¿Qué es la Estadistica Bayesiana?

Edilberto Nájera Rangel

División académica de Ciencias Básicas, Universidad Juárez Autómona de Tabasco,

Carr. Cunduacán-Jalpa Km 1, Cunduacán Tabasco, México. A.P. 24, C.P. 86690

cbgj`cprm, l_hcp_>sh_r, k v

Recibido el 15 de enero de 2015. Aceptado el 25 de junio 2015.

El objetivo de la estadística, y en particular de la estadística bayesiana, es proporcionar una metodología para analizar adecuadamente la información con la que se cuenta (análisis de datos), y decidir de manera razonable sobre la mejor forma de actuar (teoría de decisión). El propósito de este trabajo es dar una introducción básica a la Inferencia Bayesiana, con el fin de que se tenga una visión general de ella.

The goal of statistics, in particular bayesian statistics, is to provide a methodology for adequately analyzing available information (data analysis), and afterwards choosing the best form of action (decision theory). The purpose of this article is to give a basic introduction to Bayesian Inference with the aim being to provide a general overview of the topic.

Palabras claves: Inferencia Bayesiana, Distribución a posteriori, Distribución a priori, Distribución a priori conjugada, Distribución no informativa de Jeffreys, Teorema de Bayes, Teoría de Decisión.
Keywords: Bayes Theorem, Bayesian Inference, Decision Theory, Jeffreys prior distribution, Posterior distribution, Prior distribution, Prior distribution conjugate.

1. Introducción

El interés por el teorema de Bayes trasciende la aplicación clásica, especialmente cuando se amplía a otro contexto en el que la probabilidad no se entiende exclusivamente como la frecuencia relativa de un suceso a largo plazo, sino como el grado de convicción personal acerca de que el suceso ocurra o pueda ocurrir (definición subjetiva de la probabilidad). Afirmaciones del tipo "es muy probable que el partido X gane las próximas elecciones", "es improbable que Juan haya sido quien llamó por teléfono" o "es probable que se encuentre un tratamiento eficaz para el sida en los próximos cinco años", normales en el lenguaje común, no pueden cuantificarse formalmente; resultan ajenas a una metodología que se desenvuelva en un marco frecuentista. Una cuantificación sobre una base subjetiva resulta, sin embargo, familiar y fecunda para el enfoque bayesiano. Al admitir un manejo subjetivo de la probabilidad, el analista bayesiano puede emitir juicios de probabilidad sobre una hipótesis H, y expresar por esa vía su grado de convicción al respecto, tanto antes como después de haber observado los datos. En su versión más elemental, el teorema de Bayes toma la forma siguiente:

$$Pr(H|datos) = \frac{Pr(datos|H)}{Pr(datos)} Pr(H).$$

La probabilidad a priori de una hipótesis, $Pr(H)$, se ve transformada en una pro-

babilidad a posteriori, $Pr(H|datos)$, una vez incorporada la evidencia que aportan los datos. El caso considarado se circunscribe a la situación más simple, aquella en la que $Pr(H)$ representa un número único; sin embargo, si se consigue expresar la convicción inicial (y la incertidumbre) mediante una distribución de probabilidades, entonces una vez observados los datos, el teorema "devuelve" una nueva distribución, que no es otra cosa que la percepción probabilística original actualizada por los datos.

Los métodos bayesianos han sido cuestionados argumentando que al incorporar las creencias o espectativas personales del investigador, éstas pueden ser caldo de cultivo para cualquier arbitrariedad o manipulación. Se podría argumentar, por una parte, que el enfoque frecuentista no está exento de decisiones subjetivas (nivel de significancia, usar una o dos colas, importancia que se concede a las diferencias, etc.). De hecho, la subjetividad (algo bien diferente de la arbitrariedad y el capricho) es un hecho inevitable, especialmente en un marco de incertidumbre como en el que operan las ciencias biológicas o sociales.

Aunque las bases de la estadística bayesiana datan de hace más de dos siglos, no es sino hasta fechas relativamente recientes cuando empieza a tener un uso creciente en el ámbito de la investigación. Una de las razones que explican esta realidad y que a la vez anuncian un desarrollo aún mayor en el futuro, es la absoluta necesidad del cálculo computarizado para la resolución de algunos problemas de mediana complejidad. Hoy ya existe software disponible que hace posible operar con estas técnicas, lo cual augura una aplicación cada vez mayor de los métodos bayesianos (ver [2] y [6]).

El marco teórico en el que se aplica la inferencia bayesiana es similar al de la clásica: hay un parámetro poblacional respecto al cual se desea realizar inferencia, y se tiene un modelo que determina la probabilidad de observar diferentes valores de un vector aleatorio X, bajo diferentes valores de los parámetros. Sin embargo, la diferencia fundamental es que la inferencia bayesiana considera al parámetro como una variable aleatoria, o vector aleatorio, según sea el caso, lo cual conduce a una aproximación diferente para realizar la modelación del problema y la inferencia propiamente dicha.

Algunos ejemplos que justifican lo anterior son: la verdadera proporción de artículos defectuosos que produce un proceso de manufactura puede fluctuar ligeramente, pues depende de numerosos factores; la verdadera proporción de casas que se pierden por concepto de hipoteca varía dependiendo de las condiciones económicas; la demanda promedio semanal de automóviles también fluctuará como una función de varios factores, incluyendo la temporada.

En esencia, la inferencia bayesiana está basada en la distribución de probabilidad del parámetro dados los datos (distribución a posteriori de probabilidad, $Pr\left(\theta \mid y\right)$), en lugar de la distribución de los datos dado el parámetro. Lo único que se requiere para el proceso de inferencia bayesiana es la especificación previa de una distribución a priori de probabilidad $Pr(\theta)$, la cual representa el conocimiento acerca del parámetro antes de obtener cualquier información a través de los datos.

La noción de la distribución a priori, o inicial, para el parámetro es el corazón del pensamiento bayesiano. El análisis bayesiano hace uso explícito de las probabilidades para cantidades inciertas (parámetros) en inferencias basadas en análisis estadístico de datos.

La inferencia bayesiana se basa en el uso de una distribución de probabilidad para describir todas las cantidades desconocidas relevantes a un problema de estimación.

2. Conceptos bayesianos básicos

En lo que resta de este trabajo f denotará diferentes funciones de densidad de probabilidad. En cada caso el contexto hará claro a cuál representa.

2.1 Teorema de Bayes

Sea $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^{'}$ un vector aleatorio cuya densidad de probabilidad $f(\mathbf{y}|\theta)$ depende de k parámetros que forman el vector $\theta = (\theta_1, \theta_2, ..., \theta_k)^{'}$. Supóngase también que θ tiene una densidad de probabilidad $f(\theta)$. Entonces la densidad de probabilidad condicional de θ dado el vector de observaciones $\mathbf{y} = (y_1, y_2, ..., y_n)^{'}$ es

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})},$$

donde $f(\mathbf{y}) \neq 0$. A esta ecuación se le conoce como el teorema de Bayes, donde $f(\mathbf{y})$ es la distribución de probabilidad marginal de $\mathbf{Y}$, la que puede ser expresada como

$$f(\mathbf{y}) = \begin{cases} \int f(\mathbf{y}|\theta)f(\theta)d\theta & \text{si } \theta \text{ es continuo,} \\ \sum_{\theta} f(\mathbf{y}|\theta)f(\theta) & \text{si } \theta \text{ es discreto.} \end{cases}$$

La suma o integral se realiza sobre el espacio paramétrico de θ. De este modo, el teorema de Bayes puede ser escrito como

$$f(\theta|\mathbf{y}) = cf(\mathbf{y}|\theta)f(\theta) \propto f(\mathbf{y}|\theta)f(\theta), \tag{1}$$

donde $f(\theta)$ representa lo que es conocido de θ antes de recolectar los datos y es llamada la distribución a priori, o inicial, de θ; $f(\theta|\mathbf{y})$ representa lo que se conoce de θ después de recolectar los datos y es llamada la distribución a posteriori, o final, de θ dado $\mathbf{Y} = \mathbf{y}$; c es una constante normalizadora necesaria para que $f(\theta|\mathbf{y})$ sume o integre uno.

Dado que se ha observado $\mathbf{Y} = \mathbf{y}$, entonces, puesto que θ es desconocido, a la función de θ definida como $l(\theta|\mathbf{y}) = f(\mathbf{y}|\theta)$ se le denomina función de verosimilitud de θ dado $\mathbf{y}$ (ver [3]). En general, otra observación de $\mathbf{Y}$ define otra función de verosimilitud distinta. Entonces la formula de Bayes puede ser expresada como

$$f(\theta|\mathbf{y}) \propto l(\theta|\mathbf{y})f(\theta).$$

Ejemplo 1. Sean el parámetro θ que a priori tiene una distribución uniforme en el intervalo [0,1], y la variable aleatoria Y que tiene una distribución de probabilidad binomial con parámetros m y θ, con m conocido. Entonces se tienen las siguientes funciones de distribución

$$\begin{aligned} f(\theta) &= 1, & 0 \leq \theta \leq 1, \\ f(y|\theta) &= \binom{m}{y}\theta^y(1-\theta)^{m-y}, & y = 0, 1, ..., m. \end{aligned}$$

Ahora, para una muestra aleatoria de tamaño n la función de verosimilitud está dada por

$$l(\theta|\mathbf{y}) = \left[\prod_{i=1}^{n}\binom{m}{y_i}\right]\theta^{\Sigma y_i}(1-\theta)^{nm-\Sigma y_i}, \qquad y_i \in \{0, 1, ..., m\}.$$

Al aplicar el teorema de Bayes dado en la ecuación (1), la distribución a posteriori de θ dada la muestra $\mathbf{y}$ queda expresada como

$$f(\theta|\mathbf{y}) = c\frac{(m!)^n}{\prod_{i=1}^n y_i! \prod_{i=1}^n (m-y_i)!}\theta^{\Sigma y_i}(1-\theta)^{nm-\Sigma y_i}.$$

Esta expresión puede escribirse de la siguiente manera,

$$f(\theta \mid \mathbf{y}) = c\frac{(m!)^n}{\prod_{i=1}^n y_i! \prod_{i=1}^n (m-y_i)!}\theta^{(\Sigma y_i+1)-1}(1-\theta)^{(nm-\Sigma y_i+1)-1},$$

que tiene la forma de una distribución beta con parámetro (Σy_i+1) y $(nm-\Sigma y_i+1)$. Luego el valor de la constante normalizadora c es

$$c = \frac{\Gamma(nm+2)}{\Gamma(\Sigma y_i+1)\Gamma(nm-\Sigma y_i+1)}\frac{\prod y_i! \prod (m-y_i)!}{(m!)^n}.$$

Nótese que es a través de $l(\theta|\mathbf{y})$ que los datos (información muestral) modifican el conocimiento previo de θ dado por $f(\theta)$.

Por último, es conveniente señalar que la información muestral $\mathbf{Y}$ por lo general será introducida en el modelo a través de estadísticas suficientes para θ, dado que éstas contienen toda la información referente a los datos. Así, dado un conjunto de estadísticas suficientes T para los parámetros en θ, $f(\mathbf{y}|\theta)$ podrá ser intercambiada por $f(t|\theta)$, donde t es una observación de T, para lo cual bastará con calcular la distribución condicional de T dado θ.

2.2 Naturaleza secuencial del Teorema de Bayes

Supóngase que se tiene una muetra inicial $\mathbf{y}_1$. Entonces, por la fórmula de Bayes dada anteriormente se tiene

$$f(\theta|\mathbf{y}_1) \propto l(\theta|\mathbf{y}_1)f(\theta).$$

Ahora supóngase que se tiene una segunda muestra $\mathbf{y}_2$ independiente de la primera muestra, entonces

$$f(\theta|\mathbf{y}_1,\mathbf{y}_2) \propto l(\theta|\mathbf{y}_1,\mathbf{y}_2)f(\theta) = l(\theta|\mathbf{y}_1)l(\theta|\mathbf{y}_2)f(\theta),$$

de donde

$$f(\theta|\mathbf{y}_1,\mathbf{y}_2) \propto l(\theta|\mathbf{y}_2)f(\theta|\mathbf{y}_1).$$

De esta manera, la distribución a posteriori obtenida con la primera muestra se convierte en la nueva distribución a priori para ser corregida por la segunda muestra. Este proceso puede repetirse indefinidamente. Así, si se tienen r muestras independientes, la distribución a posteriori puede ser recalculada secuencialmente para cada muestra de la siguiente manera,

$$f(\theta|\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_m) \propto l(\theta|\mathbf{y}_m)f(\theta|\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_{m-1}), \text{ para } m = 2,3,...,r.$$

Nótese que $f(\theta|\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_r)$ puede también ser obtenida partiendo de $f(\theta)$ y considerando al total de las r muestras como una sola gran muestra.

2.3 Distribución a priori difusa o no informativa

La distribución a priori cumple un papel importante en el análisis bayesiano, ya que mide el grado de conocimiento inicial que se tiene de los parámetros bajo estudio. Si bien su influencia disminuye a medida que más información muestral está disponible, el uso de una u otra distribución a priori determinará ciertas diferencias en la distribución a posteriori.

Si se tiene un conocimiento previo sobre los parámetros, éste se traducirá en una distribución a priori. Así, será posible plantear tantas distribuciones a priori como estados iniciales de conocimiento existan, y los diferentes resultados obtenidos en la distribución a posteriori bajo cada uno de los enfoques adquirirán una importancia en relación con la convicción que tenga el investigador sobre cada estado inicial. Sin embargo, cuando nada es conocido sobre los parámetros, la selección de una distribución a priori adecuada adquiere una connotación especial, pues será necesario elegir una distribución a priori que no influya sobre ninguno de los posibles valores de los parámetros en cuestión. Estas distribuciones a priori reciben el nombre de difusas o no informativas.

Método de Jeffreys

En situaciones generales, para un parámetro θ el método más usado es el de Jeffreys (ver [5]), que sugiere que si un investigador es ignorante con respecto al vector de parámetros $\theta = (\theta_1, \theta_2, ..., \theta_n)'$, entonces se debe cumplir la siguiente condición de invariancia: su opinión acerca de θ dada la evidencia $\mathbf{Y}$ debe ser la misma que para una transformación diferenciable uno a uno de θ, $\eta = g(\theta)$. Esto significa que los resultados a posteriori que obtenga el investigador deben ser los mismos, ya sea que use a θ o que use a η al analizar al mismo conjunto de datos $\mathbf{Y}$ con el mismo modelo. Una condición suficiente (ver [4]) para que se cumpla esta condición de invariancia es que

$$\sqrt{|detI(\theta)|}d\theta = \sqrt{|detI(\eta)|}d\eta,$$

donde $I(\theta)$ es la matriz de información de Fisher de tamaño nxn, cuyo elemento (i, j)-ésimo es

$$I_{ij} = -E_\theta \left[\frac{\partial^2 \log f(Y|\theta)}{\partial\theta_i \partial\theta_j} \right].$$

Una a priori invariante propuesta por Jeffreys es

$$f(\theta) \propto \sqrt{|I(\theta)|}. \tag{2}$$

Ejemplo 2. Sea la variable Y con una distribución $B(n, \theta)$,

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Entonces

$$\log f(y|\theta) = \log \binom{n}{y} + y \log\theta + (n-y)\log(1-\theta),$$

$$\frac{d \log f(y|\theta)}{d\theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta},$$

$$\frac{d^2 \log f(y|\theta)}{d\theta^2} = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}.$$

De aquí

$$E\left[-\frac{Y}{\theta^2}-\frac{n-Y}{(1-\theta)^2}\right]=\left[-\frac{n\theta}{\theta^2}-\frac{E(n-Y)}{(1-\theta)^2}\right],$$
$$E\left[-\frac{Y}{\theta^2}-\frac{n-Y}{(1-\theta)^2}\right]=-\frac{n}{\theta(1-\theta)}.$$

Así, por (2),

$$f(\theta)\propto\frac{\sqrt{n}}{\sqrt{\theta}\sqrt{1-\theta}}\ .$$

Prescindiendo de n se obtiene que la distribución a priori de θ es

$$f(\theta)\propto\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}\ ,$$

esto es, $\theta\sim Beta(0{,}5,0{,}5)$.

Ejemplo 3. Se aplicará el método de Jeffreys para calcular una distribución conjunta a priori para los parámetros de un modelo normal. Sea $Y\sim N(\mu,\sigma^2)$, con ambos parámetros desconocidos. Entonces

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

luego

$$\log f(y|\mu,\sigma)=\log\frac{1}{\sqrt{2\pi}}-\log\sigma-\frac{(y-\mu)^2}{2\sigma^2}\ .$$

La matriz de información de Fisher está dada por

$$I(\theta)=-E\left[\begin{array}{cc}\frac{\partial^2}{\partial\mu^2}\log f(Y|\mu,\sigma) & \frac{\partial^2}{\partial\mu\partial\sigma}\log f(Y|\mu,\sigma)\\ \frac{\partial^2}{\partial\sigma\partial\mu}\log f(Y|\mu,\sigma) & \frac{\partial^2}{\partial\sigma^2}\log f(Y|\mu,\sigma)\end{array}\right],$$

es decir,

$$I(\theta)=-E\left[\begin{array}{cc}-\frac{1}{\sigma^2} & -\frac{2(Y-\mu)}{\sigma^3}\\ -\frac{2(Y-\mu)}{\sigma^3} & \frac{1}{\sigma^2}-\frac{3(Y-\mu)^2}{\sigma^4}\end{array}\right].$$

De aquí

$$I(\theta)=\left[\begin{array}{cc}\frac{1}{\sigma^2} & 0\\ 0 & \frac{2}{\sigma^2}\end{array}\right].$$

Ahora, por (2), la distribución a priori no informativa para $\theta=(\mu,\sigma)$ es

$$f(\mu,\sigma)\propto\sqrt{\frac{2}{\sigma^4}}\propto\frac{1}{\sigma^2}.$$

Observación 1. Siguiendo el mismo procedimiento, se comprueba que las distribuciones a priori no informativas de Jeffreys para μ (con σ conocida) y σ (con μ conocida) son $f(\mu)\propto 1$ y $f(\sigma)\propto\sigma^{-1}$, por lo que si se supone independencia entre ambos parámetros se tendría que $f(\mu,\sigma)\propto f(\mu)f(\sigma)=\sigma^{-1}$, en vez de σ^{-2}.

2.4 Distribución a priori conjugada

En este caso, la distribución a priori es determinada completamente por una función de densidad conocida. Berger (ver [1]) presenta la siguiente definición para una familia conjugada.

Definición 1. Una clase P de distribuciones a priori es una familia conjugada para la clase de funciones de densidad R, si $f(\theta|y)$ está en la clase P para toda $f(y|\theta) \in R$ y toda $f(\theta) \in P$.

En este caso, la distribución inicial dominará a la función de verosimilitud y $f(\theta|y)$ tendrá la misma forma que $f(\theta)$, con los parámetros corregidos por la información muestral.

Ejemplo 4. Sean el parámetro θ que a priori tiene una distribución beta con parámetros α y β, Y una variable aleatoria con distribución binomial con parámetros m y θ, con m conocido. Entonces se tienen las siguientes funciones de distribución

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{(0,1)}(\theta),$$
$$f(y|\theta) = \binom{m}{y}\theta^{y}(1-\theta)^{m-y}, \qquad y = 0, 1, ..., m.$$

Ahora, para una muestra aleatoria de tamaño n la función de verosimilitud está dada por

$$l(\mathbf{y}|\theta) = \left(\prod_{i=1}^{n}\binom{m}{y_i}\right)\theta^{\Sigma y_i}(1-\theta)^{mn-\Sigma y_i}, \qquad y_i \in \{0, 1, ..., m\}.$$

Al aplicar el teorema de Bayes, la distribución final de θ dada la muestra $\mathbf{y}$ queda expresada como

$$f(\theta|\mathbf{y}) \propto l(\mathbf{y}|\theta)f(\theta),$$

o sea,

$$f(\theta|\mathbf{y}) \propto \theta^{\alpha+\Sigma y_i-1}(1-\theta)^{\beta+mn-\Sigma y_i-1},$$

que tiene la forma de una distribución beta con parámetros $(\alpha + \sum y_i)$ y $(\beta + nm - \sum y_i)$. Luego, la distribución final de θ tiene la misma forma que la distribución a priori, por lo que la clase de distribuciones a priori beta es una familia conjugada para la clase de funciones de densidad binomial.

Ejemplo 5. Sean el parámetro θ con una distribución $N(\mu_0, \tau_0^2)$, donde μ_0 y τ_0 son conocidos, y la variable X con una distribución $N(\theta, \sigma^2)$, con σ^2 conocido. Entonces se tienen las funciones de densidad

$$f(\theta) = \frac{1}{\sqrt{2\pi}\tau_o}\exp\left(-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)$$

y

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

Al aplicar el teorema de Bayes, se tiene

$$f(\theta|x) \propto \exp\left[-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{(x-\theta)^2}{\sigma^2}\right)\right].$$

Pero

$$\begin{aligned}
\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{(x-\theta)^2}{\sigma^2} &= \frac{\sigma^2(\theta-\mu_0)^2+\tau_0^2(x-\theta)^2}{\tau_0^2\sigma^2} \\
&= \frac{\sigma^2(\theta^2-2\theta\mu_0+\mu_0^2)+\tau_0^2(x^2-2x\theta+\theta^2)}{\tau_0^2\sigma^2} \\
&= \frac{(\sigma^2+\tau_0^2)\theta^2-2(\sigma^2\mu_0+\tau_0^2x)\theta}{\tau_0^2\sigma^2}+\frac{\sigma^2\mu_0^2+\tau_0^2x^2}{\sigma^2\tau_0^2} \\
&= \frac{\theta^2-\frac{2(\sigma^2\mu_0+\tau_0^2x)\theta}{\sigma^2+\tau_0^2}}{\frac{\tau_0^2\sigma^2}{\sigma^2+\tau_0^2}}+\frac{\sigma^2\mu_0^2+\tau_0^2x^2}{\sigma^2\tau_0^2} \\
&= \frac{\theta^2-\frac{2\left(\frac{\mu_0}{\tau_0^2}+\frac{x}{\sigma^2}\right)\theta}{\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}}}{\frac{1}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}}+\frac{\sigma^2\mu_0^2+\tau_0^2x^2}{\sigma^2\tau_0^2} \\
&= \frac{\left(\theta-\frac{\frac{\mu_0}{\tau_0^2}+\frac{x}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}}\right)^2-\left(\frac{\frac{\mu_0}{\tau_0^2}+\frac{x}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}}\right)^2}{\frac{1}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}}+\frac{\sigma^2\mu_0^2+\tau_0^2x^2}{\sigma^2\tau_0^2},
\end{aligned}$$

de donde se tiene que

$$f(\theta|x) \propto \exp\left(-\frac{1}{2}\frac{\left(\theta-\frac{\frac{\mu_0}{\tau_0^2}+\frac{x}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}}\right)^2}{\frac{1}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}}\right),$$

es decir,

$$f(\theta|x)=\frac{1}{\sqrt{2\pi}\tau_1}\exp\left(-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\tau_1^2}\right),$$

donde

$$\mu_1=\frac{\frac{1}{\tau_0^2}\mu_0+\frac{1}{\sigma^2}x}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}} \quad \text{y} \quad \frac{1}{\tau_1^2}=\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}.$$

Así $f(\theta|x) \sim N(\mu_1, \tau_1^2)$.

Observación 2. Precisiones de las distribuciones a priori y a posteriori.

(i) A $\frac{1}{\tau_0^2}$ se le conoce como la precisión de la distribución.

(ii) En este ejemplo se tiene: precisión a posteriori = precisión a priori+precisión de los datos.

Ejemplo 6. Sea $\mathbf{X} = (X_1, X_2, ..., X_n)$ un vector aleatorio, con las X_i independientes con distribución $N(\theta, \sigma^2)$, con σ conocido y $\theta \sim N(\mu_0, \tau_0^2)$. Entonces, al aplicar el teorema de Bayes, la distribución a posteriori de θ dada la muestra $\mathbf{x}$ queda expresada como

$$f(\theta|\mathbf{x}) \propto f(\theta)f(\mathbf{x}|\theta) = f(\theta)f(x_1|\theta)f(x_2|\theta)...f(x_n|\theta),$$
$$f(\theta|\mathbf{x}) \propto f(\theta)f(\mathbf{x}|\theta) = f(\theta)\prod_{i=1}^{n} f(x_i|\theta),$$
$$f(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2}\left(\frac{(x_i-\theta)^2}{\sigma^2}\right)\right],$$
$$f(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\right].$$

Ahora,

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\theta)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\theta)^2 = \sum_{i=1}^{n}[(x_i-\bar{x})^2 + 2(x_i-\bar{x})(\bar{x}-\theta) + (\bar{x}-\theta)^2] \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-\theta)\sum_{i=1}^{n}(x_i-\bar{x}) + n(\bar{x}-\theta)^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2,\end{aligned}$$

porque $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$. De aquí

$$f(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{1}{\sigma^2}(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2)\right)\right],$$

es decir,

$$f(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{(\theta-\mu_0)^2}{\tau_0^2} + \frac{(\bar{x}-\theta)^2}{\frac{\sigma^2}{n}}\right)\right].$$

Del ejemplo anterior se tiene que

$$f(\theta|x_1, x_2, ..., x_n) \sim N(\theta|\mu_n, \tau_n^2),$$

donde,

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{y} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

3. Inferencia Bayesiana

Dado que la distribución final contiene toda la información concerniente al parámetro de interés θ (información a priori y muestral), cualquier inferencia con respecto a θ consistirá en afirmaciones hechas a partir de dicha distribución.

3.1 Estimación puntual

La distribución final reemplaza a la función de verosimilitud como una expresión que incorpora toda la información. Así $f(\theta|\mathbf{y})$ es un resumen completo de la información acerca del parámetro θ; sin embargo, para algunas aplicaciones es deseable (o necesario) resumir esta información de alguna forma, especialmente si se desea proporcionar un simple "mejor" estimador del parámetro desconocido.

En el contexto bayesiano (ver por ejemplo [2]) existen dos formas de reducir la información contenida en $f(\theta|\mathbf{y})$ a un simple "mejor" estimador, a saber,

(a) A través del estimador de Bayes a posteriori.

(b) A través de la teoría de decisión.

Estimador de Bayes a posteriori.

El estimador de Bayes a posteriori se define de la siguiente manera.

Definición 2. Sea $(X_1, X_2, ..., X_n)$ una muestra de $f(x|\theta)$, donde θ es una variable aleatoria con función de densidad $g_\theta(\bullet)$. El estimador de Bayes a posteriori de $\tau(\theta)$ con respecto a la a priori $g_\theta(\bullet)$ está dado como $E(\tau(\theta)|x_1, x_2, ..., x_n)$.

Ejemplo 7. Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de $f(x|\theta) = \theta^x(1-\theta)^{1-x}$, con $x \in \{0, 1\}$ y $g_\theta(\theta) = I_{(0,1)}(\theta)$. Determinaremos los estimadores de θ y $\theta(1-\theta)$. En este caso se tiene,

$$f(\theta|x_1, x_2, ..., x_n) = \frac{g_\theta(\theta)\prod_{i=1}^n f(x_i|\theta)}{\int_0^1 g_\theta(\theta)\prod_{i=1}^n f(x_i|\theta)d\theta} = \frac{\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} I_{(0,1)}(\theta)}{\int_0^1 \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} d\theta}.$$

De aquí,

$$\begin{aligned}
E(\theta|x_1, x_2, ..., x_n) &= \frac{\int_0^1 \theta\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}d\theta}{\int_0^1 \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}d\theta} \\
&= \frac{\int_0^1 \theta^{\Sigma x_i+1}(1-\theta)^{n-\Sigma x_i}d\theta}{\int_0^1 \theta^{\Sigma x_i+1-1}(1-\theta)^{n-\Sigma x_i+1-1}d\theta} \\
&= \frac{\int_0^1 \theta^{\Sigma x_i+2-1}(1-\theta)^{n-\Sigma x_i+1-1}d\theta}{\int_0^1 \theta^{\Sigma x_i+1-1}(1-\theta)^{n-\Sigma x_i+1-1}d\theta} \\
&= \frac{\frac{\Gamma(\sum_{i=1}^n x_i+2)\Gamma(n-\sum_{i=1}^n x_i+1)}{\Gamma(n+3)}}{\frac{\Gamma(\sum_{i=1}^n x_i+1)\Gamma(n-\sum_{i=1}^n x_i+1)}{\Gamma(n+2)}} \\
&= \frac{\Gamma(n+2)\Gamma(\sum_{i=1}^n x_i+2)\Gamma(n-\sum_{i=1}^n x_i+1)}{\Gamma(n+3)\Gamma(\sum_{i=1}^n x_i+1)\Gamma(n-\sum_{i=1}^n x_i+1)} \\
&= \frac{\Gamma(n+2)(\sum_{i=1}^n x_i+1)\Gamma(\sum_{i=1}^n x_i+1)}{(n+2)\Gamma(n+2)\Gamma(\sum_{i=1}^n x_i+1)} \\
&= \frac{\sum_{i=1}^n x_i+1}{}.
\end{aligned}$$

Luego el estimador de Bayes a posteriori de θ, con respecto a la a priori uniforme, es

$$\frac{\sum_{i=1}^{n} x_i + 1}{n+2}.$$

Ahora,

$$\begin{aligned}
E(\theta(1-\theta)|x_1, x_2, ..., x_n) &= \frac{\int_0^1 \theta(1-\theta)\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} d\theta}{\int_0^1 \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i} d\theta} \\
&= \frac{\int_0^1 \theta^{\Sigma x_i+1}(1-\theta)^{n-\Sigma x_i+1} d\theta}{\int_0^1 \theta^{\Sigma x_i+1-1}(1-\theta)^{n-\Sigma x_i+1-1} d\theta} \\
&= \frac{\int_0^1 \theta^{\Sigma x_i+2-1}(1-\theta)^{n-\Sigma x_i+2-1} d\theta}{\int_0^1 \theta^{\Sigma x_i+1-1}(1-\theta)^{n-\Sigma x_i+1-1} d\theta} \\
&= \frac{\frac{\Gamma(\sum_{i=1}^n x_i+2)\Gamma(n-\sum_{i=1}^n x_i+2)}{\Gamma(n+4)}}{\frac{\Gamma(\sum_{i=1}^n x_i+1)\Gamma(n-\sum_{i=1} x_i+1)}{\Gamma(n+2)}} \\
&= \frac{\Gamma(n+2)\Gamma(\sum_{i=1}^n x_i+2)\Gamma(n-\sum_{i=1}^n x_i+2)}{\Gamma(n+4)\Gamma(\sum_{i=1}^n x_i+1)\Gamma(n-\sum_{i=1}^n x_i+1)}.
\end{aligned}$$

Simplificando la última expresión resulta,

$$E(\theta(1-\theta)|x_1, x_2, ..., x_n) = \frac{(\sum_{i=1}^n x_i + 1)(n - \sum_{i=1}^n x_i + 1)}{(n+3)(n+2)}. \tag{3}$$

Así, el estimador de Bayes a posteriori de $\theta(1-\theta)$, con respecto a la a priori uniforme, es el dado por (3).

Ejemplo 8. Sean $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución normal $N(\theta, 1)$ y $\theta \sim N(\mu_0, 1)$. Puesto que

$$f(\theta|x) \propto f(\theta) f(x|\theta),$$

donde $x = (x_1, x_2, ..., x_n)$ es una observación de $(X_1, X_2, ..., X_n)$, entonces,

$$f(\theta|x) \propto e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2 - \frac{1}{2}(\theta-\mu_0)^2}.$$

Haciendo $x_0 = \mu_0$,

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\theta)^2 + (\theta-\mu_0)^2 &= \sum_{i=0}^{n}(x_i-\theta)^2 = \sum_{i=0}^{n}(\theta^2 - 2\theta x_i + x_i^2) \\
&= (n+1)\theta^2 - 2\theta(n+1)\bar{x} + \sum_{i=0}^{n} x_i^2 \\
&= (n+1)(\theta-\bar{x})^2 - (n+1)\bar{x}^2 + \sum_{i=0}^{n} x_i^2 \ .
\end{aligned}$$

De aquí

$$f(\theta|x) \propto e^{-\frac{n+1}{2}(\theta-\bar{x})^2},$$

de donde

$$f(\theta|x) = \frac{1}{\sqrt{\frac{2\pi}{n+1}}} e^{-\frac{n+1}{2}(\theta-\bar{x})^2} = \frac{1}{\sqrt{\frac{2\pi}{n+1}}} e^{-\frac{n+1}{2}(\theta-\sum_{i=0}^{n}\frac{x_i}{n+1})^2}.$$

Así,

$$E(\theta|x_1, x_2, ..., x_n) = \frac{\sum_{i=0}^{n} x_i}{n+1} = \frac{\mu_0 + \sum_{i=1}^{n} x_i}{n+1}.$$

Aproximación a la teoría de la decisión

Para los bayesianos, el problema de estimación es un problema de decisión. Asociada con cada estimador a hay una función de pérdida $L(\theta, a)$ que refleja la diferencia entre θ y a. $L(\theta, a)$ cuantifica las posibles penalizaciones al estimar θ por a. A través de la minimización de la pérdida esperada final

$$E(L(\theta, a)) = \int L(\theta, a) f(\theta|x) d\theta, \tag{4}$$

se obtiene el estimador puntual de θ, para la elección particular de la función de pérdida.

Definición 3. La regla de Bayes con respecto a la función de pérdida $L(\theta, a)$ y a la función $f(\theta|x)$, es la acción a que minimiza (4) y la cual se denotará por $d(x)$.

Hay muchas funciones de pérdida que se pueden usar. La elección en particular de una de ellas dependerá del contexto del problema, como se ilustra en los ejemplos siguientes.

Ejemplo 9. Un ingeniero debe construir un muro contra las inundaciones debidas a las crecidas de un río. Si decide construir un muro con altura a (medida a partir del nivel medio del río), entonces tendrá un costo de ac millones de pesos, donde c es una constante. Si hay una crecida del río, no habrá ningún daño si su altura θ (medida a partir del nivel medio del río) es menor que a; pero habrá pérdidas valuadas en $C(\theta - a)$ millones de pesos (donde C es una constante), si $\theta > a$. En este caso la función de pérdida es

$$L(\theta, a) = ac + C(\theta - a) I_{(\theta>a)},$$

donde $I_{(\theta>a)}$ es la función indicadora del conjunto $\{\theta \mid \theta > a\}$.

Ejemplo 10. Consideremos ahora un juego entre dos jugadores, donde cada jugador elige un número positivo. Desde el punto de vista del jugador 1, $\Theta = \{1, 2, ...\}$, que es el mismo conjunto de estimadores que tal jugador usará. El jugador que elija el número mayor gana 100 pesos del otro. En esta situación la función de pérdida es

$$L(\theta, a) = 100 I_{(\theta>a)} - 100 I_{(\theta<a)}.$$

En 3.3 veremos otros ejemplos. Las funciones de pérdida más usadas son:

1. Pérdida cuadrática,

$$L(\theta, a) = (\theta - a)^2.$$

2. Pérdida error absoluto o lineal absoluta,

$$L(\theta, a) = |\theta - a|.$$

3. Pérdida lineal. Para c_1, $c_2 > 0$,

$$L(\theta, a) = \begin{cases} c_1(a-\theta) & \text{si } a \geq \theta, \\ c_2(\theta - a) & \text{si } a < \theta. \end{cases}$$

Ejemplo 11. Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución $f(x|\theta)$, con $\theta \in \mathbb{R}$. Sea $L(\theta, a) = (\theta - a)^2$. Entonces

$$E(L(\theta, a)) = E(\theta - a)^2 = \int (\theta - a)^2 f(\theta|x) d\theta.$$

Ahora, si

$$\theta_m = \int \theta f(\theta|x) d\theta,$$

$$\begin{aligned} E(\theta - a)^2 &= E((\theta - \theta_m + \theta_m - a)^2) \\ &= E((\theta - \theta_m)^2 + 2(\theta - \theta_m)(\theta_m - a) + (\theta_m - a)^2) \\ &= E((\theta - \theta_m)^2) + 2E((\theta - \theta_m)(\theta_m - a)) + E((\theta_m - a)^2) \\ &= E((\theta - \theta_m)^2) + (\theta_m - a)^2 \\ &= Var(\theta) + (\theta_m - a)^2. \end{aligned}$$

Por lo tanto la pérdida esperada es mínima cuando

$$a = \theta_m.$$

De aquí se concluye que la regla de Bayes con respecto a la función de pérdida cuadrática es $d(x) = E[\theta|x]$.

Ejemplo 12. Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución $f(x|\theta)$, con $\theta \in \mathbb{R}$, θ variable aleatoria continua. Sea ahora $L(\theta, a) = |\theta - a|$. Entonces

$$E(L(\theta, a)) = \int |\theta - a| f(\theta|x) d\theta,$$

donde

$$|\theta - a| = \begin{cases} \theta - a & \text{si } a \leq \theta, \\ -(\theta - a) & \text{si } \theta < a, \end{cases}$$

Ahora,

$$\begin{aligned}\int_{-\infty}^{\infty}|\theta-a|f(\theta|x)d\theta &= \int_{-\infty}^{a}-(\theta-a)f(\theta|x)d\theta+\int_{a}^{\infty}(\theta-a)d\theta\\ &=\int_{-\infty}^{a}(a-\theta)f(\theta|x)d\theta+\int_{a}^{\infty}(\theta-a)f(\theta|x)d\theta\\ &=\int_{-\infty}^{a}af(\theta|x)d\theta-\int_{-\infty}^{a}\theta f(\theta|x)d\theta\\ &+\int_{a}^{\infty}\theta f(\theta|x)d\theta-\int_{a}^{\infty}af(\theta|x)d\theta\\ &=a\int_{-\infty}^{a}f(\theta|x)d\theta-\int_{-\infty}^{a}\theta f(\theta|x)d\theta\\ &+\int_{a}^{\infty}\theta f(\theta|x)d\theta-a\int_{a}^{\infty}f(\theta|x)d\theta,\end{aligned}$$

de aquí,

$$\begin{aligned}\frac{d}{da}\left(\int_{-\infty}^{\infty}|\theta-a|f(\theta|x)d\theta\right) &= \int_{-\infty}^{a}f(\theta|x)d\theta+af(a|x)-af(a|x)-af(a|x)\\ &-\int_{a}^{\infty}f(\theta|x)d\theta+af(a|x)\\ &=\int_{-\infty}^{a}f(\theta|x)d\theta-\int_{a}^{\infty}f(\theta|x)d\theta.\end{aligned}$$

Si $\frac{d}{da}\left(\int_{-\infty}^{\infty}|\theta-a|f(\theta|x)d\theta\right)=0$, entonces

$$\int_{-\infty}^{a}f(\theta|x)d\theta=\int_{a}^{\infty}f(\theta|x)d\theta,$$

de donde se sigue que

$$\int_{-\infty}^{a}f(\theta|x)d\theta=\frac{1}{2}.$$

Por lo tanto, el único valor crítico de $\int_{-\infty}^{\infty}|\theta-a|f(\theta|x)d\theta$ es la mediana, la cual se denotará por a_m.

Ahora, como

$$\begin{aligned}\frac{d^2}{da^2}\left(\int_{-\infty}^{\infty}|\theta-a|f(\theta|x)d\theta\right) &= \frac{d}{da}\left(\int_{-\infty}^{a}f(\theta|x)d\theta-\int_{a}^{\infty}f(\theta|x)d\theta\right)\\ &=f(a|x)+f(a|x)\\ &=2f(a|x),\end{aligned}$$

entonces

$$\frac{d^2}{da^2}\left(\int_{-\infty}^{\infty}|\theta-a|f(\theta|x)d\theta\right)|_{a=a_m}=2f(a_m|x)>0.$$

Por lo tanto a_m es un punto de mínimo local de esta función, y como es el único punto de extremo relativo, entonces la función alcanza su mínimo absoluto en a_m.

De aquí se tiene que la regla de Bayes, con respecto a la función de pérdida lineal absoluta, es la mediana de la distribución final.

Ejemplo 13. Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución $f(x|\theta)$, con $\theta \in \mathbb{R}$, θ variable aleatoria continua. Sea ahora $L(\theta, a)$ la función de pérdida lineal. Entonces

$$E(L((\theta, a)) = \int_{-\infty}^{\infty} L(\theta - a) f(\theta|x) d\theta,$$

donde

$$L(\theta, a) = \begin{cases} c_1(a - \theta) & \text{si } a \geq \theta \ , \\ c_2(\theta - a) & \text{si } a < \theta \ . \end{cases}$$

Ahora,

$$\begin{aligned} \int_{-\infty}^{\infty} L(\theta, a) f(\theta|x) d\theta &= \int_{-\infty}^{a} c_1(a - \theta) f(\theta|x) d\theta + \int_{a}^{\infty} c_2(\theta - a) f(\theta|x) d\theta \\ &= \int_{-\infty}^{a} (c_1 a - c_1\theta) f(\theta|x) d\theta + \int_{a}^{\infty} (c_2\theta - c_2 a) f(\theta|x) d\theta \\ &= \int_{-\infty}^{a} c_1 a f(\theta|x) d\theta - \int_{-\infty}^{a} c_1\theta f(\theta|x) d\theta \\ &+ \int_{a}^{\infty} c_2\theta f(\theta|x) d\theta - \int_{a}^{\infty} c_2 a f(\theta|x) d\theta \\ &= a \int_{-\infty}^{a} c_1 f(\theta|x) d\theta - \int_{-\infty}^{a} c_1\theta f(\theta|x) d\theta \\ &+ \int_{a}^{\infty} c_2\theta f(\theta|x) d\theta - a \int_{a}^{\infty} c_2 f(\theta|x) d\theta, \end{aligned}$$

de aquí, procediendo igual que en el ejemplo anterior,

$$\frac{d}{da} \left(\int_{-\infty}^{\infty} L(\theta, a) f(\theta|x) d\theta \right) = \int_{-\infty}^{a} c_1 f(\theta|x) d\theta - \int_{a}^{\infty} c_2 f(\theta|x) d\theta.$$

Si $\frac{d}{da} \left(\int_{-\infty}^{\infty} L(\theta, a) f(\theta|x) d\theta \right) = 0$, entonces

$$\int_{-\infty}^{a} c_1 f(\theta|x) d\theta = \int_{a}^{\infty} c_2 f(\theta|x) d\theta,$$

de donde se tiene

$$\int_{a}^{\infty} f(\theta|x) d\theta = \frac{c_1}{c_2} \int_{-\infty}^{a} f(\theta|x) d\theta. \tag{5}$$

Sustituyendo en

$$1 = \int_{-\infty}^{a} f(\theta|x) d\theta + \int_{a}^{\infty} f(\theta|x) d\theta$$

la expresión de $\int_a^{\infty} f(\theta|x) d\theta$ dada por (5), se obtiene

$$\int_{-\infty}^{a} f(\theta|x) d\theta = \frac{c_2}{c_1 + c_2}.$$

Así, el percentil $\hat{a} = \frac{c_2}{c_1 + c_2}$ es el valor crítico de $E[L(\theta, a)]$.

Por otro lado,

$$\begin{aligned} \frac{d^2}{da^2} \left(\int_{-\infty}^{\infty} L(\theta, a) f(\theta|x) d\theta \right) &= \frac{d}{da} \left(\int_{-\infty}^{a} c_1 f(\theta|x) d\theta - \int_{a}^{\infty} c_2 f(\theta|x) d\theta \right) \\ &= c_1 f(a|x) + c_2 f(a|x) \\ &= (c_1 + c_2) f(a|x), \end{aligned}$$

luego,

$$\frac{d^2}{da^2}\left(\int_{-\infty}^{\infty} L(\theta,a)f(\theta|x)d\theta\right)|_{a=\hat{a}} = (c_1+c_2)f(\hat{a}|x) > 0.$$

Por lo tanto, el percentil $\frac{c_2}{c_1+c_2}$ es un punto mínimo local de esta función, y como es el único punto de extremo relativo, entonces la función alcanza en él su mínima pérdida lineal.

Así, la regla de Bayes con respecto a la funciónde pérdida lineal es el percentil

$$\frac{c_2}{c_1+c_2}.$$

3.2 Regiones de credibilidad

La idea de una región de credibilidad es proporcionar el análogo de un intervalo de confianza en estadística clásica. El razonamiento es que los estimadores puntuales no proporcionan una medida de la precisión de la estimación. Esto causa problemas en la estadística clásica dado que los parámetros no son considerados como variables aleatorias, y por lo tanto no es posible dar un intervalo con la interpretación de que existe una cierta probabilidad de que el parámetro esté en el intervalo. En la teoría bayesiana no hay dificultad para realizar esta aproximación, porque los parámetros son tratados como variables aleatorias.

Definición 4. : Sea Θ el espacio parametral de θ. Una región $C \subseteq \Theta$ tal que

$$\int_C f(\theta|x)d\theta = 1-\alpha,$$

se llama una región del $100(1-\alpha)\,\%$ de credibilidad de θ. Si θ es discreta, en la expresión anterior reemplazamos la integral por una suma. Si $\Theta \subseteq \mathbb{R}$, las regiones de credibilidad conexas se llaman intervalos de credibilidad.

Un aspecto importante con las regiones de credibilidad (y lo mismo sucede con los intervalos de confianza) es que no están definidos de manera única. Cualquier región con probabilidad $(1-\alpha)$ cumple la definición. Sin embargo, generalmente se desea el intervalo que contiene únicamente los valores "más" posibles del parámetro, por lo que es usual imponer una restricción adicional que indica que el ancho del intervalo debe ser tan pequeño como sea posible. Para hacer esto se deben considerar sólo aquellos puntos con $f(\theta|x)$ más grandes. Esto conduce a un intervalo (o región) de la forma

$$C = C_\alpha(x) = \{\theta : f(\theta|x) \geq \gamma\},$$

donde γ es elegido tal que $\int_C f(\theta|x)d\theta = 1-\alpha$.

La región C que cumple las anteriores condiciones se denomina "región de densidad de probabilidad más grande" (HPD).

Ejemplo 14. (Media de una normal) Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución $N(\theta, \sigma^2)$, con σ^2 conocido y $\theta \sim N(b, d^2)$. Del ejemplo 6 se sabe que

$$\theta|x \sim N\left(\frac{\frac{b}{d^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{d^2}+\frac{n}{\sigma^2}}\right).$$

Sean

$$\mu_n = \frac{\frac{b}{d^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}},$$

y

$$\tau_n^2 = \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}.$$

Si z_α es el percentil α de la distribución normal estándar, entonces un intervalo del $100(1-\alpha)\,\%$ de credibilidad de θ es

$$\left[\mu_n - \tau_n z_{1-\frac{\alpha}{2}}, \mu_n + \tau_n z_{1-\frac{\alpha}{2}}\right].$$

Si n es grande, entonces $\mu_n \approx \bar{x}$ y $\tau_n^2 \approx \frac{\sigma^2}{n}$, luego el intervalo del $100(1-\alpha)\,\%$ de credibilidad de θ es aproximadamente igual al intervalo del $100(1-\alpha)\,\%$ de confianza de θ, de la estadística clásica. Sin embargo, las interpretaciones de ambos intervalos son distintas.

Ejemplo 15. Sea $(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución de Poisson, $P(\lambda)$, con $\lambda \sim$gamma (a, b), donde a es un entero positivo. Entonces

$$f(\lambda|x) \propto \lambda^{\Sigma x_i} e^{-n\lambda} \lambda^{a-1} e^{-\frac{\lambda}{b}},$$

o sea

$$f(\lambda|x) \propto \lambda^{\Sigma x_i + a - 1} e^{-\lambda(n+\frac{1}{b})},$$

de donde se tiene que

$$\lambda|x \sim gamma\left(a + \Sigma x_i, \left(n + \frac{1}{b}\right)^{-1}\right).$$

Si

$$\theta = \frac{2(nb+1)}{b}\lambda,$$

entonces

$$f(\theta) \propto \theta^{\Sigma x_i + a - 1} e^{-\frac{\theta}{2}},$$

es decir,

$$\theta \sim \chi^2_{2(\Sigma x_i + a)}.$$

Si $\chi^2_{2(\Sigma x_i+a),\alpha}$ es el percentil α de la distribución $\chi^2_{2(\Sigma x_i+a)}$, un intervalo del $100(1-\alpha)\,\%$ de credibilidad de λ es

$$\left[\frac{b}{2(nb+1)}\chi^2_{2(\Sigma x_i+a),\frac{\alpha}{2}}, \frac{b}{2(nb+1)}\chi^2_{2(\Sigma x_i+a),1-\frac{\alpha}{2}}\right].$$

Si $a = b = 1, n = 10$ y $\Sigma_{i=1}^n x_i = 6$, entonces, ya que $\chi^2_{14,0,05} = 6{,}571$ y $\chi^2_{14,0,95} = 23{,}685$, un intervalo del $90\,\%$ de credibilidad de λ es

$$[0{,}299, 1{,}077].$$

Definición 5. Sea $f(x)$ una función de densidad. Se dice que $f(x)$ es unimodal si existe $a \in \mathbb{R}$ que satisface las dos condiciones siguientes:

(a) Si $y \leq x \leq a$, entonces $f(y) \leq f(x) \leq f(a)$.

(b) Si $a \leq x \leq y$, entonces $f(a) \geq f(x) \geq f(y)$.

El número a se llama moda de $f(x)$.

Observación 3. Si $f(x)$ es una función de densidad unimodal, entonces la moda no necesariamente es única.

Ejemplo 16. Sea $X \sim U(0,1)$. Entonces la función de densidad de X es

$$f(x) = \begin{cases} 1 & \text{si } x \in (0,1)\,, \\ 0 & \text{en otro caso}\,. \end{cases}$$

En este caso todo $x \in (0,1)$ es una moda de $f(x)$.

El siguiente teorema (ver [3]) nos dice cómo obtener el intervalo de credibilidad de mínima longitud, cuando la densidad $f(\theta|x)$ es unimodal.

Teorema 4. Sea $f(x)$ una función de densidad unimodal. Si el intervalo $[a,b]$ satisface

(i) $\int_a^b f(x)dx = 1-\alpha$,

(ii) $f(a) = f(b) > 0$, y

(iii) $a \leq x^* \leq b$, donde x^* es una moda de $f(x)$,

entonces $[a,b]$ es el intervalo más corto que satisface (i).

Prueba. Sea $[a^{'}, b^{'}]$ cualquier intervalo con $b^{'} - a^{'} < b - a$. Se probará que esto implica que $\int_{a'}^{b'} f(x)dx < 1-\alpha$. El resultado sólo se demostrará para $a^{'} \leq a$; la prueba es similar si $a < a^{'}$. También se consideran dos casos, $b^{'} \leq a$ y $b^{'} > a$.

Si $b^{'} \leq a$, entonces $a^{'} \leq b^{'} \leq a \leq x^*$, luego, si $a^{'} \leq x \leq b^{'} \leq x^*$, tenemos $f(a^{'}) \leq f(x) \leq f(b^{'})$, por lo tanto

$$\int_{a'}^{b'} f(x)dx \leq f(b^{'})(b^{'} - a^{'}).$$

Como $b^{'} \leq a \leq x^*$, entonces $f(b^{'}) \leq f(a)$; también $b^{'} - a^{'} < b - a$. De aquí

$$\int_{a'}^{b'} f(x)dx \leq f(a)(b^{'} - a^{'}) < f(a)(b-a).$$

Ahora, como $a \leq x^* \leq b$ y $f(a) = f(b)$, entonces $f(a) \leq f(x)$ si $x \in [a,b]$, de donde

$$f(a)(b-a) \leq \int^{b} f(x)dx.$$

Así,

$$\int_{a'}^{b'} f(x)dx < f(a)(b-a) \leq \int_{a}^{b} f(x)dx = 1-\alpha.$$

Esto completa la demostración del primer caso.

Si $b' \geq b$, entonces $b' - a' \geq b - a$, luego en este caso nada hay que hacer. Supongamos ahora que $a' \leq a < b' < b$. En este caso se puede escribir

$$\begin{aligned}\int_{a'}^{b'} f(x)dx &= \int_{a}^{b} f(x)dx + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right] \\ &= (1-\alpha) + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right],\end{aligned}$$

y el teorema quedará demostrado si mostramos que la expresión entre corchetes es negativa. Ahora, usando el hecho de que f es unimodal, del ordenamiento $a' \leq a < b' < b$ y (ii) tenemos

$$\int_{a'}^{a} f(x)dx \leq f(a)(a-a')$$

y

$$\int_{b'}^{b} f(x)dx \geq f(b)(b-b'),$$

de donde

$$\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx \leq f(a)(a-a') - f(b)(b-b').$$

Ya que $f(a) = f(b),\ b' - a' < b - a\ y\ f(a) > 0,$

$$\begin{aligned}f(a)(a-a') - f(b)(b-b') &= f(a)[(a-a') - (b-b')] \\ &= f(a)[(b'-a') - (b-a)] \\ &< 0,\end{aligned}$$

o sea,

$$\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx < 0.$$

■

Corolario 1. Si la función de densidad final $f(\theta|x)$ es unimodal, entonces, para un valor dado de α, el intervalo del $100(1-\alpha)\,\%$ de credibilidad de longitud mínima de θ está dado por

$$\{\theta : f(\theta|x) \geq k\},$$

donde

$$\int_{\{\theta: f(\theta|x) \geq k\}} f(\theta|x)d\theta = 1-\alpha.$$

Ejemplo 17. Sea $(X_1, X_2, ..., X_n)$ como en el ejemplo 15. Por el teorema anterior, la región HPD está dada por

$$\{\lambda : f(\lambda|x) \geq k\},$$

donde k es elegido tal que

$$\int_{\{\lambda: f(\lambda|x)\geq k\}} f(\lambda|x)dx = 1-\alpha.$$

Como $\lambda|x \sim$ gamma $(a+\Sigma x_i, (n+\frac{1}{b})^{-1})$, se requiere encontrar λ_L y λ_U tales que

$$f(\lambda_L|x) = f(\lambda_U|x)$$

y

$$\int_{\lambda_L}^{\lambda_U} f(\lambda|x)d\lambda = 1-\alpha.$$

Si $a=b=1, n=10$ y $\Sigma x_i = 6$, mediante un procedimiento numérico se encuentra que la región HPD del 90 % de credibilidad de λ es $[0{,}253, 1{,}005]$.

3.3 Pruebas de hipótesis

Las pruebas de hipótesis son decisiones en las que se debe elegir una de dos hipótesis diferentes,

$$\begin{cases} H_0 : \theta \in \Theta_0, \\ H_1 : \theta \in \Theta_0^c, \end{cases}$$

donde Θ es el espacio parametral, $\Theta_0 \subset \Theta$, $\Theta_0 \neq \emptyset$, $\Theta_0^c \neq \emptyset$.

En un problema de prueba de hipótesis sólo se pueden realizar dos acciones, aceptar H_0 o rechazar H_0. Estas dos acciones se denotan por medio de a_0 y a_1, respectivamente.

En un problema de prueba de hipótesis, la función de pérdida debe reflejar el hecho de que si $\theta \in \Theta_0$ pero se toma la decisión a_1, o si $\theta \in \Theta_0^c$ pero se toma la decisión a_0, entonces se ha cometido un error. Sin embargo, en los otros dos casos se ha tomado la decisión correcta.

La función de pérdida más simple es la función de pérdida $0-1$, la cual está definida como

$$L(\theta, a_0) = \begin{cases} 0 & \text{si } \theta \in \Theta_0, \\ 1 & \text{si } \theta \in \Theta_0^c, \end{cases}$$

y

$$L(\theta, a_1) = \begin{cases} 1 & \text{si } \theta \in \Theta_0, \\ 0 & \text{si } \theta \in \Theta_0^c. \end{cases}$$

Con la función de pérdida $0-1$, si se toma una decisión correcta se incurre en una pérdida de 0; si se toma una decisión incorrecta se incurre en una pérdida de 1. Esta es una situación en la cual ambos tipos de errores tienen la misma penalización. Una función de pérdida más realista es la función de pérdida $0-1$ generalizada, la cual está dada como

$$L(\theta, a_0) = \begin{cases} 0 & \text{si } \theta \in \Theta_0, \\ c_0 & \text{si } \theta \in \Theta_0^c, \end{cases}$$

y

$$L(\theta, a_1) = \begin{cases} c_1 & \text{si } \theta \in \Theta_0, \\ 0 & \text{si } \theta \in \Theta_0^c, \end{cases}$$

donde $c_0 > 0$ y $c_1 > 0$.

Con esta función de pérdida, c_1 es el costo de un error de tipo I (el error de equivocarse al rechazar H_0), y c_0 es el costo de un error de tipo II (el error de equivocarse al aceptar H_0).

Sea k la importancia relativa del error de tipo I con respecto al error de tipo II, es decir,

$$k = \frac{c_1}{c_0}.$$

Si $R_a(\theta)$ es la pérdida esperada final,

$$R_a(\theta) = E(L(\theta, a)) = \int_\Theta L(\theta, a) f(\theta|x) d\theta,$$

entonces la regla de decisión de Bayes es la acción que minimiza $R_a(\theta)$. Pero,

$$\begin{aligned} R_{a_0}(\theta) &= E(L(\theta, a_0) = \int_\Theta L(\theta, a_0) f(\theta|x) d\theta = \int_{\Theta_0^c} c_0 f(\theta|x) d\theta \\ &= c_0 P(\theta \in \Theta_0^c|x), \\ R_{a_1}(\theta) &= E(L(\theta, a_1)) = \int_\Theta L(\theta, a_1) f(\theta|x) d\theta = k c_0 P(\theta \in \Theta_0|x). \end{aligned}$$

Por lo tanto, rechazamos H_0, o sea se lleva a cabo la acción a_1, si y sólo si

$$R_{a_1}(\theta) < R_{a_0}(\theta),$$

si y sólo si

$$k c_0 P(\theta \in \Theta_0|x) < c_0 P(\theta \in \Theta_0^c|x),$$

si y sólo si

$$k P(\theta \in \Theta_0|x) < P(\theta \in \Theta_0^c|x),$$

si y sólo si

$$P(\theta \in \Theta_0^c|x) + P(\theta \in \Theta_0|x) < P(\theta \in \Theta_0^c|x) + \frac{1}{k} P(\theta \in \Theta_0^c|x),$$

es decir, rechazamos H_0 si y sólo si

$$P(\theta \in \Theta_0^c|x) > \frac{k}{k+1}. \tag{6}$$

Si $k = 1$, esto es si ambos tipos de errores tienen la misma importancia, de (6) se tiene que rechazamos H_0 si y sólo si

$$P(\theta \in \Theta_0^c|x) > \frac{1}{2}.$$

Si $k = 9$, rechazamos H_0 si y sólo si

$$P(\theta \in \Theta_0^c|x) > 0{,}9.$$

Análogamente, aceptamos H_0 si y sólo si

$$P(\theta \in \Theta_0|x) > \frac{1}{k+1}. \tag{7}$$

Ejemplo 18. Sean $x_1, x_2, ..., x_n$ los valores observados de las v.a.i.i.d. $X_1, X_2, ..., X_n$ con distribución $N(\theta, \sigma^2)$, donde θ tiene distribución inicial $N(\mu, \tau^2)$, con σ^2, μ y τ^2 conocidos. Queremos probar $H_0 : \theta \leq \theta_0$ contra $H_1 : \theta > \theta_0$. Del ejemplo 6, la densidad final $f(\theta|x)$ es normal con media

$$\frac{n\tau^2\bar{x} + \sigma^2\mu}{n\tau^2 + \sigma^2}$$

y varianza

$$\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}.$$

Si en (7) $k = 1$, es decir si ambos tipos de errores tienen la misma importancia, entonces aceptamos H_0 si y sólo si

$$\frac{1}{2} < P(\theta \in \Theta_0|x) = P(\theta \leq \theta_0|x). \tag{8}$$

Puesto que $f(\theta|x)$ es simétrica con respecto a su media, la desiguadad (8) se cumple si y sólo si

$$\frac{n\tau^2\bar{x} + \sigma^2\mu}{n\tau^2 + \sigma^2} < \theta_0,$$

si y sólo si

$$\bar{x} < \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}.$$

Por lo tanto, H_0 será aceptada como verdadera si y sólo si

$$\bar{x} < \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}.$$

En caso contrario H_1 será aceptada como verdadera.

En particular, si $\mu = \theta_0$, y previo a la evidencia muestral a H_0 y a H_1 se les asigna una probabilidad de $\frac{1}{2}$, entonces H_0 será aceptada como verdadera si y sólo si $\bar{x} < \theta_0$; en caso contrario H_1 es aceptada como verdadera.

4. Comentarios finales

Dada la relevancia que cada vez más están teniendo los métodos estadísticos bayesianos, es conveniente que las personas interesadas en la estadística tengan un conocimiento por lo menos básico de la Estadística Bayesiana. En particular, los planes de estudio de las licenciaturas cuyo propósito es formar profesionistas con una buena preparación estadística, deberían tener al menos una asignatura sobre Inferencia Bayesiana.

Agradecimientos

Gracias a los revisores por sus críticas y sugerencias, las cuales contribuyeron significativamente a mejorar el contenido y la presentación de este trabajo de divulgación.

Referencias

[1] Berger, J. O. (1985). Statistical Decisión Theory and Bayesian Analysis. *Springer-Verlag, New York*.

[2] Bernardo, J. M. and A. F. M. Smith (2004) Bayesian Theory. *John Wiley and Sons, England*.

[3] Casella, G and R. L. Berger (2002). Statistical Inference. *Duxbury, USA*.

[4] Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186(1007), 453-461.

[5] O'Hagan, A. (1994). Kendall's Advanced Theory of Satatistics, Volume 2B, Bayesian Inference. *John Wiley y Sons, USA*.

[6] Robert, C. P. and G. Casella (2004). Monte Carlo Statistical Methods. *Springer-Verlag, USA*.