Modelos Bayesianos con R y STAN

Andrés Gutiérrez - Hanwen Zhang

2021-05-30

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

[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.

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

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.

Apéndice A

Algunas distribuciones de probabilidad

ss

A.1. Distribuciones discretas

ss

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.

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.

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.

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.

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.