\gamma^\theta}\right\} I_{[0,\infty)}(y) \tag{A.8}$$

Resultado A.10. *Si Y es una variable aleatoria con distribución Weibull, entonces*

- $E(Y) = \gamma\Gamma\left(1 + \frac{1}{\theta}\right)$.
- $Var(Y) = \gamma^2\left[\Gamma\left(1 + \frac{2}{\theta}\right) + \Gamma^2\left(1 + \frac{1}{\theta}\right)\right]$.
- $m_Y(t) = \sum_{n=0}^{\infty} \frac{t^n\gamma^n}{n!}\Gamma\left(1 + \frac{n}{\theta}\right),\ \theta \geq 1$.

A.2.3. Distribución valor-extremo

Definición A.9. Una variable aleatoria Y tiene distribución valor-extremo si su función de densidad está dada por:

$$p(y \mid \theta, \lambda) = \theta \exp(\theta y) \exp\{\lambda - \exp(\lambda + \theta y)\} \tag{A.9}$$

Resultado A.11. *Si Y es una variable aleatoria con distribución valor-extremo, entonces*

- $E(Y) = -\frac{\lambda}{\theta} - \frac{\epsilon}{\theta}$.
- $Var(Y) = \frac{\pi^2}{6\theta^2}$.

Donde $\pi \approx 3.1416$ es el número Pi y $\epsilon = 0.5772$ es la constante de Euler.

A.2.4. Distribución Gamma

Definición A.10. Una variable aleatoria Y tiene distribución Gamma con parámetro de forma $\alpha > 0$ y parámetro de escala $\theta > 0$ si su función de densidad está dada por:

$$p(\theta \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}I_{(0,\infty)}(\theta). \tag{A.10}$$

donde $\Gamma(k) = \int_0^\infty u^{k-1}\exp(-u)\ du$.

La distribución Gamma tiene dos parámetros: α y β, en este caso, el vector de hiper-parámetros es $\boldsymbol{\theta} = (\alpha, \theta)'$ donde el espacio paramétrico está dado por $\Theta = (0, \infty) \times (0, \infty)$.

Resultado A.12. *Si Y es una variable aleatoria con distribución Gamma con parámetro de forma α y parámetro de escala θ, entonces*

- $E(Y) = \alpha/\beta$.
- $Var(Y) = \alpha/\theta^2$.

Resultado A.13. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes con distribución Gamma con parámetro de forma α_i y parámetro de escala β para $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n X_i$ tiene distribución Gamma con parámetro de forma $\sum_{i=1}^n \alpha_i$ y parámetro de escala θ.*

A.2.5. Distribución Gamma-inversa

Definición A.11. Una variable aleatoria Y tiene distribución Gamma-inversa con parámetro de forma $\alpha > 0$ y parámetro de escala $\beta > 0$ si su función de densidad está dada por:

$$p(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{-\alpha-1} e^{-\beta/y} I_{(0,\infty)}(y). \tag{A.11}$$

donde $\Gamma(k) = \int_0^{\infty} u^{k-1} \exp(-u)\, du$.

La distribución Gamma-inversa tiene dos parámetros: α y β; en este caso, el vector de hiper-parámetros es $\boldsymbol{\theta} = (\alpha, \beta)'$ donde el espacio paramétrico está dado por $\Theta = (0, \infty) \times (0, \infty)$.

Resultado A.14. *Si Y es una variable aleatoria con distribución Gamma-inversa con parámetro de forma α y parámetro de escala β, entonces*

- $E(Y) = \beta/(\alpha - 1)$.
- $Var(Y) = \theta^2/(\alpha - 1)^2(\alpha - 2)$.

Resultado A.15. *Si X es una variable aleatoria con distribución $Gamma(\alpha, \beta)$, entonces $1/X$ tiene distribución $Gamma - inversa(\alpha, 1/\beta)$.*

A.2.6. Distribución exponencial

Definición A.12. Una variable aleatoria Y tiene distribución exponencial con parámetro de escala $\theta > 0$ si su función de densidad está dada por:

$$f_Y(y) = \frac{1}{\theta} e^{-y/\theta} I_{(0,\infty)}(y) \tag{A.12}$$

La distribución exponencial es un caso particular de la distribución Gamma cuando el parámetro de forma k toma el valor 1, y usualmente se utiliza para describir la vida útil de un componente eléctrico o el tiempo necesario para la ocurrencia de algún evento.

Resultado A.16. *Si Y es una variable aleatoria con distribución exponencial con parámetro θ, entonces*

- $E(Y) = \theta$.
- $Var(Y) = \theta^2$.
- $m_Y(t) = \frac{1}{1-\theta t}$ *para $t < 1/\theta$, y no existe para otros valores de t.*

Resultado A.17. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes e idénticamente distribuidas con distribución exponencial con parámetro de escala θ, entonces la variable $\sum_{i=1}^{n} X_i$ tiene distribución Gamma con parámetro de forma n y parámetro de escala θ.*

A.2.7. Distribución Beta

Definición A.13. Una variable aleatoria Y tiene distribución Beta con parámetro de forma $\alpha > 0$ y parámetro de escala $\beta > 0$ si su función de densidad está dada por:

$$f_Y(y) = \frac{1}{Beta(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} I_{[0,1]}(y). \tag{A.13}$$

donde $Beta(\alpha, \beta) = \dfrac{\gamma(\alpha)\gamma(\beta)}{\gamma(\alpha+\beta)}$.

La distribución Beta tiene dos parámetros: α y β; en este caso, el vector de parámetros es $\boldsymbol{\theta} = (\alpha, \beta)'$ donde el espacio paramétrico está dado por $\Theta = (0, \infty) \times (0, \infty)$. Pero cuando uno de los dos parámetros es fijo, por ejemplo θ, entonces la distribución tendría un sólo parámetro: k.

Resultado A.18. *Si Y es una variable aleatoria con distribución Gamma con parámetro de forma k y parámetro de escala θ, entonces*

- $E(Y) = \frac{\alpha}{\alpha+\beta}$.
- $Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

A.2.8. Distribución normal

La distribución normal también es llamada la distribución gaussiana, rindiendo homenaje al matemático alemán Carl Friedrich Gauss (1777-1855). La distribución normal es, sin duda, una de las distribuciones más importantes, puesto que una gran parte de la teoría estadística fue desarrollada inicialmente para variables con esta distribución; por el otro lado, gracias al teorema central del límite, muchas distribuciones ajenas a la normal puede ser aproximadas por esta.

Definición A.14. Una variable aleatoria Y tiene distribución normal con parámetros μ y σ^2 si su función de densidad está dada por:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} I_{\mathbb{R}}(y), \tag{A.14}$$

donde $\sigma > 0$ y se nota como $Y \sim N(\mu, \sigma^2)$.

La distribución normal tiene dos parámetros, representado como $\boldsymbol{\theta} = (\mu, \sigma^2)$, mientras que su espacio paramétrico es $\Theta = \mathbb{R} \times (0, \infty)$.

Resultado A.19. *Si Y es una variable aleatoria con distribución normal con parámetros μ y σ^2, entonces*

- $E(Y) = \mu$.
- $Var(Y) = \sigma^2$.
- $m_Y(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$.

Cuando $\mu = 0$ y $\sigma = 1$, se dice que Y tiene distribución normal estándar y usualmente se denota por Z.
Resultado A.20. *Si $Y \sim N(\mu, \sigma^2)$, y α, β son constantes, entonces la variable $\alpha Y + \beta$ tiene distribución $N(\alpha\mu + \beta, \alpha^2\sigma^2)$.*

Demostración. Se usará el hecho de que la función generadora de momentos caracteriza la distribución probabilística. Se tiene que:

$$\begin{aligned} m_{\alpha Y+\beta}(t) &= E(e^{t(\alpha Y+\beta)}) \\ &= E(e^{\alpha t Y})e^{\beta t} \\ &= m_Y(\alpha t)e^{\beta t} \\ &= e^{\mu\alpha t+\sigma^2\alpha^2 t/2}e^{\beta t} \\ &= e^{(\alpha\mu+\beta)t+\sigma^2\alpha^2 t/2} \end{aligned}$$

la cual es la función generadora de momentos de una distribución $N(\alpha\mu+\beta, \alpha^2\sigma^2)$, y el resultado queda demostrado. □

Como consecuencia inmediata del anterior resultado, se define la estandarización, que es fundamental en la teoría relacionada con las distribuciones normales. Si $Y \sim N(\mu, \sigma^2)$, entonces la variable $Z = \frac{Y-\mu}{\sigma}$ tiene distribución normal estándar, y la anterior transformación se conoce como la normal estandarizada.
Resultado A.21. *Sea $Y_1, \cdots, Y_n$ variables aleatorias independientes, donde $Y_i \sim N(\mu_i, \sigma_i^2)$ con $i = 1, \cdots, n$, entonces la variable $\sum_{i=1}^n Y_i$ tiene distribución $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.*

A.2.9. Distribución log-normal

Definición A.15. Una variable aleatoria Y tiene distribución log-normal si su función de densidad está dada por:

$$p(Y \mid \mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\{\frac{-1}{2\sigma^2(\ln(y) - \mu)^2}\} \tag{A.15}$$

Nótese que si μ y σ^2 son la media y la varianza de $\ln(Y)$, entonces $\ln(Y)$ tiene distribución normal de media μ y varianza σ^2.
Resultado A.22. *Si Y es una variable aleatoria con distribución log-normal, entonces*

- $E(Y) = \exp(\mu + \sigma^2/2)$.
- $Var(Y) = (\exp(\sigma^2 - 1)) \exp(2\mu + \sigma^2)$.

A.2.10. Distribución Ji-cuadrado

Definición A.16. Una variable aleatoria Y tiene distribución Ji-cuadrado con n grados de libertad, con n entero positivo, si su función de densidad está dada por:

$$f_Y(y) = \frac{y^{(n/2)-1}e^{-y/2}}{2^{n/2}\Gamma(n/2)} I_{(0,\infty)}(y), \tag{A.16}$$

y se nota como $Y \sim \chi^2_n$.

La distribución Ji-cuadrado con n grados de libertad es un caso particular de la distribución Gamma cuando el parámetro de forma k toma el valor $n/2$ y el parámetro de escala toma el valor 2. También, en la literatura estadística existe la siguiente definición para la distribución Ji-cuadrado.

Definición A.17. Si $Z_1, \cdots, Z_n$ son variables aleatorias independientes e idénticamente distribuidas con distribución normal estándar, entonces la variable $\sum_{i=1}^n Z_i^2$ tiene distribución Ji-cuadrado con n grados de libertad.

Resultado A.23. *Si Y es una variable aleatoria con distribución Ji-cuadro con n grados de libertad, entonces*

- $E(Y) = n$.
- $Var(Y) = 2n$.
- $m_Y(t) = \left(\frac{1}{1-2t}\right)^{n/2}$ *para $t < 1/2$, y no existe para otros valores de t.*

Resultado A.24. *Sea $Z_1, \cdots, Z_m$ variables aleatorias independientes con distribución $\chi^2_{n_i}$ para $i = 1, \cdots, m$, entonces la variable $\sum_{i=1}^m Z_i$ tiene distribución Ji-cuadrado con $\sum_{i=1}^m n_i$ grados de libertad.*

A.2.11. Distribución t-student

El descubrimiento de la distribución t-student fue publicado por el estadístico inglés William Sealy Gosset (1876-1937) en el año 1908 cuando trabajaba en la famosa empresa cervecera *Guinness*. La publicación lo hizo de forma anónimo bajo el nombre de Student, pues Guinness le prohibía la publicación por ser el descubrimiento parte de resultados de investigación realizado por la empresa.

Definición A.18. Una variable aleatoria Y tiene distribución t-student con n grados de libertad si su función de densidad está dada por:

$$f_Y(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2} I_{\mathbb{R}}(y), \tag{A.17}$$

donde $n > 0$ y se nota como $Y \sim t_n$.

Otra definición que se encuentra frecuentemente en la literatura estadística es la siguiente.

Definición A.19. Sea Z una variable aleatoria con distribución normal estándar y Y una variable aleatoria con distribución Ji-cuadrado con n grados de libertad, si Z y Y son independientes, entonces la variable $\frac{Z}{\sqrt{Y/n}}$ tiene distribución t-student con n grados de libertad.

La función de densidad de la distribución t-student es muy parecida a la de distribución normal estándar, entre más grande sea el grado de libertad, más se parece a la distribución normal estándar.

Resultado A.25. *Si Y es una variable aleatoria con distribución t-student con n grados de libertad, entonces*

- $E(Y) = 0$ *para* $n > 1$.
- $Var(Y) = \frac{n}{n-2}$ *para* $n > 2$.

La distribución t-student no tiene función generadora de momentos.

A.2.12. Distribución t-student generalizada

Definición A.20. Una variable aleatoria Y tiene distribución t-student con n grados de libertad, parámetro de centralidad θ y parámetro de escala σ^2, si su función de densidad está dada por:

$$f_Y(y) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}\sigma}\left[1 + \frac{1}{n}\left(\frac{y-\theta}{\sigma}\right)^2\right]^{-(n+1)/2} I_{\mathbb{R}}(y), \tag{A.18}$$

donde $n > 0$ y se nota como $Y \sim t_n(\theta, \sigma^2)$.

Resultado A.26. *Si Y es una variable aleatoria con distribución t-student generalizada, entonces*

- $E(Y) = \theta$ *para* $n > 1$.
- $Var(Y) = \frac{n}{n-2}\sigma^2$ *para* $n > 2$.

A.2.13. Distribución F

La distribución F también se conoce como la distribución F de Fisher o distribución de Fisher-Snedecor, refiriendo al gran estadístico Ronald Aylmer Fisher (1890-1962) y el fundador del primer departamento de estadística en los Estados Unidos, George Waddel Snedecor (1881-1974).

Definición A.21. Una variable aleatoria Y tiene distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador si su función de densidad está dada por:

$$f_Y(y) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{m/2}\frac{z^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}z\right)^{\frac{m+n}{2}}}, \tag{A.19}$$

y se nota como $Y \sim F_n^m$.

Otra definición para la distribución F es como sigue:
Definición A.22. Sea Y y Y variables aleatorias independientes con distribuciones Ji-cuadrado con m y n grados de libertad, respectivamente, entonces la variable $\frac{Y/m}{Y/n}$ tiene distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador.
Resultado A.27. *Si Y es una variable aleatoria con distribución F con m grados de libertad en el numerador y n grados de libertado en el denominador, entonces*

- $E(Y) = \frac{n}{n-2}$ *para* $n > 2$.
- $Var(Y) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ *para* $n > 4$.

La distribución F no tiene función generadora de momentos.

A.3. Distribuciones multivariadas

A.3.1. Distribución Multinomial

Definición A.23. Un vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_p')$ tiene distribución multinomial si su función de densidad está dada por:

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \binom{n}{y_1, \ldots, y_p} \theta_1^{y_1} \cdots \theta_p^{y_p} \qquad \theta_i > 0\,, \ \sum_{i=1}^{p} \theta_i = 1 \ \text{ y } \ \sum_{i=1}^{p} y_i = p \quad \text{(A.20)}$$

donde

$$\binom{p}{y_1, \ldots, y_p} = \frac{p!}{y_1! \cdots y_p!}. \quad \text{(A.21)}$$

Como Gelman et al. (2003), afirma esta distribución es una generalización de la distribución binomial. La distribución marginal de una sola variable Y_i es $Binomial(p, \theta_i)$
Resultado A.28. *Si $\mathbf{Y}$ es una vector aleatorio con distribución multinomial, entonces*

- $E(\mathbf{Y}) = p(\theta_1, \ldots, \theta_p)'$.
- $Var(\mathbf{Y})_{ij} = \begin{cases} p\theta_i(1-\theta_i) & si\ i = j \\ -p\theta_i\theta_j & si\ i \neq j \end{cases}$

A.3.2. Distribución Dirichelt

Definición A.24. Un vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_p')$ tiene distribución Dirichelt si su función de densidad está dada por:

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \frac{\Gamma(\theta_1 + \cdots + \theta_p)}{\Gamma(\theta_1) \cdots \Gamma(\theta_p)} y^{\theta_1 - 1} \cdots y^{\theta_p - 1} \qquad \theta_i > 0 \text{ y } \sum_{i=1}^{p} \theta_i = 1. \tag{A.22}$$

Esta distribución es una generalización de la distribución beta. La distribución marginal de una sola variable Y_i es $Beta(\theta_i, (\sum_{i=1}^{p} \theta_i) - \theta_i)$

Resultado A.29. *Si* $\mathbf{Y}$ *es una vector aleatorio con distribución Dirichlet, entonces*

- $E(\mathbf{Y}) = (\sum_{i=1}^{p} \theta_i)^{-1}(\theta_1, \ldots, \theta_p)'$.
- $Var(\mathbf{Y})_{ij} = \begin{cases} \frac{\theta_i(\sum_{i=1}^{p} \theta_i - \theta_i)}{(\sum_{i=1}^{p} \theta_i)^2(\sum_{i=1}^{p} \theta_i + 1)} & si\ i = j \\ -\frac{\theta_i \theta_j)}{(\sum_{i=1}^{p} \theta_i)^2(\sum_{i=1}^{p} \theta_i + 1)} & si\ i \neq j \end{cases}$

A.3.3. Distribución Normal Multivariante

Definición A.25. Un vector aleatorio $\mathbf{Y} = (Y_1, \ldots, Y_p')$ tiene distribución normal multivariante de orden p, denotada como $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, si su función de densidad está dada por:

$$p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\theta})' \boldsymbol{\Sigma} (\mathbf{y} - \boldsymbol{\theta})\right\} \tag{A.23}$$

donde $\mid \boldsymbol{\Sigma} \mid$ se refiere al determinante de la matriz $\boldsymbol{\Sigma}$, la cual es simétrica y definida positiva de orden $p \times p$.

La distribución Normal Multivariante es el baluarte de una gran cantidad de técnicas y métodos estadísticos como son los modelos lineales, los modelos lineales generalizados, el análisis factorial, etc. Algunas de sus propiedades se citan a continuación.

Resultado A.30. *Si* $\mathbf{Y} = (Y_1, \ldots, Y_p')$ *es una vector aleatorio con distribución normal multivariante, entonces*

- *La distribución marginal de cualquier subconjunto de componentes de* $\mathbf{Y}$ *es también normal multivariante. Por ejemplo si* $\mathbf{Y}$ *es particionado en* $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')$*, entonces* $p(\mathbf{Y}_1)$ *seguiría una distribución normal multivariante, al igual que* $p(\mathbf{Y}_2)$*.*
- *Cualquier transformación lineal de* $\mathbf{Y}$ *es normal multivariante y su dimensión equivale al rango de la transformación. en particular, la suma de las*

componentes del vector, dada por $\sum_{i=1}^{p} Y_i$ *sigue una distribución normal univariada.*

- *La distribución condicional de* $\mathbf{Y}$*, restringida a un subespacio lineal es normal.*
- *La distribución condicional de cualquier sub-vector de elementos de* $\mathbf{Y}$ *dados los restantes elementos es normal multivariante. Más aún, si* $\mathbf{Y}$ *es particionado en* $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')$*, entonces* $p(\mathbf{Y}_1 \mid \mathbf{Y}_2)$ *es normal multivariada con*

$$E(\mathbf{Y}_1 \mid \mathbf{Y}_2) = E(\mathbf{Y}_1) + Cov(\mathbf{Y}_1, \mathbf{Y}_2)(Var(\mathbf{Y}_2))^{-1}(\mathbf{Y}_2 - E(\mathbf{Y}_2))$$
$$Var(\mathbf{Y}_1 \mid \mathbf{Y}_2) = Var(\mathbf{Y}_1) - Cov(\mathbf{Y}_1, \mathbf{Y}_2)(Var(\mathbf{Y}_2))^{-1}Cov(\mathbf{Y}_2, \mathbf{Y}_1)$$

- *Si* $\mathbf{X}$ *es un vector con distribución normal multivariante, entonces* $\mathbf{X} + \mathbf{Y}$ *tiene una distribución normal multivariante. En particular si* $\mathbf{X}$ *es independiente de* $\mathbf{Y}$*, comparten el mismo orden* p *y* $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Gamma})$*, entonces* $\mathbf{X} + \mathbf{Y} \sim N_p(\boldsymbol{\mu} + \boldsymbol{\theta}, \boldsymbol{\Gamma} + \boldsymbol{\Sigma})$*.*

Resultado A.31. *Si* $\mathbf{Y}$ *es una vector aleatorio con distribución Normal Multivariante, entonces*

- $E(\mathbf{Y}) = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$.
- $Var(\mathbf{Y}) = \boldsymbol{\Sigma}$

Resultado A.32. *Dado* $\mathbf{Y}$ *un vector aleatorio particionado como* $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')$ *con esperanza* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ *y matrix de varianzas y covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Si $\mathbf{Y}_1 \mid \mathbf{Y}_2 \sim N(\boldsymbol{\theta}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\theta}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$ *y* $\mathbf{Y}_2 \sim N(\boldsymbol{\theta}_2, \boldsymbol{\Sigma}_{22})$*, entonces se tiene que*

$$\mathbf{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}).$$

Resultado A.33. *Si* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ *es una muestra aleatoria de vectores con distribución Normal Multivariante, entonces la verosimilitud de la muestra se puede escribir como*

$$\prod_{i=1}^{n} p(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2} traza(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}})\right\} \tag{A.24}$$

Donde $\mathbf{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})'$.

Demostración. La verosimilitud de la muestra aleatoria está dada por

$$\prod_{i=1}^{n} p(\mathbf{Y}_i \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right\}$$
$$= \mid \boldsymbol{\Sigma} \mid^{-n/2} \exp\left\{-\frac{1}{2} traza\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}}\right)\right\}$$

Puesto que, por las propiedades del operador *traza*, se tiene que

- Si c es un escalar, entonces $c = traza(c)$.
- Si $\mathbf{A}$ y $\mathbf{B}$ son dos matrices, entonces $traza(\mathbf{AB}) = traza(\mathbf{BA})$
- Si $\mathbf{A}_i$ (i=1,. . . ,n) son matrices del mismo tamaño, entonces $\sum_{i=1}^{n} traza(\mathbf{A}_i) = traza\left(\sum_{i=1}^{n} \mathbf{A}_i\right)$

Por lo anterior,

$$\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta}) = traza\left[\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})\right]$$
$$= \sum_{i=1}^{n} traza[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})')]$$
$$= traza\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\theta})(\mathbf{Y}_i - \boldsymbol{\theta})')\right]$$
$$= traza(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\boldsymbol{\theta}})$$

□

A.3.4. Distribución Wishart

Definición A.26. Sea $\boldsymbol{\Sigma}$ una matriz aleatoria simétrica y definida positiva de tamaño $p \times p$. Se dice que $\boldsymbol{\Sigma}$ tiene distribución Wishart con v grados de libertad, denotada como $\mathbf{Y} \sim Wishart_v(\boldsymbol{\Lambda})$, si su función de densidad está dada por:

$$p(\boldsymbol{\Sigma}) = \left(2^{vp/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{v+1-i}{2}\right)\right)^{-1}$$
$$\times \mid \boldsymbol{\Lambda} \mid^{-v/2} \mid \boldsymbol{\Sigma} \mid^{(v-p-1)/2} \exp\left\{-\frac{1}{2} traza(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma})\right\} \quad \text{(A.25)}$$

donde $\mid \boldsymbol{\Lambda} \mid$ se refiere al determinante de la matriz $\boldsymbol{\Lambda}$, la cual es simétrica y definida positiva de orden $p \times p$.

Resultado A.34. *Si $\boldsymbol{\Sigma}$ es una matriz aleatoria con distribución Wishart con v grados de libertad, entonces $E(\boldsymbol{\Sigma}) = v\boldsymbol{\Lambda}$*

A.3.5. Distribución inversa-Wishart

Definición A.27. Sea $\mathbf{\Sigma}$ una matriz aleatoria simétrica y definida positiva de tamaño $p \times p$. Se dice que $\mathbf{\Sigma}$ tiene distribución Wishart con v grados de libertad, denotada como $\mathbf{Y} \sim Wishart_v(\mathbf{\Lambda})$, si su función de densidad está dada por:

$$p(\mathbf{\Sigma}) = \left(2^{vp/2} \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\frac{v+1-i}{2} \right) \right)^{-1} \times \mid \mathbf{\Lambda} \mid^{v/2} \mid \mathbf{\Sigma} \mid^{-(v+p+1)/2} \exp\left\{ -\frac{1}{2} traza(\mathbf{\Lambda}\mathbf{\Sigma}^{-1}) \right\} \quad \text{(A.26)}$$

donde $\mid \mathbf{\Lambda} \mid$ se refiere al determinante de la matriz $\mathbf{\Lambda}$, la cual es simétrica y definida positiva de orden $p \times p$.

Resultado A.35. *Si $\mathbf{\Sigma}$ es una matriz aleatoria con distribución inversa-Wishart con v grados de libertad, entonces* $E(\mathbf{\Sigma}) = \dfrac{1}{v-p-1}\mathbf{\Lambda}$

Resultado A.36. *Si $\mathbf{\Sigma}^{-1}$ es una matriz aleatoria con distribución inversa-Wishart, entonces con $\mathbf{\Sigma}$ tiene distribución Wishart.*

Apéndice B

Matriz de información

Definición B.1. Dada X una variable aleatoria con función de densidad $f(x,\theta)$, donde θ es el parámetro de la distribución, y además existe $\dfrac{\partial}{\partial\theta}\ln f(x,\theta)$, entonces se define la información contenida en X acerca de θ como

$$I_X(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X,\theta)\right]^2\right\}. \tag{B.1}$$

Resultado B.1. *En la anterior definición, si además existe* $\dfrac{\partial^2}{\partial\theta^2}\ln f(x,\theta)$, *entonces se tiene que*

$$I_X(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right\}. \tag{B.2}$$

Las anteriores definiciones introducen la información contenida en una variable; sin embargo, cuando tenemos disponible una muestra aleatoria, es necesario definir la información contenida en una muestra aleatoria acerca de algún parámetro.

Definición B.2. Dada $X_1, \cdots, X_n$ variables aleatorias con función de densidad $f(x_i,\theta)$, donde θ es el parámetro de la distribución, y además existe $\dfrac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(x_i,\theta)$, entonces se define la información contenida en la muestra aleatoria acerca de θ como

$$I_{X_1,\cdots,X_n}(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i,\theta)\right]^2\right\}. \tag{B.3}$$

Resultado B.2. *Dada $X_1, \cdots, X_n$ una muestra aleatoria, entonces*

$$I_{X_1,\cdots,X_n}(\theta) = nI_X(\theta),$$

donde $I_X(\theta) = I_{X_i}(\theta)$, con $i = 1, \cdots, n$. Es decir, en una muestra aleatoria, cada variable aporta la misma cantidad de información, y la cantidad total de información en la muestra es la suma de la información en cada variable.

Demostración.

$$\begin{aligned}
I_{X_1,\cdots,X_n}(\theta) &= E\left\{\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(X_i,\theta)\right]^2\right\} \\
&= E\left\{\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} \\
&= E\left\{\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} + \\
&\qquad \underbrace{E\left\{\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\frac{\partial}{\partial\theta}\ln f(X_j,\theta)\right]\right\}}_{=0,\text{ por la independencia entre } X_i \text{ y } X_j} \\
&= \sum_{i=1}^{n} E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right]^2\right\} \\
&= \sum_{i=1}^{n} I_X(\theta) = nI_X(\theta).
\end{aligned}$$

□

Ejemplo B.1. Sea $X_1, \cdots, X_n$ una muestra aleatoria proveniente de la distribución $N(\mu, \sigma^2)$, la información contenida en la muestra acerca de μ es n/σ^2. Para verificar esta afirmación, calculamos la información acerca de μ en una variable X con distribución $N(\mu, \sigma^2)$. Tenemos:

$$
\begin{aligned}
I_X(\mu) &= -E\left\{\frac{\partial^2}{\partial\mu^2}\ln f(X,\theta)\right\} \\
&= -E\left\{\frac{\partial^2}{\partial\mu^2}\left[-\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2\right]\right\} \\
&= -E\left\{\frac{\partial}{\partial\mu}\left[\frac{X-\mu}{\sigma^2}\right]\right\} \\
&= -E\left\{-\frac{1}{\sigma^2}\right\} \\
&= \frac{1}{\sigma^2}.
\end{aligned}
$$

Ahora, usando el Resultado 2.3.4, se tiene que $I_{X_1,\cdots,X_n}(\mu) = n/\sigma^2$.

Nótese que esta información, en primer lugar, depende del tamaño n de manera que entre más grande sea la muestra, hay mayor información acerca de μ; en segundo lugar, entre más pequeña sea la varianza σ^2, la cantidad de información acerca de μ también incrementa, esto es natural, puesto que si σ^2 es pequeña, los datos de la muestra están muy concentrados alrededor de μ, entonces estos datos aportan más información que otros datos con más dispersión.

Definición B.3. Dada una variable aleatoria X con función de densidad $f(x,\boldsymbol{\theta})$, la matriz de información contenida en X acerca de $\boldsymbol{\theta}$ se define como

$$
I_X(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\left(\frac{\partial \ln f(X,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'\right\} \tag{B.4}
$$

Definición B.4. Dada una muestra aleatoria $X_1, \cdots, X_n$ con función de densidad $f(x_i,\boldsymbol{\theta})$, la matriz de información contenida en la muestra acerca de $\boldsymbol{\theta}$ se define como

$$
I_{X_1,\cdots,X_n}(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln \prod_{i=1}^n f(X_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\left(\frac{\partial \ln \prod_{i=1}^n f(X_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'\right\}
$$

Ejemplo B.2. Dada una muestra aleatoria $X_1, \cdots, X_n$ con distribución común $N(\mu,\sigma^2)$, vamos a hallar la matriz de información contenida en la muestra acerca del vector de parámetros (μ,σ^2). Tenemos que

$$
\begin{aligned}
& I_{X_1,\cdots,X_n}(\mu,\sigma^2) \\
&= E\left\{\begin{pmatrix} \dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \mu} \\ \dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \sigma^2} \end{pmatrix} \begin{pmatrix} \dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \mu} & \dfrac{\partial \ln \prod_{i=1}^n f(X_i,\mu,\sigma^2)}{\partial \sigma^2} \end{pmatrix}\right\} \\
&= E\left\{\begin{pmatrix} \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma^2} \\ \dfrac{\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2}{2\sigma^4} \end{pmatrix} \begin{pmatrix} \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma^2} & \dfrac{\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2}{2\sigma^4} \end{pmatrix}\right\} \\
&= E\left\{\begin{pmatrix} \dfrac{(\sum_{i=1}^n X_i - n\mu)^2}{\sigma^4} & \dfrac{(\sum_{i=1}^n X_i - n\mu)(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)}{2\sigma^6} \\ \dfrac{(\sum_{i=1}^n X_i - n\mu)(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)}{2\sigma^6} & \dfrac{(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)^2}{4\sigma^8} \end{pmatrix}\right\}
\end{aligned}
$$

Donde el primer elemento diagonal de la anterior matriz está dada por

$$
\begin{aligned}
E\left\{\frac{(\sum_{i=1}^n X_i - n\mu)^2}{\sigma^4}\right\} &= \left[Var\left(\sum_{i=1}^n X_i - n\mu\right) + (E\left(\sum_{i=1}^n X_i - n\mu\right))^2\right]/\sigma^4 \\
&= n\sigma^2/\sigma^4 = n/\sigma^2.
\end{aligned}
$$

El segundo elemento diagonal está dada por

$$
\begin{aligned}
& E\left\{\frac{(\sum_{i=1}^n (X_i-\mu)^2 - n\sigma^2)^2}{4\sigma^8}\right\} && \text{(B.5)} \\
&= \frac{1}{4\sigma^8} E\left\{\left[\sum_{i=1}^n (X_i-\mu)^2\right]^2 + n^2\sigma^4 - 2n\sigma^2 \sum_{i=1}^n (X_i-\mu)^2\right\} && \text{(B.6)} \\
&= \frac{1}{4\sigma^8}\left\{Var(\sum_{i=1}^n (X_i-\mu)^2) + \left[E(\sum_{i=1}^n (X_i-\mu)^2)\right]^2 + n^2\sigma^4 - 2n\sigma^2 E\left[\sum_{i=1}^n (X_i-\mu)^2\right]\right\} && \text{(B.7)}
\end{aligned}
$$

Usando el hecho de que

$$
\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} \sim \chi^2_n
$$

y la esperanza y varianza de la distribución χ^2_n, tenemos que la expresión (B.5) está dada por

$$
\frac{1}{4\sigma^8}\left\{2n\sigma^4 + \left[n\sigma^2\right]^2 + n^2\sigma^4 - 2n\sigma^2 n\sigma^2\right\} = \frac{n}{2\sigma^4}.
$$

Finalmente, el elemento fuera de la diagonal de la matriz $I_{X_1,\cdots,X_n}(\mu,\sigma^2)$ está dado por

$$
\begin{aligned}
& E\left\{\left(\sum_{i=1}^{n} X_i - n\mu\right)\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2\right)\right\} \\
&= E\left\{\sum_{i=1}^{n} X_i\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2\right) - n\mu\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2\right)\right\} \\
&= E\left\{\sum_{i=1}^{n} X_i \sum_{i=1}^{n}(X_i-\mu)^2\right\} - n\sigma^2 E\left(\sum_{i=1}^{n} X_i\right) - n\mu E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) + n^2\mu\sigma^2 \\
&= E\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i^2\right) - 2\mu E\left[(\sum_{i=1}^{n} X_i)^2\right] + n^2\mu^3 - n^2\mu\sigma^2 - n^2\mu\sigma^2 + n^2\mu\sigma^2 \\
&= E\left(\sum_{i=1}^{n} X_i^3 + \sum_{i\neq j} X_i X_j^2\right) - 2\mu(n\sigma^2 + n^2\mu^2) + n^2\mu^3 - n^2\mu\sigma^2 \\
&= \sum_{i=1}^{n}\left[3\mu E(X_i^2) - 2\mu^3\right] + \sum_{i\neq j} E(X_i)E(X_j^2) - 2n\mu\sigma^2 - 2n^2\mu^3 + n^2\mu^3 - n^2\mu\sigma^2 \\
&= 3n\mu(\sigma^2+\mu^2) - 2n\mu^3 + \mu(\sigma^2+\mu^2)(n^2-n) - 2n\mu\sigma^2 - 2n^2\mu^3 + n^2\mu^3 - n^2\mu\sigma^2 \\
&= 0
\end{aligned}
$$

De donde obtenemos finalmente la matriz de información $I_{X_1,\cdots,X_n}(\mu,\sigma^2)$ dada por

$$
I_{X_1,\cdots,X_n}(\mu,\sigma^2) = \begin{pmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{pmatrix}
$$

Apéndice C

Elementos de simulación estadística

Como lo afirma Gelman et al. (2003) la simulación numérica es parte central del análisis bayesiano puesto que la generación de datos provenientes de una distribución de probabilidad se puede realizar fácilmente, incluso cuando la forma estructural de ésta no es conocida o es muy complicada computacionalmente. A lo largo de la historia del desarrollo de la teoría estadística, la simulación de distribuciones de probabilidad ha jugado un papel importante. Aunque son innumerables los métodos de generación de datos, en este apartado, se da cuenta de unos pocos, quizás lo más usados en este auge computacional.

`R` es un software de uso libre que maneja un ambiente de programación enfocado al manejo de matrices y por lo tanto muy apropiado para realizar la simulación de las distribuciones posteriores y predictivas necesarias para la inferencia bayesiana. En este capítulo, usando una serie de ejemplos, se describe cómo `R` puede ser usado como una herramienta efectiva. Estos ejemplos tienen un énfasis especial en tópicos bayesianos, específicamente en el uso de las cadenas de Markov para simular distribuciones posteriores.

C.1. Métodos directos

C.1.1. Método de la transformación uniforme

Al momento de la simulación estocástica de observaciones provenientes de alguna distribución de interés, la distribución uniforme es quizás la más usada y la más importante. El siguiente resultado adaptado de Robert and Casella (1999) así lo confirma.

Resultado C.1. *Si U es una variable aleatoria con distribución uniforme en el intervalo $(0, 1)$, entonces la la variable aleatoria $F^{-1}(U)$ tiene distribución F.*

Aunque la función F no necesariamente es una función uno a uno (por lo menos no lo es en el caso discreto) sí se puede verificar que $F^{-1}(U)$ es única con probabilidad uno. Una definición general, que encaja en el caso continuo o discreto, de la función F inversa es la siguiente

Definición C.1. Para cualquier función F definida sobre $\mathbb{R}$, se define la función inversa generalizada de F como

$$F^{-1}(u) = \text{ínf}\{x \mid F(x) \geq u\} \tag{C.1}$$

Ejemplo C.1. Suponga que X es una variable aleatoria con distribución exponencial. De esta forma, su función de densidad acumulativa viene dada por

$$F(x) = 1 - \exp\{-\theta x\}$$

Del anterior resultado se tiene que si u es una realización de una variable $U \sim Uniforme(0, 1)$, entonces $F^{-1}(u)$ es una realización de una variable con distribución exponencial. Como $x = F^{-1}(u)$, entonces $F(x) = u$ y despejando x, se llega a que la siguiente expresión

$$F^{-1}(u) = -\frac{\ln(1-u)}{\theta}$$

entrega una forma diáfana para la simulación de una observación con distribución exponencial. Para simular una muestra de n observaciones, simplemente se repite el anterior procedimiento n veces. En `R`, el código necesario para la simulación de una muestra de tamaño 1000 proveniente de una distribución exponencial con parámetro $\theta = 5$ es

```
theta <- 5
u <- runif(1000)
rexpo <- log(1 - u)/(-theta)
1/mean(rexpo)
```

```
## [1] 5.350282
```

```
hist(rexpo, freq = FALSE)
lines(density(rexpo), col = 2)
```

C.1.2. Método de la grilla

Existen distribuciones de probabilidad cuya forma estructural es muy compleja. Mas aún, existen distribuciones de probabilidad conocidas para las cuales la

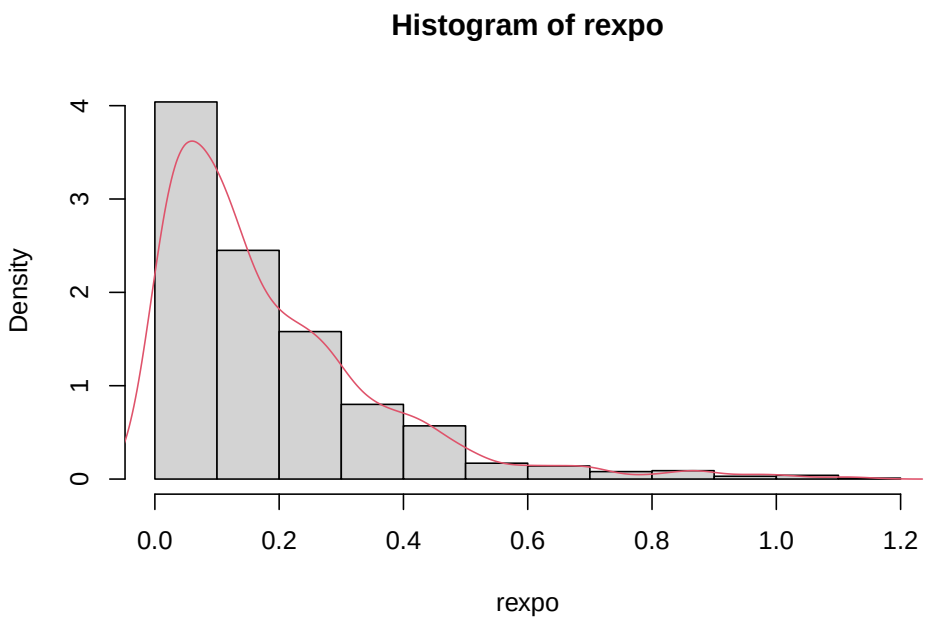


Figura C.1: Histograma de observaciones con distribución exponencial

inversa de la función de de densidad acumulativa es difícil de solucionar analíticamente. En los anteriores casos, el método analítico dado por el teorema de la transformación integral de probabilidad no siempre resulta efectivo. Sin embargo, es posible realizar una variante, manteniendo el espíritu de la anterior técnica.

El presente método utiliza una distribución discreta para aproximar cualquier tipo de distribución (discreta o continua) sin importar su nivel de complejidad. El algoritmo que enmarca este método se da a continuación:

1. Escribir la densidad de interés como $f(\cdot)$ y establecer el rango de la variable aleatoria de interés.
2. Fijar un conjunto de n valores $x_1 < \cdots < x_n$ equiespaciados que cubran una gran parte del rango de la variable aleatoria.
3. Para x_k $(k = 1, \ldots, n)$ calcular $f(x_k)$ que equivale al valor de la densidad en el punto x_k. Nótese que si $f(\cdot)$ es una función de densidad continua, entonces $f(x_k)$ no corresponde a una probabilidad;
4. Calcular la probabilidad asociada al punto x_k definida por la aproximación discreta a $f(\cdot)$ y dada por

$$p(x_k) = \frac{f(x_k)}{\sum_{k=1}^{n} f(x_k)}$$

5. Calcular la función de densidad acumulativa aproximada definida como

$$F(x) = \begin{cases} 0, & \text{si } x < x_1 \\ \sum_{l=1}^{k} p(x_l), & \text{si } x_k \leq x < x_{k+1} \\ 1, & \text{si } x > x_n \end{cases}$$

6. Simular una observación u proveniente de una distribución uniforme continua en el intervalo $(0, 1)$.
7. Si $F(x_k) < u \leq F(x_{k+1})$, entonces $F^{-1}(u) = x_{k+1}$ y por consiguiente el valor x_{k+1} es una pseudo-observación proveniente de la densidad de interés.

Nótese que en el anterior proceso, la unidad x_{k+1} es seleccionada con probabilidad p_{k+1}; puesto que

$$\begin{aligned} P(F(x_k) < U \leq F(x_{k+1})) &= F(x_{k+1}) - F(x_k) \\ &= \sum_{l=1}^{k+1} p(x_l) - \sum_{l=1}^{k} p(x_l) = p_{k+1} \end{aligned}$$

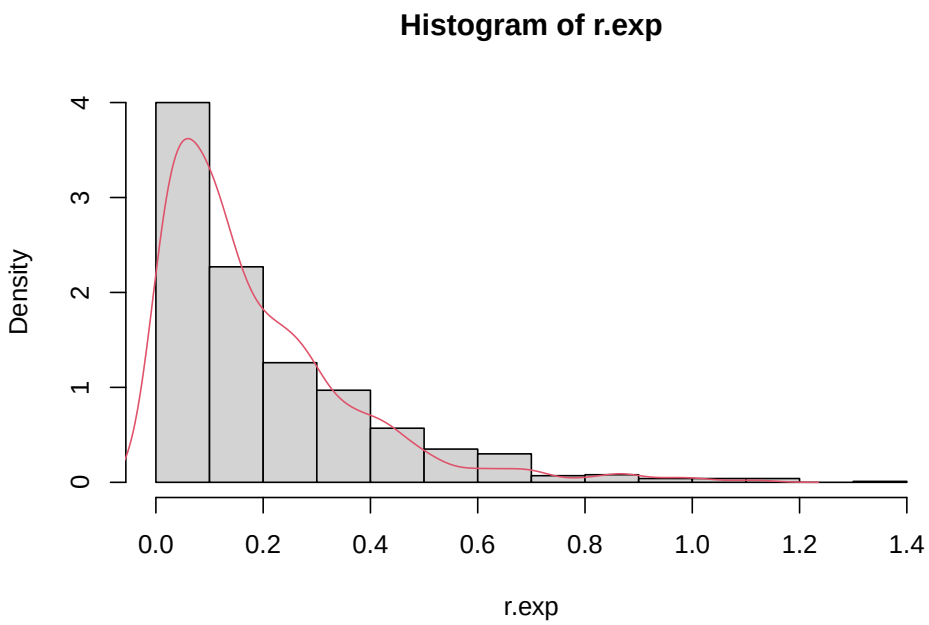
Si se quiere extraer una muestra aleatoria de N observaciones provenientes de la distribución de interés, entonces basta con repetir el anterior proceso N veces. Por supuesto, como se trata de una muestra aleatoria cada selección se debe realizar con repetición; de esta manera no importa si $N > n$. Suponiendo que el conjunto $x_1, \ldots, x_n$ conforma una grilla de puntos lo suficientemente cercanos y que no sucede nada importante entre cada uno de ellos, entonces esta técnica debe tener un buen funcionamiento.

Ejemplo C.2. El siguiente código computacional permite utilizar el método de la grilla para simular mil valores provenientes de una distribución exponencial con parámetro $\theta = 5$.

```
theta <- 5
x.grid <- seq(0, 100, by=0.01)
p.exp <- theta * exp(-theta * x.grid)
r.exp <- sample(x.grid, 1000, prob = p.exp, replace = T)
1/mean(r.exp)
```

```
## [1] 4.805151
```

```
hist(r.exp, freq = FALSE)
lines(density(rexpo), col = 2)
```



Ejemplo C.3. De la misma manera, el método de la grilla permite simular valores de un distribución discreta. El siguiente código computacional permite utilizar el método de la grilla para simular mil valores provenientes de una distribución Poisson con parámetro $\theta = 2$.

```
p.poisson <-function(theta, x.grid){
  N <- length(x.grid)
  res <- rep(NA, N)
  for(k in 1:N){
    P1 <- exp(-theta) * theta^(x.grid[k])
    P2 <- factorial(x.grid[k])
    res[k] <- P1/P2
  }
  return(res)
}

theta <- 2
x.grid <- seq(0, 100, by = 1)
f.x <- p.poisson(theta, x.grid)
p.x <- f.x/sum(f.x)
sum(p.x)
```

```
## [1] 1
```

```
rpois <- sample(x.grid, 1000, prob=p.x, replace = T)
mean(rpois)
```

```
## [1] 2.013
```

```
var(rpois)
```

```
## [1] 1.972804
```

```
hist(rpois, freq = FALSE)
```

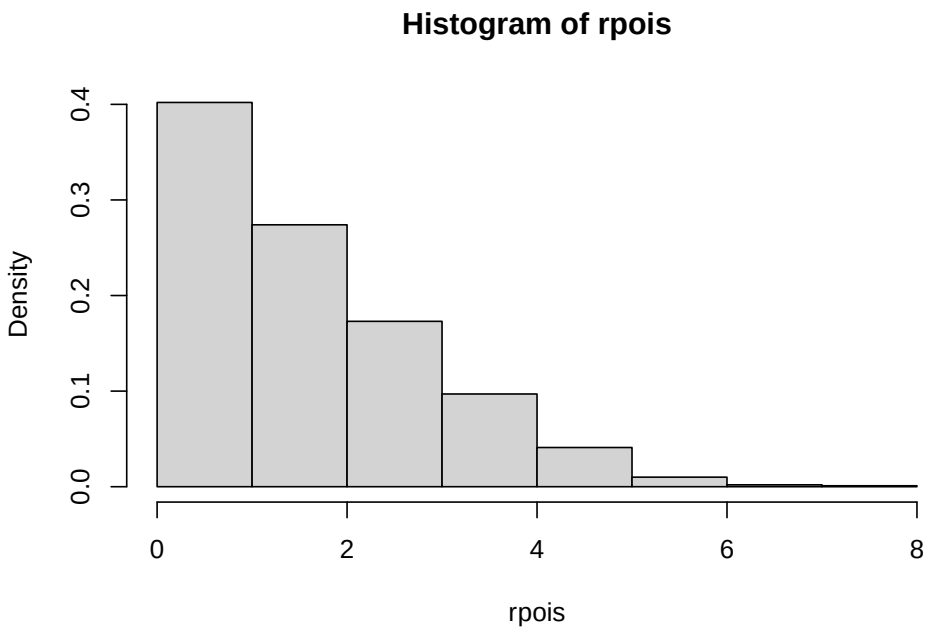


Figura C.2: Histograma de observaciones con distribución Poisson

Ejemplo C.4. El método de la grilla también puede utilizarse para simular observaciones de una distribución biparamétrica, univariada y continua. El siguiente código computacional permite utilizar el método de la grilla para simular mil valores provenientes de una distribución Gamma con parámetros $\theta = 4$, $\beta = 2$.

```
p.gamma <- function(a, b, x.grid){
  N <-length(x.grid)
  res <- rep(NA, N)
  for(k in 1:N){
    P1 <- (b^a)/gamma(a)
    P2 <- x.grid[k]^(a - 1)
    P3 <- exp(-b * x.grid[k])
    res[k] <- P1*P2*P3
  }
  return(res)
}
```

```r
alpha <- 4
beta <- 2
x.grid <- seq(0, 100, by = 0.1)
f.x <- p.gamma(alpha, beta, x.grid)
p.x <- f.x / sum(f.x)

rgamma <- sample(x.grid, 1000, prob = p.x, replace = T)
mean(rgamma)
```

```
## [1] 1.9975
```

```r
var(rgamma)
```

```
## [1] 1.031415
```

```r
hist(rgamma, freq = F)
lines(density(rgamma), col = 2)
```

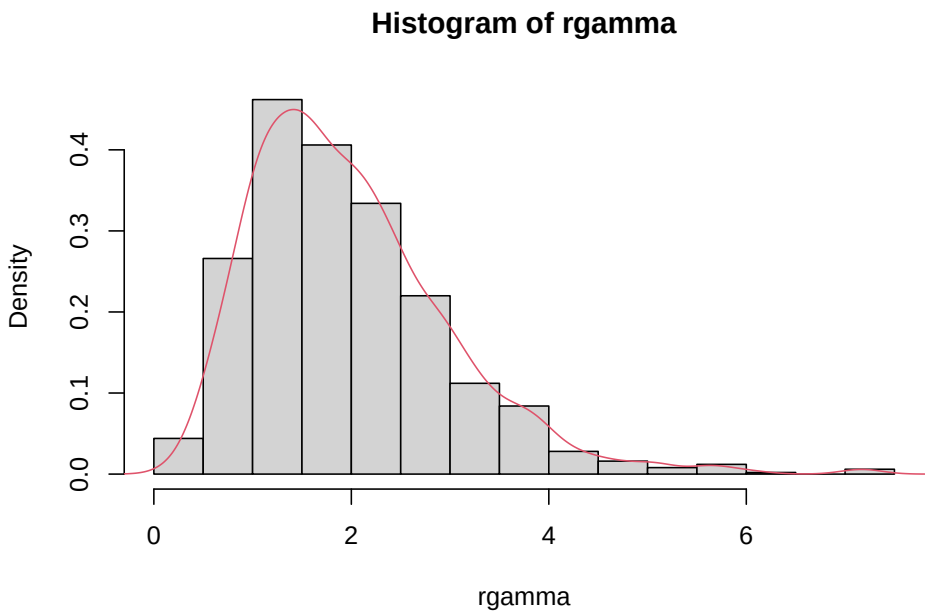


Figura C.3: Histograma de observaciones con distribución Gamma.

Ejemplo C.5. Para comprobar el poder de este método de simulación, se presenta el siguiente código que permite simular valores de una distribución multiparamétrica, bivariada y continua. En particular, se simulan valores de la distribución Normal multivariante con vector de medias $\boldsymbol{\mu} = (2, 4)'$ y matriz de varianzas covarianzas $\boldsymbol{\Sigma} = \begin{bmatrix} 25 & 30 \\ 30 & 16 \end{bmatrix}$

```
p.normal2 <- function(mu, Sigma, x, y){
  P1  <- 1/(2 * pi)
  P2  <- 1/sqrt(det(Sigma))
  P3a <- t((c(x, y) - mu)) %*% solve(Sigma) %*% (c(x, y) - mu)
  P3  <- exp((-1/2) * P3a)
  res <- P1 * P2 * P3
  return(res)
}

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)
}

mu1 <- c(2, 4)
Sigma1 <- matrix(c(25, 10, 10, 16), nrow=2)

x.grid <- seq(mu1[1] - 3 * sqrt(Sigma1[1, 1]),
              mu1[1] + 3 * sqrt(Sigma1[1, 1]),
              by = 0.5)
y.grid <- seq(mu1[2] - 3 * sqrt(Sigma1[2, 2]),
              mu1[2] + 3 * sqrt(Sigma1[2, 2]),
              by = 0.5)
xy.grid <- grilla(x.grid, y.grid)
N.grid <- dim(xy.grid)[1]

f.xy <- rep(NA, N.grid)
for(j in 1:N.grid){
  f.xy[j] <- p.normal2(mu1, Sigma1,
                       xy.grid[j, 1],
                       xy.grid[j, 2])
}

p.xy <- as.vector(f.xy/sum(f.xy))
sum(p.xy)
```

```
## [1] 1
```

```
rnormal2 <- sample(N.grid, 1000, prob = p.xy, replace = T)
rxy.normal2 <- xy.grid[rnormal2, ]
```

```
rx.normal <- rxy.normal2[, 1]
ry.normal <- rxy.normal2[, 2]

colMeans(rxy.normal2)
```

```
## [1] 1.678 3.976
```

```
var(rxy.normal2)
```

```
##           [,1]      [,2]
## [1,] 24.955271  9.292064
## [2,]  9.292064 14.929353
```

```
hist(rx.normal, freq = F)
lines(density(rx.normal), col = 2)
```

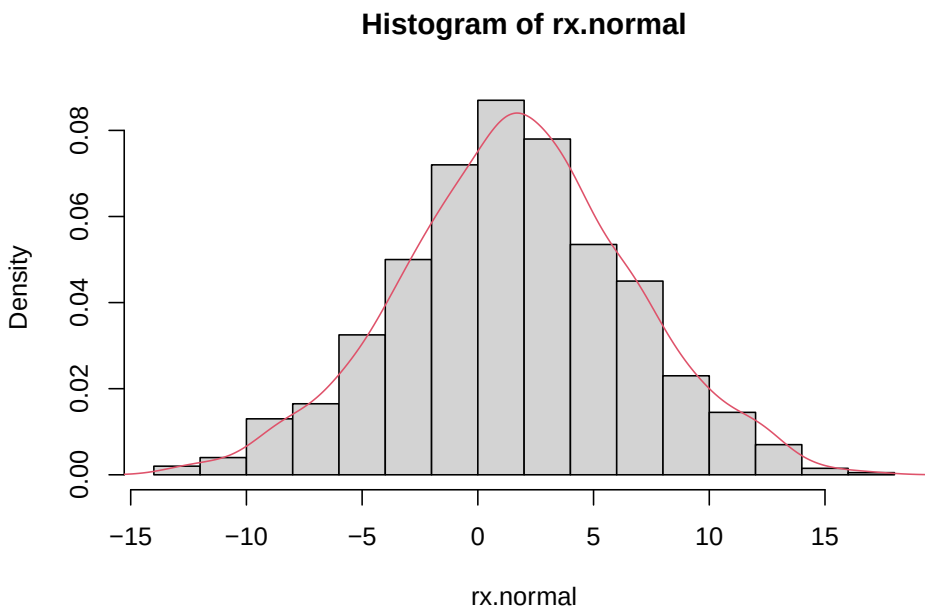


Figura C.4: Histogramas de observaciones con distribución Normal bivariada.

```
hist(ry.normal, freq = F)
lines(density(ry.normal), col = 2)
```

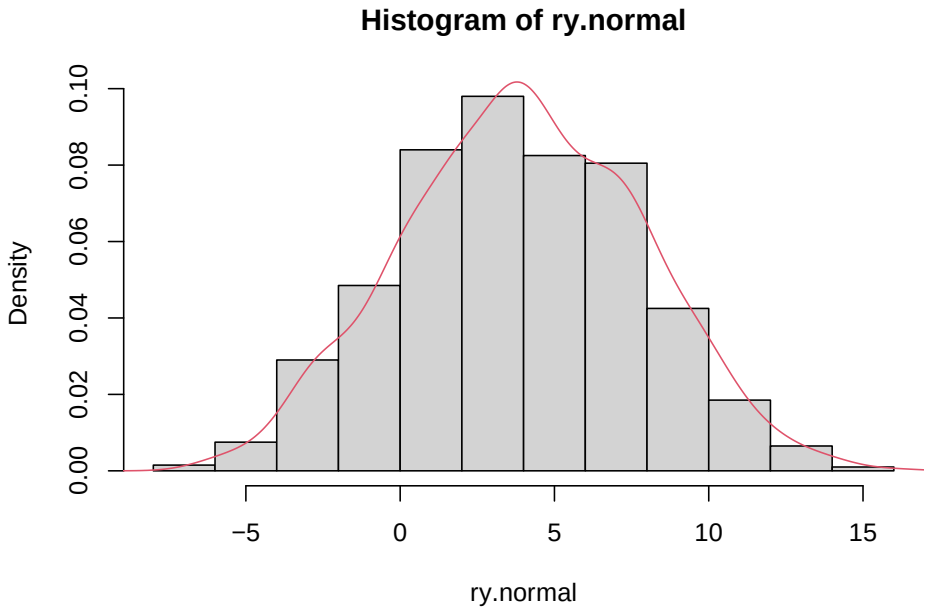


Figura C.5: Histogramas de observaciones con distribución Normal bivariada.

C.2. Métodos de Monte Carlo vía cadenas de Markov

C.2.1. El muestreador de Gibbs

Tal como lo afirma Peña (2002), este procedimiento es apropiado para obtener muestras de una distribución conjunta cuando es fácil muestrear de las distribuciones condicionadas. El algoritmo se implementa asumiendo que $\boldsymbol{\theta}_i = (\theta_i^{(1)}, ..., \theta_i^{(d)})$ representa a los valores actuales de $\boldsymbol{\theta}$. Entonces $\boldsymbol{\theta}_{i+1}$ se obtiene así:

- Generar $\theta_{i+1}^{(1)}$ de $p(\theta^{(1)} \mid \theta_i^{(2)}, \ldots, \theta_i^{(d)}, x)$
- Generar $\theta_{i+1}^{(2)}$ de $p(\theta^{(2)} \mid \theta_{i+1}^{(1)}, \theta_i^{(3)}, \ldots, \theta_i^{(d)}, x)$
- ...
- Generar $\theta_{i+1}^{(d)}$ de $p(\theta^{(d)} \mid \theta_{i+1}^{(1)}, \theta_{i+1}^{(2)}, \ldots, \theta_{i+1}^{(d-1)}, x)$

La idea de este esquema es renovar cada componente por medio de la simulación de la correspondiente distribución condicional. Una vez que la cadena converge, se tiene que los valores de $\boldsymbol{\theta}$ corresponden a observaciones de la distribución requerida, $p(\boldsymbol{\theta} \mid x)$. Sin embargo, en general, no se garantiza una muestra variables aleatorias *totalmente* independientes provenientes de la distribución $p(\theta \mid x)$, dado que el esquema del muestreador de Gibbs usa el valor actual para construir el siguiente valor; por ende, la secuencia de valores que se obtiene

estará correlacionada.

Ejemplo C.6. Se puede implementar el muestreador de Gibbs para generar una secuencia de observaciones con densidad conjunta

$$(x, y) \sim N_2\Big(0, \begin{pmatrix} \rho & 0 \\ 0 & \rho \end{pmatrix}\Big)$$

Teniendo en cuenta que la media de ambas variables es cero y su varianza uno, entonces la covarianza entre ambas variables será ρ (Robert and Casella, 2009). Por ende, partiendo de valores iniciales (x_t, y_t), el algoritmo se centra en actualizar las distribuciones condicionales según el resultado A.30.

$$x_{t+1} \mid y_t \sim N(\rho y_t, 1 - \rho^2)$$
$$y_{t+1} \mid x_{t+1} \sim N(\rho x_{t+1}, 1 - \rho^2)$$

```
bivariate.gibbs <- function (n, rho, x, y) {
  mat <- matrix(ncol = 2, nrow = n)
  mat[1, ] <- c(x, y)
  for (i in 2:n){
    x <- rnorm(1, rho * y, sqrt(1 - rho^2))
    y <- rnorm(1, rho * x, sqrt(1 - rho^2))
    mat[i, ] <- c(x, y)
  }
  mat<-as.data.frame(mat)
  return(mat)
}

biv <- bivariate.gibbs(n=2000, rho=0.5, x= 0, y = 0)
colMeans(biv)
```

```
##           V1           V2
## -0.009751821 -0.028181553
```

```
var(biv)
```

```
##           V1        V2
## V1 0.9705958 0.4782972
## V2 0.4782972 0.9944298
```

```
cor(biv)
```

```
##           V1        V2
## V1 1.0000000 0.4868459
## V2 0.4868459 1.0000000
```

```
plot(biv)
```

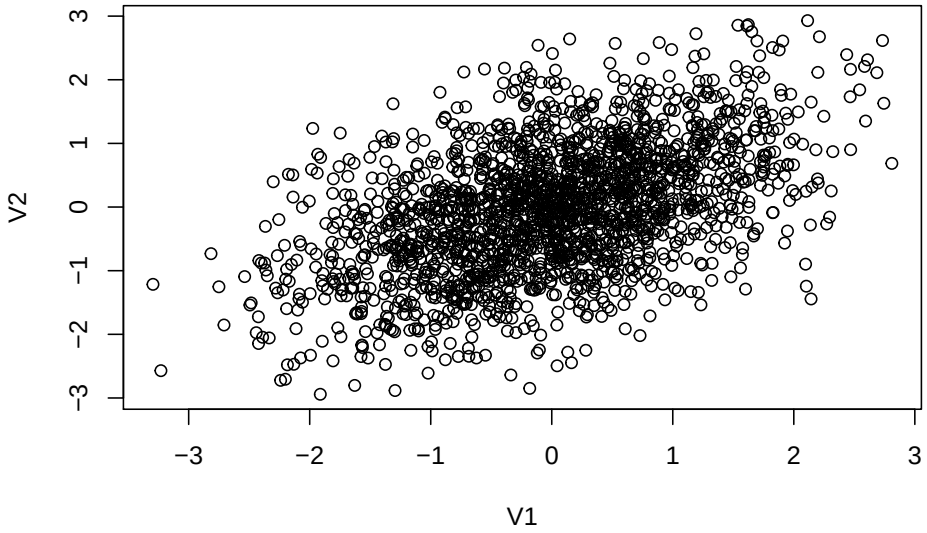


Figura C.6: Generación de valores para una distribución normal bivariada.

Ejemplo C.7. Un problema común es el de descartar los primeros valores, puesto que el algoritmo puede demorar en obtener convergencia; esto se puede resolver en forma empírica utilizando las medias y varianzas acumuladas y graficándolas se puede tomar una decisión acerca del valor óptimo en el que la cadena converge.

Con el siguiente código computacional, es posible corroborar que un punto de corte óptimo desde el cual se consideraría que las cadenas simuladas anteriormente es a partir de la iteración **600**.

```
g.diag <- function(sample){
  n <- length(sample)
  res <- matrix(nrow=2, ncol=n)
  for(i in 1:n){
    res[1, i] <- mean(sample[1 : i])
    res[2, i] <- var(sample[1 : i])
  }
  return(res)
}

m1 <- g.diag(biv[, 1])
m2 <- g.diag(biv[, 2])

par(mfcol = c(1, 2))
plot(m1[1, ], type = 'l', ylim=c(-0.6, 0.6), col=4)
```

```
lines(m2[1, ], lty = 2, col = 2)
title("Diagnóstico - Media acumulada")

plot(m1[2, ], type = 'l', ylim = c(0.5, 1.5), col=4)
lines(m2[2, ], lty = 2, col = 2)
title("Diagnóstico - Varianza acumulada")
```

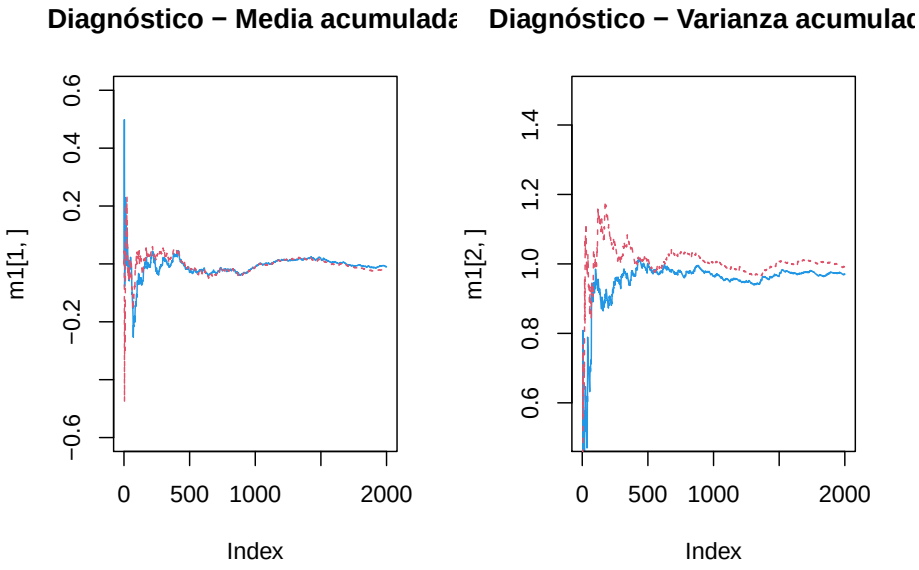


Figura C.7: Convergencia de la media y varianza usando el muestreador de Gibbs.

El muestrador de Gibbs también funciona en una "segunda fase", cuando queremos seleccionar una muestra de $f(\theta \mid x)$, es decir, la distribución de los parámetros dada la información observada x.

Ejemplo C.8. Suponga que y tiene distribución $N(\mu, \sigma^2 = 1/\phi)$ y queremos obtener una muestra de la distribución posterior del vector aleatorio $\boldsymbol{\theta} = (\mu, 1/\phi)$. Para este caso supongamos que conocemos las distribuciones previas; para la media μ se asume una distribución uniforme y para la varianza ϕ una distribución Gamma con parámetros a y b. La distribución posterior de (μ, ϕ) satisface:

$$p(\mu, \phi \mid y) \propto (\phi)^{n/2} \exp\left\{-\phi \frac{\sum_{j=1}^{n}(y_j - \mu)^2}{2}\right\} (\phi)^{a-1} exp(-b/\phi) \qquad \text{(C.2)}$$

En donde la primera parte después del signo de proporcionalidad, corresponde a la verosimilitud de la información observada y la segunda parte corresponde a la distribución posterior de ϕ; la distribución posterior de μ no aparece pues es una constante. Por tanto, ésta se puede escribir como:

$$p(\mu, \phi \mid y) \propto (\phi)^{n/2+a-1} exp\left\{-\phi\Big(\frac{\sum_{j=1}^{n}(y_j-\mu)^2}{2}+b\Big)\right\}$$

Acudiendo al resultado A.15, la distribución condicional de la varianza σ^2 dado (μ, y) es Gamma-inversa con parámetros $a + n/2$ y $\sum_{j=1}^{n}(y_j - \mu)^2/2 + b$. Por tanto,

$$\sigma^2 \mid \mu, x \sim Gamma - inversa\left(\theta + n/2, \sum_{j=1}^{n}(y_j-\mu)^2/2 + b\right) \tag{C.3}$$

Análogamente, la distribución de μ dado (σ^2, y) es normal con media $\bar{y}$ y varianza σ^2/n, es decir,

$$\mu \mid \sigma^2, y \sim N(\bar{y}, \sigma^2/n) \tag{C.4}$$

Para implementar el muestreador de Gibbs con estas distribuciones, primero se deben escoger valores apropiados para a y b, con el propósito de representar correctamente la distribución previa, y luego

- Defninir un valor inicial para la media y la varianza, (μ_0, σ_0^2).
- Generar $(\mu_{i+1}, \sigma_{i+1}^2)$ simulando μ_{i+1} de (C.3) y luego σ_{i+1}^2 de (C.4).
- Iterar para obtener $(\mu_0, \sigma_0^2), (\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \cdots$,.
- Suponiendo que el algoritmo converge después de m iteraciones, descartar los m primeros valores.

Entonces $(\mu_{m+1}, \sigma_{m+1}^2), (\mu_{m+2}, \sigma_{m+2}^2), \cdots$, es una muestra (correlacionada) de $p(\mu, \sigma^2 \mid x)$.

La siguiente función en `R` implementa el muestreador de Gibbs para el anterior ejemplo.

```
library(invgamma)

normal2 <- function(datos, a, b, nsim, inicial){
  n <- length(datos)
  xbar <- mean(datos)
  mu.now <- inicial[1]
  var.now <- inicial[2]
  dummy <- matrix(ncol = 2, nrow = nsim)
  dummy[1, 1] <- mu.now
  dummy[1, 2] <- var.now

  for (i in 2 : nsim){
    alp <- a + (n/2)
    bet <- b + (sum((datos - mu.now)^2)/2)
```

```
    var.next <- rinvgamma(1, shape = alp, rate = bet)
    mu.next <- rnorm(1, xbar, sqrt(var.now/n))
    dummy[i, 1] <- mu.next
    dummy[i, 2] <- var.next
    mu.now <- mu.next
    var.now <- var.next
  }
  return(dummy)
}

datos <- rnorm(100, 5, 2)
mc1.vals <- normal2(datos, a = 2, b = 5,
                    nsim = 1000, inicial = c(2, 2))
mc1.vals <- mc1.vals[101: 1000, ]
colMeans(mc1.vals)
```

```
## [1] 4.718260 4.056971
```

```
par(mfcol = c(1, 2))
plot(mc1.vals[, 1], type = 'l', ylab = 'mu')
plot(mc1.vals[, 2], type = 'l', ylab = 'sigma^2')
```

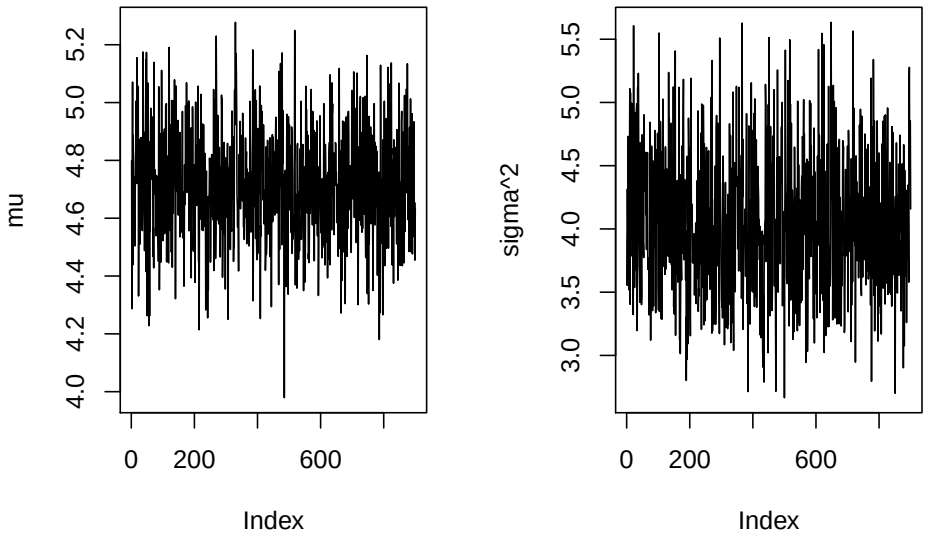


Figura C.8: Cadenas generadas desde el muestreador de Gibbs.

```
par(mfcol = c(1, 2))
hist(mc1.vals[, 1], prob = T, xlab='mu', main = "")
lines(density(mc1.vals[, 1], kernel='gaussian'))
hist(mc1.vals[, 2], prob = T, xlab='sigma^2', main = "")
```

```
lines(density(mc1.vals[, 2], kernel='gaussian'))
```

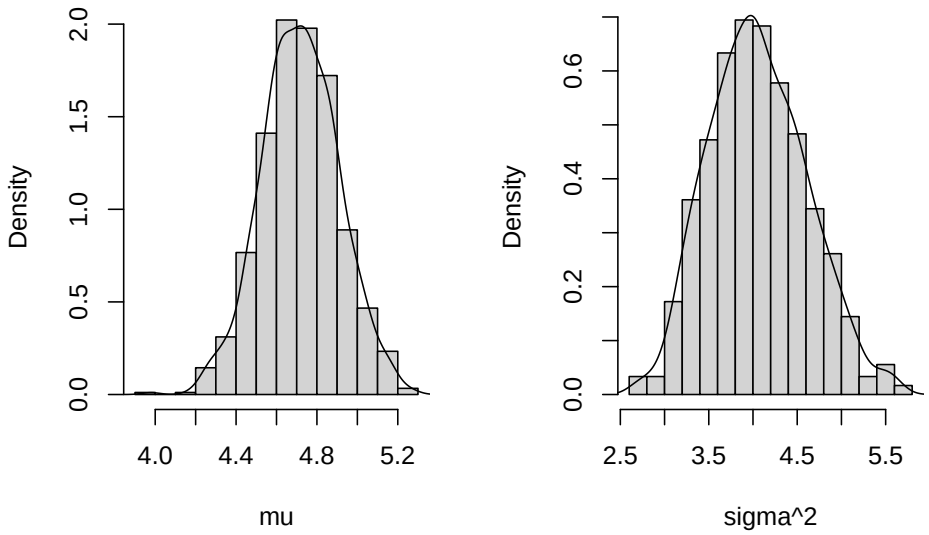


Figura C.9: Densidades posteriores generadas con el muestreador de Gibbs.

C.2.2. El algoritmo de Metrópolis-Hastings

Este algoritmo se basa en proponer un nuevo punto de acuerdo a una función de densidad adecuada y aceptar este nuevo valor propuesto con una probabilidad que depende del punto actual, del nuevo punto y de la densidad de la cual fue propuesto el nuevo punto.

Suponga que deseamos simular valores de una distribución multivariada $p(\theta \mid y)$. Sea la función de densidad propuesta $q(\theta, \theta')$, una función de densidad de probabilidad arbitraria que describe la probabilidad de aceptación de θ' a partir de la posición actual de θ. El algoritmo de Metropolis-Hastings está dado por los siguientes pasos:

- Siendo el valor actual θ_i, genere un valor candidato θ' obtenido como una observación de la densidad $q(\theta_i, \theta')$.
- Calcule

$$T(\theta_i, \theta') = \begin{cases} \text{mín}\left(1, \frac{p(\theta'|y)q(\theta', \theta_i)}{p(\theta_i|y)q(\theta_i, \theta')}\right), & \text{si } p(\theta_i \mid y)q(\theta_i, \theta') > 0, \\ 1, & \text{si } p(\theta_i \mid y)q(\theta_i, \theta') = 0 \end{cases}$$

- Acepte el nuevo valor y actualícelo a $\theta_{i+1} = \theta'$ con probabilidad $T(\theta_i, \theta')$. De otra forma, rechazar el valor candidato y defina $\theta_{i+1} = \theta_i$.
- Repita el paso anterior para obtener la secuencia $\theta_0, \theta_1, ...$, donde θ_0 denota un valor arbitrario de arranque.

- Descarte los primeros m valores obtenidos.

Siguiendo el anterior algoritmo, entonces se tiene que $\theta_{m+1}, \theta_{m+2}, \ldots$ es una secuencia (correlacionada) de la distribución requerida. En principio, puede ser usada cualquier densidad q, pero si ésta es escogida ingenuamente, la eficiencia de la cadena puede ser muy pobre. La relación más importante entre el muestreados de Gibbs y el algorítmo de Metropolis-Hastings, está dada como un teorema en el libro de Robert and Casella (2009, pág. 296).

Resultado C.2. *El muestreador de Gibbs es equivalente al algoritmo de Metropolis-Hastings, con la probabilidad de aceptación igual a uno para todos los puntos propuestos.*

Lo anterior implica que la convergencia para ambos métodos no es la misma. Para cerrar la sección de cadenas de Markov vía Monte Carlo, es importante hacernos la siguiente pregunta: ¿Son independientes las muestras simuladas? En principio no se puede hablar de independencia, pues es claro que la observación $\{i+1\}$ depende de la observación $\{i\}$. Dado que las observaciones resultantes se encuentran en estricto orden de medición, podríamos utilizar algunos criterios como la función de auto-correlación (ACF) y la función de auto-correlación parcial (PACF), para conocer sobre la correlación entre observaciones.

Siguiendo con el ejemplo C.8 del apartado de Gibbs, se ha escogido usar como como distribuciones propuestas q para la media y para la varianza, densidades normales centradas en el actual parámetro, ambas con varianza igual a uno. Dadas las distribuciones propuestas, algunos valores de la varianza pueden ser negativos; aunque este no es un problema porque la distribución posterior le asignará el valor cero, por tanto este valor será rechazado con un probabilidad de uno.

```
library(invgamma)

met.hast <- function(datos, a, b, iter, ini){
  mu0 <- ini[1]
  var0 <- ini[2]
  resul <- matrix(ncol = 2, nrow = iter)
  resul[1, 1] <- mu0
  resul[1, 2] <- var0
  for (i in 2 : iter){
    mu.prop <- rnorm(1, mu0, 1)
    var.prop <- rnorm(1, var0, 1)
    if (var.prop <= 0){ T.val <- 0 }
    else{
      p1 <- prod(dnorm(datos, mu.prop, sqrt(var.prop))) *
        dinvgamma(var.prop, shape = a, rate = b)
      q1 <- dnorm(mu0, mu.prop, 1) *
        dnorm(var0, var.prop, 1)
      p2 <- prod(dnorm(datos, mu0, sqrt(var0))) *
```

```
        dinvgamma(var0, shape = a, rate = b)
      q2 <- dnorm(mu.prop, mu0, 1) *
        dnorm(var.prop, var0, 1)
      T.val <- min(1, (p1 * q1)/(p2 * q2))
    }
    u <- runif(1)
    if (u <= T.val){
      resul[i, 1] <- mu.prop
      resul[i, 2] <- var.prop
      }
    else{
      resul[i, 1] <- mu0
      resul[i, 2] <- var0
      }
    mu0 <- resul[i, 1]
    var0 <- resul[i, 2]
  }
  return(resul)
}

datos <- rnorm(100, 5, 2)
mc2 <- met.hast(datos, a = 2, b = 5,
                iter = 1000, ini = c(2, 2))
colMeans(mc2)
```

```
## [1] 4.729000 4.387919
```

```
par(mfrow=c(2,2))
pacf(mc2[, 1], 100)
pacf(mc2[, 2], 100)
acf(mc2[, 1], 100)
acf(mc2[, 2], 100)
```

C.2.3. Buenas prácticas en la aplicación de métodos MCMC

Dado que una gran parte de la inferencia bayesiana está ligada a la programación e implementación de los métodos MCMC para realizar inferencias posteriores de los parámetros de interés, se sugiere seguir el razonamiento y recomendaciones de Gelman and Shirley (2010), que puede ser resumido en los siguientes ítemes para cada parámetro de interés:

1. Simulación de tres o más cadenas de forma paralela. Los valores iniciales de cada cadena deben estar dispersos entre sí.
2. Comprobación de la convergencia de las cadenas mediante el descarte de

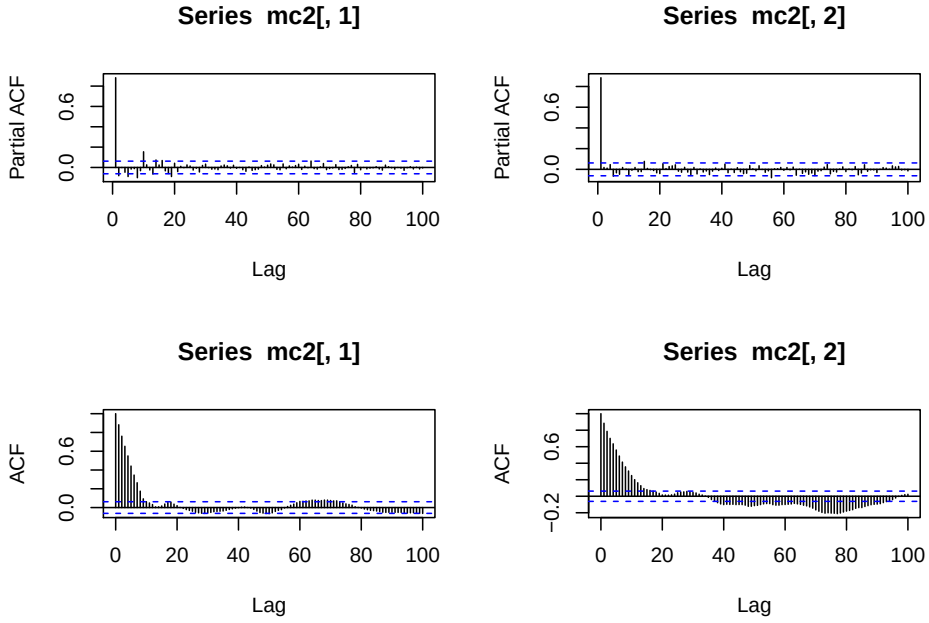


Figura C.10: Autocorrelación y autocorrelación parcial para las cadenas simuladas del algoritmo MH.

la primera mitad de los valores generados en las cadenas. Esta etapa se conoce como *burning stage*.

3. Una vez que las cadenas converjan, mezclar los tres conjuntos de valores generados por las cadenas. Esto garantiza, en primera instancia, que las cadenas no estén auto-correlacionadas.
4. Además de realizar esta mezcla, descartar valores intermedios mediante un muestreo sistemático. Esta etapa se conoce como *thining stage*. Al final se recomienda almacenar una cantidad elevada de valores simulados.
5. Calibrar el algoritmo si la convergencia de las cadenas no se presenta rápidamente.
   - Para los algoritmos de Metropolis-Hastings, escoger una distribución de salto acorde con la distribución de la cual se desea simular. Por ejemplo, Cepeda and Gamerman (2001) presentan dos distribuciones de salto para el problema de la modelación de la varianza (cada una de las propuestas presenta tasas de aceptación diferentes).
6. Comparación y contraste de los resultados con modelos simples que permitan examinar posibles discrepancias y corregir errores de programación.

En términos de inferencia bayesiana, se tienen dos tipos de procesos: el primero y más común, que trata de realizar inferencias acerca de un vector de parámetros de interés $\boldsymbol{\theta}$; el segundo trata con los momentos del parámetro, por ejemplo su esperanza. Nótese que el primer proceso se presenta con seguridad en ejercicios

empíricos simulados; sin embargo, el segundo se presenta en los ejercicios prácticos con datos reales, en donde se quiere contrastar alguna hipótesis.

Las anteriores dos opciones tienen tratamientos muy diferentes en términos de la cantidad de simulaciones requeridas. Por ejemplo, si el objetivo es inferir acerca de $\boldsymbol{\theta}$, para conocer su comportamiento estructural, basta con realizar una simulación que genere una cantidad mediana de valores y que se resumen en un promedio y una desviación estándar. Por otro lado, si el objetivo es inferir acerca de $E(\boldsymbol{\theta})$, se requieren muchas más simulaciones para obtener una buena precisión. Siguiendo a Gelman and Shirley (2010), una vez terminado el proceso de *burning* y *thining*, se sugiere que se dividan los valores simulados en las cadenas paralelas y se formen k grupos; de esta forma, una estimación de $E(\boldsymbol{\theta})$ será la gran media de las medias muestrales de cada grupo y el error estándar será su desviación estándar dividida por $\sqrt{k}$.

Referencias

Bibliografía

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley, 1 edition.

Brewer, K. (2002). *Combined Survey Sampling Inference: Weighing Basu's Elephants*. A Hodder Arnold Publication. Arnold.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes for Data Analysis*. Chapman and Hall/CRC, 1 edition.

Cavanaugh, J. E. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 31:201–208.

Cepeda, E. and Gamerman, D. (2001). Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics*, 14:207 – 221.

Dempster, A. P. (1974). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, pages 335 – 352. Department of Theoretical Statistics: University of Aarhus.

Efron, B. and Morris, C. (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70:311 – 319.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, 3(3):445 – 449.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC, 1 edition.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2 edition.

Gelman, A. and Shirley, K. (2010). *Handbook of Markov Chain Monte Carlo*, chapter Inference from Simulations and Monitoring Convergence. CRC.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer.

Jordan, M. I. (2004). The exponential family and generalized linear models.

Migon, H. S. and Gamerman, D. (1999). *Statistical Inference: An Integrated Approach*. Arnold.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer.

Robert, C. P. and Casella, G. (2009). *Introducing Monte Carlo Methods with R*. Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461 – 464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and VanderLinde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B 64:583 – 639.

Wikipedia (2011). 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.