Estadística Bayesiana

Inferencia Bayesiana

Lizbeth Naranjo Albarrán

Facultad de Ciencias, UNAM

March 7, 2022

Índice

Capítulo 1

Inferencia

1.1 Inferencia Estadística

El objetivo es resolver problemas de Inferencia Estadística como Problemas de Decisión, considerando:
(1) Espacio de acciones potenciales disponibles $\mathcal{A}$.
(2) Espacio parametral, que contiene los posibles estados de la naturaleza Θ.
(3) Espacio de consecuencias $\mathcal{C} = \mathcal{A} \times \Theta$.

Para poder resolver un problema de decisión es necesario cuantificar tanto la incertidumbre sobre Θ como las consecuencias en $\mathcal{C}$.

La única forma 'racional' de cuantificar la incertidumbre es a través de una medida de probabilidad $f(\theta)$, y que las consecuencias deben cuantificarse por medio de una función de pérdida $L(a, \theta)$, o análogamente de una función de utilidad $U(a, \theta)$.

Debe elegirse aquella acción que minimice la pérdida esperada:

$$\begin{aligned} L^*(a) &= \int_\Theta L(a,\theta) f(\theta) \mathrm{d}\theta \\ L^*_X(a) &= \int_\Theta L(a,\theta) f(\theta|\underline{x}) \mathrm{d}\theta \end{aligned}$$

o equivalentemente que maximice la utilidad esperada:

$$\begin{aligned} U^*(a) &= \int_\Theta U(a,\theta) f(\theta) \mathrm{d}\theta \\ U^*_X(a) &= \int_\Theta U(a,\theta) f(\theta|\underline{x}) \mathrm{d}\theta \end{aligned}$$

1.2 Estimación puntual

En estimación puntual

- Espacio de acciones potenciales disponibles es $\mathcal{A} = \Theta$.
- Se tiene la unción de utilidad $U : \Theta \times \Theta \to \mathbb{R}$, o la función de pérdida $L : \Theta \times \Theta \to \mathbb{R}$.

Definición 1.1. La estimación puntual de θ respecto a la función de utilidad $U(\theta, \hat{\theta})$ y a la distribución de probabilidad $f(\theta)$ sobre Θ, es la acción óptima $\hat{\theta}^* \in \mathcal{A} = \Theta$, tal que,

$$\overline{U}(\hat{\theta}^*) \;=\; \max_{\hat{\theta} \in \Theta} \overline{U}(\hat{\theta})$$

donde $\overline{U}(\hat{\theta})$ es la utilidad esperada,

$$\overline{U}(\hat{\theta}) \;=\; \int_{\Theta} U(\hat{\theta}, \theta) f(\theta) \mathrm{d}\theta \;=\; \mathbb{E}_{\theta}\left[U(\hat{\theta}, \theta)\right]$$

1.2.1 Funciones de pérdida y reglas de decisión

A continuación se presentan las funciones de pérdida y utilidad más comúnes.

Ejemplo 1.1 (Utilidad o Pérdida Cuadrática)**.** Suponga que la función de utilidad es la utilidad cuadrática:

$$U(\hat{\theta}, \theta) \;=\; -(\hat{\theta} - \theta)^2$$

o de manera equivalente la pérdida cuadrática:

$$L(\hat{\theta}, \theta) \;=\; (\hat{\theta} - \theta)^2$$

Nos interesa que la utilidad sea máxima, o equivalentemente, que la pérdida sea mínima. Si la diferencia $\hat{\theta} - \theta$ es 'grande' implicará que la utilidad sea 'pequeña' y que la pérdida sea 'grande'.

La utilidad esperada es:

$$\begin{aligned}
\overline{U}(\hat{\theta}) &= \mathbb{E}_\theta\left[U(\hat{\theta},\theta)\right] \\
&= \mathbb{E}_\theta\left[-(\hat{\theta}-\theta)^2\right] \\
&= -\mathbb{E}_\theta\left[\left(\hat{\theta}-\mathbb{E}(\theta)+\mathbb{E}(\theta)-\theta\right)^2\right] \\
&= -\mathbb{E}_\theta\left[\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2+2\left(\hat{\theta}-\mathbb{E}(\theta)\right)(\mathbb{E}(\theta)-\theta)+(\mathbb{E}(\theta)-\theta)^2\right] \\
&= -\mathbb{E}_\theta\left[\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2\right]-2\mathbb{E}_\theta\left[\left(\hat{\theta}-\mathbb{E}(\theta)\right)(\mathbb{E}(\theta)-\theta)\right]-\mathbb{E}_\theta\left[(\mathbb{E}(\theta)-\theta)^2\right] \\
&= -\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2-2\left(\hat{\theta}-\mathbb{E}(\theta)\right)\mathbb{E}_\theta\left[(\mathbb{E}(\theta)-\theta)\right]-\mathbb{V}ar_\theta(\theta) \\
&= -\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2-2\left(\hat{\theta}-\mathbb{E}(\theta)\right)(\mathbb{E}(\theta)-\mathbb{E}_\theta(\theta))-\mathbb{V}ar_\theta(\theta) \\
\overline{U}(\hat{\theta}) &= -\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2-\mathbb{V}ar_\theta(\theta)
\end{aligned}$$

O quivalentemente, la pérdida esperada:

$$\overline{L}(\hat{\theta}) = \left(\hat{\theta}-\mathbb{E}(\theta)\right)^2+\mathbb{V}ar_\theta(\theta)$$

Se busca el estimador puntual de θ, denotado por $\hat{\theta}^*$, tal que se maximice la utilidad esperada, es decir,

$$\overline{U}(\hat{\theta}^*) = \max_{\hat{\theta}\in\Theta}\overline{U}(\hat{\theta})$$

o que se minimice la pérdida esperada,

$$\overline{L}(\hat{\theta}^*) = \min_{\hat{\theta}\in\Theta}\overline{L}(\hat{\theta})$$

Como la varianza de θ es fija, no depende de $\hat{\theta}$, la utilidad esperada se maximizará cuando $-\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2$ sea máximo, o equivalentemente, la pérdida esperada se minimizará cuando $\left(\hat{\theta}-\mathbb{E}(\theta)\right)^2$ sea mínimo, y esto implica que $\hat{\theta}=\mathbb{E}(\theta)$.

Por lo tanto, el estimador puntual es $\hat{\theta}^*=\mathbb{E}(\theta)$. ■

Ejemplo 1.2 (Pérdida $0-1$)**.** Suponga que la función de pérdida $0-1$ es:

$$L(\hat{\theta},\theta) \;=\; \begin{cases} 0 & \text{si } |\hat{\theta}-\theta| \leq \varepsilon \\ 1 & \text{si } |\hat{\theta}-\theta| > \varepsilon \end{cases}$$

para $\varepsilon > 0$.

Análogamente,la función de utilidad es:

$$U(\hat{\theta},\theta) \;=\; \begin{cases} 0 & \text{si } |\hat{\theta}-\theta| \leq \varepsilon \\ -1 & \text{si } |\hat{\theta}-\theta| > \varepsilon \end{cases}$$

La pérdida esperada es:

$$\begin{aligned} \overline{L}(\hat{\theta}) &= \mathbb{E}_\theta\left[L(\hat{\theta},\theta)\right] \\ &= \int L(\hat{\theta},\theta) f(\theta) \mathrm{d}\theta \\ &= 0 \times \mathbb{P}_\theta\left[|\hat{\theta}-\theta| \leq \varepsilon\right] + 1 \times \mathbb{P}_\theta\left[|\hat{\theta}-\theta| > \varepsilon\right] \\ &= \mathbb{P}_\theta\left[|\hat{\theta}-\theta| > \varepsilon\right] \\ &= 1 - \mathbb{P}_\theta\left[|\hat{\theta}-\theta| \leq \varepsilon\right] \end{aligned}$$

Entonces, minimizar la pérdida esperada, equivale a minimizar $\overline{L}(\hat{\theta})$, y equivale a maximizar $\mathbb{P}_\theta\left[|\hat{\theta}-\theta| \leq \varepsilon\right]$.

Definimos un intervalo modal de longitud 2ε igual a $\theta \in [\hat{\theta}-\varepsilon, \hat{\theta}+\varepsilon]$ que tiene la probabilidad más alta de suceder.

Esto implica que $\hat{\theta}$ corresponde al punto medio del intervalo modal. Así, si elegimos valores arbitrariamente pequeños de ε, este produciría:

$$\hat{\theta}^* \;=\; \text{Moda}(\theta)$$

que sería el estimador puntual Bayesiano bajo la función de pérdida $0-1$.

Cuando se tiene una muestra observada $x_1, \ldots, x_n$, el estimador Bayesiano es:

$$\hat{\theta}^* \;=\; \text{Moda}(\theta|\underline{x})$$

Este estimador es un 'estimador máximo verosímil penalizado' en el sentido clásico. ■

Ejemplo 1.3 (Utilidad o Pérdida Valor Absoluto)**.** La función de pérdida valor absoluto es:

$$L(\hat{\theta},\theta) \;=\; |\hat{\theta}-\theta|$$

para $\varepsilon > 0$.

La pérdida esperada es:

$$\begin{aligned}
\overline{L}(\hat{\theta}) &= \mathbb{E}_\theta\left[L(\hat{\theta},\theta)\right] \\
&= \int L(\hat{\theta},\theta) f(\theta)\mathrm{d}\theta \\
&= \int |\hat{\theta}-\theta| f(\theta)\mathrm{d}\theta \\
&= \int_{-\infty}^{\hat{\theta}} \left(\hat{\theta}-\theta\right) f(\theta)\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} \left(\theta-\hat{\theta}\right) f(\theta)\mathrm{d}\theta \\
&= \int_{-\infty}^{\hat{\theta}} \hat{\theta} f(\theta)\mathrm{d}\theta - \int_{-\infty}^{\hat{\theta}} \theta f(\theta)\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} \theta f(\theta)\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} \hat{\theta} f(\theta)\mathrm{d}\theta \\
&= \hat{\theta}\int_{-\infty}^{\hat{\theta}} f(\theta)\mathrm{d}\theta - \int_{-\infty}^{\hat{\theta}} \theta f(\theta)\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} \theta f(\theta)\mathrm{d}\theta - \hat{\theta}\int_{\hat{\theta}}^{\infty} f(\theta)\mathrm{d}\theta
\end{aligned}$$

Se busca encontrar el valor $\hat{\theta}$ tal que minimice la pérdida esperada.

Se calcula la derivada:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}\overline{L}(\hat{\theta}) &= \frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}\left\{\hat{\theta}\int_{-\infty}^{\hat{\theta}} f(\theta)\mathrm{d}\theta - \int_{-\infty}^{\hat{\theta}} \theta f(\theta)\mathrm{d}\theta + \int_{\hat{\theta}}^{\infty} \theta f(\theta)\mathrm{d}\theta - \hat{\theta}\int_{\hat{\theta}}^{\infty} f(\theta)\mathrm{d}\theta\right\} \\
&= 1\times\int_{-\infty}^{\hat{\theta}} f(\theta)\mathrm{d}\theta + \hat{\theta}f(\hat{\theta}) - \hat{\theta}f(\hat{\theta}) - \hat{\theta}f(\hat{\theta}) - 1\times\int_{\hat{\theta}}^{\infty} f(\theta)\mathrm{d}\theta + \hat{\theta}f(\hat{\theta}) \\
&= \int_{-\infty}^{\hat{\theta}} f(\theta)\mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} f(\theta)\mathrm{d}\theta \\
&= F\left(\hat{\theta}\right) - \left[1-F\left(\hat{\theta}\right)\right] \\
&= 2F\left(\hat{\theta}\right) - 1
\end{aligned}$$

Igualando a cero:

$$\begin{aligned}
2F\left(\hat{\theta}\right) - 1 &= 0 \\
F\left(\hat{\theta}\right) &= \frac{1}{2}
\end{aligned}$$

Obteniendo segunda derivada:

$$\frac{\mathrm{d}^2}{\mathrm{d}\hat{\theta}^2}\overline{L}(\hat{\theta}) \;=\; \frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}2F\left(\hat{\theta}\right) \;=\; 2f\left(\hat{\theta}\right) \;>\; 0$$

Por lo tanto, se alcanza un mínimo en $F\left(\hat{\theta}\right) = \frac{1}{2}$. Lo cual implica que $\hat{\theta}$ es tal que $F\left(\hat{\theta}\right) = \frac{1}{2}$, y este valor corresponde a la mediana.

Por lo tanto, el estimador Bayesiano bajo le función de pérdida valor absoluto es

$$\hat{\theta}^* \;=\; \text{Mediana}(\theta)$$

Cuando se tiene una muestra observada $x_1, \ldots, x_n$, el estimador Bayesiano es

$$\hat{\theta}^* \;=\; \text{Mediana}(\theta|\underline{x})$$

■

Ejemplo 1.4 (Pérdida Lineal)**.** La función de pérdida es:

$$L(\hat{\theta}, \theta) \;=\; \begin{cases} g\left(\hat{\theta} - \theta\right) & \text{si } \hat{\theta} > \theta \\ h\left(\theta - \hat{\theta}\right) & \text{si } \hat{\theta} < \theta \end{cases}$$

para $g, h > 0$.

Entonces, el estimador Bayesiano es $\hat{\theta}^*$ el cuantil de orden $\frac{h}{g+h}$ de $f(\theta|\underline{x})$. ■

Ejemplo 1.5 (Binomial)**.** Sea $X \sim \text{Binomial}(n, \theta)$ y $\theta \sim \text{Beta}(a, b)$. Sabemos que la distribución final de θ es $\theta \sim \text{Beta}(x + a, n - x + b)$. Bajo diferences funciones de pérdida, los estimadores puntuales para θ serían:

$$\hat{\theta}^* \;=\; \begin{cases} \text{E}[\theta] = \frac{x+a}{n+a+b} & \text{pérdida cuadrática} \\ \text{Moda}(\theta|x) = \frac{x+a-1}{n+a+b-2} & \text{pérdida 0-1} \\ \text{Mediana}(\theta|x) \approx \frac{x+a-\frac{1}{3}}{n+a+b-\frac{2}{3}} & \text{pérdida valor absoluto} \end{cases}$$

para $a > 1$ y $b > 1$.

Note que bajo el enfoque clásico, el estimador máximo verosímil es $\hat{\theta} = \frac{x}{n}$. ■

Definición 1.2 (Utilidad Cuadrática en $\mathbb{R}^k$)**.** Se tiene el espacio paramétrico $\Theta \cap \mathbb{R}^k$, y la utilidad cuadrática:

$$U(\hat{\theta}, \theta) \;=\; -(\hat{\theta} - \theta)' H (\hat{\theta} - \theta)$$

entonces, la utilidad esperada de $\hat{\theta}$ es

$$\overline{U}(\hat{\theta}) \;=\; -\int_{\Theta} (\hat{\theta} - \theta)' H (\hat{\theta} - \theta) f(\theta) \mathrm{d}\theta$$

Derivando $\overline{U}(\hat{\theta})$ respecto a $\hat{\theta}$, e igualando a cero, resulta que:

$$-2H \int_{\Theta} (\hat{\theta} - \theta) f(\theta) \mathrm{d}\theta \;=\; 0$$

Lo que implica que

$$H\hat{\theta} \;=\; H\mathrm{E}[\theta]$$

Si H^{-1} existe y la esperanza existe, entonces:

$$\hat{\theta}^* \;=\; \mathrm{E}[\theta]$$

■

1.2.2 Estimar una observación futura

Si se desea estimar puntualmente una observación futura de X, denotada por $\hat{x}^*$, el espacio de estados es $X = \mathcal{X}$ (posibles valores que puede tomar X).

La estimación puntual de una observación futura respecto a la función de utilidad $U(\hat{x}, x)$ y a la distribución predictica $f(x)$, es la acción óptima $\hat{x}^* \in \mathcal{X}$ tal que

$$\overline{U}(\hat{x}^*) \;=\; \max_{\hat{x} \in \mathcal{X}} \overline{U}(\hat{x})$$

donde $\overline{U}(\hat{x})$ es la uitlidad esperada definida como:

$$\overline{U}(\hat{x}) \;=\; \int_{\mathcal{X}} U(\hat{x}, x) f(x) \mathrm{d}x \;=\; \mathbb{E}_X \left[U(\hat{x}, x) \right]$$

Analogamente se puede obtener una observación futura usando la distribución predictiva final $f(x|\underline{x})$ dada una muestra observada $\underline{x}$.

1.2.3 Precisión del estimador Bayesiano

Se puede evaluar la precisión del estimador Bayesiano a través de ciertas medidas, por ejemplo, usando el error cuadrático medio posterior:

$$\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2 |\underline{x}\right]$$

donde, si el estimador puntual es $\hat{\theta} = \mathbb{E}[\theta|\underline{x}]$, entonces

$$\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2 |\underline{x}\right] \;\; = \;\; \mathbb{V}ar\left[\theta|\underline{x}\right]$$

1.3 Estimación por regiones

En el enfoque bayesiano la estimación por intervalo para el (los) parámetro(s) desconocidos, θ, de un modelo se basa en la distribución posterior de los mismos $p(\theta|y)$.

Un intervalo del $100(1-\alpha)\%$ de credibilidad es cualquier intervalo (L, U) que satisface que

$$\int_L^U p(\theta|y)d\theta = 1-\alpha$$

Estos intervalos de probabilidad no son únicos. Se puede adoptar por ejemplo un intervalo de colas iguales donde

$$\int_{-\infty}^{L} p(\theta|y)d\theta = \int_U^{\infty} p(\theta|y)d\theta = \alpha/2$$

o uno unilateral donde $L = -\infty$ o $U = \infty$. En los casos donde la distribución posterior del parmetro de interés es unimodal, también podemos adoptar un intervalo de *alta densidad posterior*, (HPD) por sus siglas en inglés, donde $p(L|y) = p(U|y)$. En este caso, este intervalo es el de menor longitud.

Ejemplo: Sea $Y \sim Bin(n, \theta)$ y $\theta \sim Be(g, h)$, entonces

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{1-g}(1-\theta)^{h-1} \\ &= \theta^{g+y}(1-\theta)^{h+n-y} \end{aligned}$$

De esta expresión podemos concluir que las constantes de normalización corresponden a aquellas de una distribución $Be(g+y, h+n-y)$, que es la distribución posterior para θ bajo esta distribución inicial conjugada. Si ahora se considera el escenario con $n = 10$ y $y = 4$ éxitos observados en el experimento de interés. Para una distribución inicial $Be(2, 2)$. Tenemos que la distribución posterior $p(\theta|y)$ es una $Be(6, 8)$, los intervalos del 99% de confiabilidad se muestran el la gráfica.

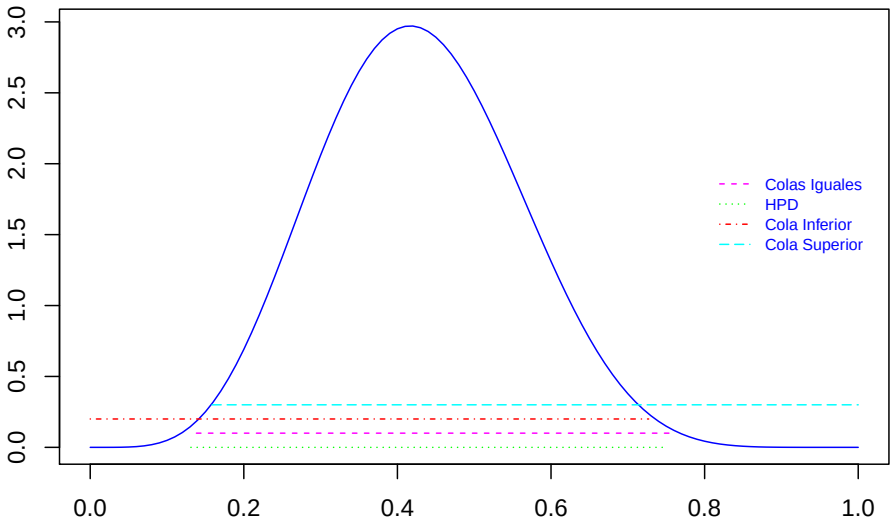

Figura 1.1: Intervalos posteriores para θ.

De manera general, se requiere obtener una región $B \in \Theta$ con probabilidad $1-\alpha$ de obtener el valor correcto, tal que

$$\int_B f(\theta|\underline{x})\mathrm{d}\theta \quad = \quad 1-\alpha$$

donde $0 \leq \alpha \leq 1$. Al conjunto B se le llama de región de probabilidad $1-\alpha$ para θ con respecto a la distribución final $f(\theta|\underline{x})$.

Note que la obtención de B No es única, y se pueden tener conjuntos de soluciones posibles. Por lo que es necesario tener criterios adicionales para elegir una B adecuada. Se puede resolver usando teoría de decisiones.

Estimación por regiones y teoría de decisión

Posibles funciones de utilidad: *Región de menor tamaño posible* $||B||$ pero que tenga el valor correcto de θ:

$$U(B,\theta) \quad = \quad -k||B|| + \mathbb{I}_B(\theta)$$

para $k > 0$.

Utilidad esperada:

$$\begin{aligned}\overline{U}(B) &= \int_{\Theta} U(B,\theta) f(\theta|\underline{x}) \mathrm{d}\theta \\ &= -k||B|| + \int_{B} f(\theta|\underline{x}) \mathrm{d}\theta \\ &= -k||B|| + (1-\alpha)\end{aligned}$$

Por lo tanto, la región óptima es aquella $B^* \in \mathcal{A}$ tal que su tamaño $||B^*||$ sea mínimo. Por tanto, B^* es la región de probabilidad $1-\alpha$ de máxima densidad.

1.4 Pruebas de hipótesis

1.4.1 Contraste de hipótesis y Teoría de Decisión

Contraste de dos hipótesis

Suponga que se desea contrastar las hipótesis

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta \in \Theta_1$$

donde los espacios paramétricos son tales que $\Theta_0 \cap \Theta_1 = \emptyset$ y $\Theta_0 \cup \Theta_1 = \Theta$.

El espacio de acciones potenciales es $\mathcal{A} = \{\Theta_0, \Theta_1\}$. Usando teoría de decisión, la regla de decisión es:

$$\text{Rechazar } H_0 \text{ si } \mathbb{P}\left[\theta \in \Theta_0 | \underline{x}\right] < \mathbb{P}\left[\theta \in \Theta_1 | \underline{x}\right].$$

Es decir, se acepta la hipótesis con mayor porbabilidad, y se rechaza la hipótesis con menor probabilidad.

Suponga que se usa la función de pérdida:

$$L(a_i, \theta) = \begin{cases} k_i & \text{si } \theta \notin \Theta_i \\ 0 & \text{si } \theta \in \Theta_i \end{cases}$$

para $i = 0, 1$, es decir,

$$L(a_0, \theta) = \begin{cases} k_0 & \text{si } \theta \notin \Theta_0 \\ 0 & \text{si } \theta \in \Theta_0 \end{cases} \quad \text{y} \quad L(a_1, \theta) = \begin{cases} k_1 & \text{si } \theta \notin \Theta_1 \\ 0 & \text{si } \theta \in \Theta_1 \end{cases}$$

La pérdida esperada es:

$$\begin{aligned} \overline{L}(a_i) &= k_i \times \mathbb{P}\left(\theta \notin \Theta_i | \underline{x}\right) \\ &= k_i \times \left[1 - \mathbb{P}\left(\theta \in \Theta_i | \underline{x}\right)\right] \end{aligned}$$

La regla de decisión es:

$$\text{Rechazar } H_0 \text{ si } \overline{L}(a_0) > \overline{L}(a_1)$$

Es decir, se rechaza la hipótesis con mayor pérdida esperada.

Esto equivaldría a que, sustituyendo el valor de la pérdida esperada,

$$\text{Rechazar } H_0 \text{ si} \quad k_0 \times [1 - \mathbb{P}(\theta \in \Theta_0 | \underline{x})] \;>\; k_1 \times [1 - \mathbb{P}(\theta \in \Theta_1 | \underline{x})]$$

equivalentemente

$$\frac{k_0}{k_1} \;>\; \frac{[1 - \mathbb{P}(\theta \in \Theta_1 | \underline{x})]}{[1 - \mathbb{P}(\theta \in \Theta_0 | \underline{x})]}$$

como los conjuntos son disjuntos y exahustivos, se tiene que $[1 - \mathbb{P}(\theta \in \Theta_0 | \underline{x})] = \mathbb{P}(\theta \in \Theta_1 | \underline{x})$ y $[1 - \mathbb{P}(\theta \in \Theta_1 | \underline{x})] = \mathbb{P}(\theta \in \Theta_0 | \underline{x})$, entonces,

$$\frac{k_0}{k_1} \;>\; \frac{\mathbb{P}(\theta \in \Theta_0 | \underline{x})}{\mathbb{P}(\theta \in \Theta_1 | \underline{x})}$$

En el caso específico de que $k_0 = k_1$, la regla de decisión es:

$$\text{Rechazar } H_0 \text{ si} \quad \mathbb{P}(\theta \in \Theta_0 | \underline{x}) \;<\; \mathbb{P}(\theta \in \Theta_1 | \underline{x})$$

Si se tienen hipótesis simples, $\Theta_i = \{\theta_i\}$, la regla de decisión es: Rechazar H_0 si

$$k_1 \mathbb{P}(\theta_0 | \underline{x}) \;<\; k_0 \mathbb{P}(\theta_1 | \underline{x})$$

pero, usando las propiedades de la distribución final

$$k_1 f(\underline{x} | \theta_0) f(\theta_0) \;<\; k_0 f(\underline{x} | \theta_1 |) f(\theta_1)$$

equivalentemente,

$$\frac{f(\underline{x} | \theta_0)}{f(\underline{x} | \theta_1 |)} \;<\; \frac{k_0}{k_1} \frac{f(\theta_1)}{f(\theta_0)}$$

Lo cual tiene la misma 'forma' que el Lema de Neyman-Pearson de Estadística Frecuentista.

El contraste de hipótesis en Estadística Bayesiana se basa en comparar las verosimilitudes integradas; en estadística frecuentista se basa en maximizar las verosimilitudes.

Contraste de hipótesis general

En el contexto Bayesiano también se realizan contrastes de hipótesis, y es posible hacer contraste de dos o más hipótesis, por ejemplo un contraste de J hipótesis:

$$H_1 : \theta \in \Theta_1, \quad H_2 : \theta \in \Theta_2, \quad \cdots \quad H_J : \theta \in \Theta_J,$$

donde $\Theta_1, \Theta_1, \ldots, \Theta_J$ denotan una partición del espacio paramétrico Θ.

Medidas de probabilidad prior o posterior:

$$\begin{aligned} \text{Prior} \quad \mathbb{P}[H_j] = \mathbb{P}[\theta \in \Theta_j] &= \int_{\Theta_j} f(\theta) \mathrm{d}\theta \\ \text{Posterior} \quad \mathbb{P}[H_j] = \mathbb{P}[\theta \in \Theta_j] &= \int_{\Theta_j} f(\theta|\underline{x}) \mathrm{d}\theta \end{aligned}$$

Suponga que se tiene la función de utilidad $U(a_j, H_j)$, entonces la utilidad esperada es:

$$\overline{U}(a_j) = \sum_{j=1}^{J} U(a_j, H_j) \mathbb{P}[H_j]$$

Se elige la acción óptima a^* en el conjunto de acciones potenciales $\mathcal{A}$, tal que se maximice la utilidad esperada,

$$\overline{U}(a^*) = \max_{a_j \in \mathcal{A}} \overline{U}(a_j)$$

Contraste de hipótesis para observaciones futuras

En caso de que se tengan hipótesis con relación a observaciones futuras de X, el procedimiento es análogo:

$$H_1 : X \in \mathcal{X}_1, \quad H_2 : X \in \mathcal{X}_2, \quad \cdots \quad H_J : X \in \mathcal{X}_J,$$

donde $\mathcal{X}$ es el espacio muestral, cuya partición es $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_J\}$.

Será necesario usar la distribución predictiva a prior o posterior:

$$\begin{aligned} \text{Prior} \quad \mathbb{P}[H_j] = \mathbb{P}[X \in \mathcal{X}_j] &= \int_{\mathcal{X}_j} f(x^*) \mathrm{d}x^* \\ \text{Posterior} \quad \mathbb{P}[H_j] = \mathbb{P}[X \in \mathcal{X}_j] &= \int_{\mathcal{X}_j} f(x^*|\underline{x}) \mathrm{d}x^* \end{aligned}$$

1.4.2 Contraste de hipótesis y Factor de Bayes

Contraste de hipótesis

En el contexto Bayesiano también se realizan contrastes de hipótesis, y es posible hacer contraste de dos o más hipótesis, por ejemplo un contraste de J hipótesis:

$$H_1 : \theta \in \Theta_1, \quad H_2 : \theta \in \Theta_2, \quad \cdots \quad H_J : \theta \in \Theta_J,$$

donde $\Theta_1, \Theta_1, \ldots, \Theta_J$ denotan una partición del espacio paramétrico Θ. En esta sección sólo se presentará el caso para cuando se requiere hacer contraste de dos hipótesis, sin embargo, la extensión a contrastar más de dos hipótesis será natural.

Suponga que Y proviene de un modelo $f(y|\theta)$ y que se desea evaluar las hipótesis

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_a : \theta \in \Theta_1,$$

donde Θ_0 y Θ_1 constituyen una partición del espacio paramétrico Θ. Recuerde que para hacer inferencia estadística desde una perspectiva Bayesiana se requiere una distribución inicial del parámetro θ. Si se asigna una distribución inicial propia con densidad $p(\theta)$, entonces es posible evaluar las dos hipótesis a priori a través del cociente:

$$\frac{\pi_0}{\pi_1} = \frac{\mathbb{P}(\theta \in \Theta_0)}{\mathbb{P}(\theta \in \Theta_1)} = \frac{\int_{\Theta_0} p(\theta) d\theta}{\int_{\Theta_1} p(\theta) d\theta}.$$

Una vez que se han observado los datos, $Y = y$, la apreciación inicial acerca del parámetro se actualiza a través de la distribución posterior

$$p(\theta|y) \propto L(\theta|y)p(\theta),$$

donde $L(\theta|y)$ denota la función de verosimilitud. Las dos hipótesis planteadas pueden evaluarse ahora considerando la distribución posterior a través del cociente

$$\frac{p_0}{p_1} = \frac{\mathbb{P}(\theta \in \Theta_0|y)}{\mathbb{P}(\theta \in \Theta_1|y)} = \frac{\int_{\Theta_0} p(\theta|y) d\theta}{\int_{\Theta_1} p(\theta|y) d\theta}.$$

En Bayesiana el contraste de hipótesis consistirá en comparar las distribuciones de los parámetros, y de manera intuitiva se elegirá aquella hipótesis con probabilidad mayor. Por ejemplo, si sólo se cuenta con información inicial entonces se podrían comparar las probabilidades π_0 y π_1 y entonces elegir H_0 ó H_a de acuerdo a aquella hipótesis con probabilidad mayor; pero si además se observó una muestra $Y = y$ entonces se podrían comparar las probabilidades p_0 y p_1 y elegir H_0 ó H_a de acuerdo a aquella hipótesis con probabilidad mayor.

Factor de Bayes

Con frecuencia se propone a la estadística conocida como factor de Bayes y definida como

$$FB = \frac{p_0/p_1}{\pi_0/\pi_1},$$

como una medida de la evidencia que proporcionan los datos en favor de la hipótesis nula.

Note que el factor de Bayes compara de manera simultánea las probabilidades obtenidas a partir de las distribuciones iniciales π_0 y π_1 y las obtenidas de las distribuciones posterior p_0 y p_1. Cuando ambas hipótesis son igualmente probables de manera inicial, es decir $\pi_0 = \pi_1$, el factor de Bayes se reduce a sólo comparar las distribuciones finales p_0 y p_1.

Distribuciones Finales

La evaluación de la probabilidad relativa de las hipótesis o los modelos (asociados a esas hipótesis) se puede hacer utilizando la probabilidad posterior del modelo asociado:

$$\begin{aligned} p(H_j|y) &= \frac{p(y|H_j)p(H_j)}{p(y)} \\ &= \frac{p(y|H_j)p(H_j)}{\sum_{k=1}^{J} p(y|H_k)p(H_k)} \\ &\propto p(y|H_j)p(H_j), \end{aligned}$$

donde $p(H_j)$ es la probabilidad inicial del modelo y $j = \{0, a\}$. Es decir, $p(H_0)$ y $p(H_a)$ son las probabilidades iniciales bajo las hipótesis H_0 y H_a, respectivamente. Además:

$$p(y|H_j) = \int p(y|\theta)p(\theta|H_j)d\theta$$

es la verosimilitud marginal bajo el modelo H_j, y $p(\theta|H_j)$ es la distribición inicial para θ cuando H_j es cierta.

Por lo tanto, si se desea contrastar dos hipótesis H_0 y H_a, para evaluar la probabilidad relativa de estas hipótesis, se deberá calcular $p(H_0|y)$ y $p(H_a|y)$, y estas probabilidades se podrían comparar usando el factor de Bayes.

Sin embargo, estas evaluaciones deben hacerse tomando en cuanta los contextos particulares de cada modelo. Los siguientes ejemplos buscan hacer algunas observaciones en este sentido.

Ejemplo 1.6. Sea Y una variable aleatoria de un modelo Normal$(\mu, 1)$, si se plantean las hipótesis: $H_0 : \mu = 0$ contra la alternativa $H_a : \mu \neq 0$; adoptando una distribución inicial conjugada para μ que sea $Normal(0, S^2)$ con $S > 0$.

Note que en este caso el espacio paramétrico es $\Theta = \{\mu; \mu \in \mathbb{R}\} = (-\infty, \infty)$, el conjunto de los reales. Bajo la hipótesis nula H_0 el espacio paramétrico es $\Theta_0 = \{0\}$, y bajo la hipótesis alternativa H_a el espacio paramétrico es $\Theta_1 = \{\mu; \mu \neq 0, \mu \in \mathbb{R}\}$.

Para hacer el contraste de hipótesis en el contexto Bayesiano, se requerirá calcular el factor de Bayes $FB = p(H_0|y)/p(H_a|y)$. Note que también puede obternerse que $p(H_0|y) = 1/\left(1 + \frac{1}{FB}\right)$.

Suponiendo que no se tiene mayor información acerca de cuál de las hipótesis tiene mayor probabilidad inicial, se supondrá que $p(H_0) = p(H_a)$. En esto caso el factor de Bayes se reduciría a calcular $FB = p(y|H_0)/p(y|H_a)$.

Bajo la hipótesis nula, $H_0 : \mu = 0$,

$$p(y|H_0) \;=\; p(y|\mu = 0) = N(y|0, 1).$$

Bajo la hipótesis alternativa $H_a : \mu \neq 0$,

$$\begin{aligned}
p(y|H_a) &= \int p(y|\mu)p(\mu|H_a)d\mu \\
&= \int N(y|\mu, 1)N(\mu|0, S^2)d\mu \\
&= \int \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\} \frac{1}{\sqrt{2\pi S^2}} \exp\left\{-\frac{1}{2S^2}\mu^2\right\} d\mu \\
&= \frac{1}{\sqrt{2\pi(1+S^2)}} \exp\left\{-\frac{1}{2}y^2 + \frac{1}{2}\frac{y^2}{\left(1+\frac{1}{S^2}\right)}\right\} \\
&\times \int \frac{\sqrt{\left(1+\frac{1}{S^2}\right)}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(1+\frac{1}{S^2}\right)\left[\mu^2 - 2\mu\frac{y}{\left(1+\frac{1}{S^2}\right)} + \frac{y^2}{\left(1+\frac{1}{S^2}\right)^2}\right]\right\} d\mu \\
&= \frac{1}{\sqrt{2\pi(1+S^2)}} \exp\left\{-\frac{1}{2(1+S^2)}y^2\right\} \\
&= N(y|0, 1+S^2).
\end{aligned}$$

Por lo tanto, el factor de Bayes resulta en

$$FB = \frac{N(y|0, 1)}{N(y|0, 1+S^2)}.$$

Observe que en este caso el Factor de Bayes crece cuando $S^2 \to \infty$ para cualquier y. ■

Ejemplo 1.7. Considere el experimento de lanzar una moneda n veces, donde las variables aleatorias Y_i son independientes con distribución $Y_i \sim Bernoulli(\theta)$, para $i = 1, \ldots, n$. La hipótesis nula se plantea como $H_0 : \theta = 0.5$ contra la alternativa $H_a : \theta \neq 0.5$, con una distribución inicial $\theta \sim Beta(a, b)$.

Para hacer el contraste de hipótesis se calculará el factor de Bayes $FB = p(H_0|y)/p(H_a|y)$. Otra vez, no se tiene mayor información acerca de cuál de las hipótesis tiene mayor probabilidad inicial, así que se supondrá que $p(H_0) = p(H_a)$, y entonces el factor de Bayes se reduciría a calcular $FB = p(y|H_0)/p(y|H_a)$.

Bajo la hipótesis nula, $H_0 : \theta = 0.5$,

$$p(y|H_0) \;=\; p(y|\theta = 0.5) = 0.5^n.$$

Bajo la hipótesis alternativa, $H_a : \theta \neq 0.5$, se tiene que dadas las observaciones, la verosimilitud es

$$L(\theta|y) = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i},$$

y distribución inicial del parámetro θ es

$$p(\theta|H_a) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},$$

donde $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ es la función beta. Entonces la probabilidad posterior de $Y = y$ dada H_a es

$$\begin{aligned}
p(y|H_a) &= \int p(y|\theta)p(\theta|H_a)d\theta \\
&= \int L(\theta|y)p(\theta|H_a)d\theta \\
&= \int \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}d\theta \\
&= \frac{B\left(\sum_{i=1}^{n} y_i + a\ ,\ n-\sum_{i=1}^{n} y_i + b\right)}{B(a,b)} \int \frac{\theta^{\sum_{i=1}^{n} y_i + a - 1}(1-\theta)^{n-\sum_{i=1}^{n} y_i + b - 1}}{B\left(\sum_{i=1}^{n} y_i + a\ ,\ n-\sum_{i=1}^{n} y_i + b\right)}d\theta \\
&= \frac{B\left(\sum_{i=1}^{n} y_i + a\ ,\ n-\sum_{i=1}^{n} y_i + b\right)}{B(a,b)}.
\end{aligned}$$

Entonces, el factor de Bayes es:

$$\begin{aligned}
FB &= \frac{0.5^n}{\frac{B\left(\sum_{i=1}^{n} y_i + a\ ,\ n-\sum_{i=1}^{n} y_i + b\right)}{B(a,b)}} \\
&= \frac{0.5^n B(a,b)}{B\left(\sum_{i=1}^{n} y_i + a\ ,\ n-\sum_{i=1}^{n} y_i + b\right)}.
\end{aligned}$$

Ahora, consideraremos unos casos particulares, con el objetivo de comparar los resultados obtenidos en contraste de hipótesis bajo los contextos de estadística Bayesiana y del cociente de verosimilitudes generalizado.

Considere dos casos $n = 10$ y $n = 50$ del modelo bajo H_0, con $P(\theta|H_a) \sim Beta(2,2)$ la Figura 1.2 muestra las probabilidades posteriores $P(H_0|y)$.

Observe ahora que la verosimilitud $L(\theta) = p(y|\theta)$ y el cociente de verosimilitudes generalizado para probar las hipótesis planteadas es:

$$\lambda(y) = \frac{\max_{\Theta_0} L(\theta)}{\max_{\Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

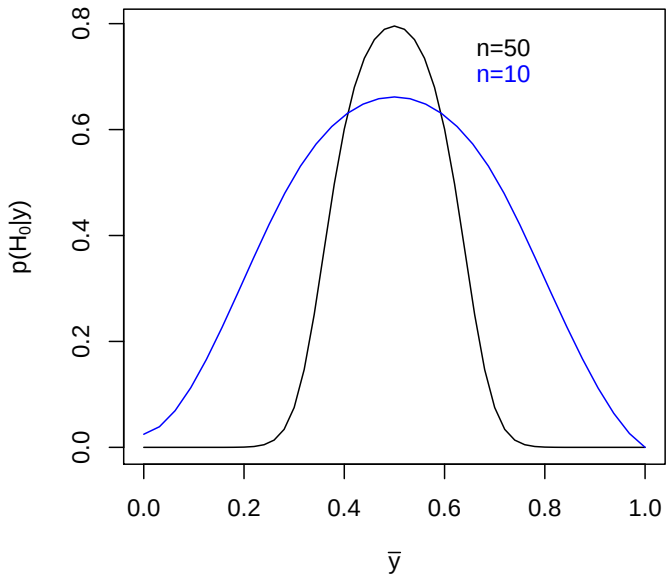

Figura 1.2: Probabilidad posterior del modelo con tamaño de muestra.

donde $\hat{\theta}_0$ y $\hat{\theta}$ son los estimadores máximo verosímiles en los espacios paramétricos correspondientes.

Para este caso en particular

$$\lambda(y) = \frac{0.5^n}{\bar{y}^{n\bar{y}}(1-\bar{y})^{n-n\bar{y}}}.$$

La región de rechazo es de la forma $\{y : \lambda(y) \leq k\}$. Se ha visto que, bajo ciertas condiciones,

$$-2\ln(\lambda(y)) \to \chi^2_{(\nu)}.$$

y si el p-value $< \alpha$, entonces se rechaza H_0 con un nivel de significancia α.

Considere ahora que $n = 10,000$ y $y = 4,900$; el p-valor se aproxima a $\mathbb{P}(\chi^2_1 > (4.000267)) = 0.04549306$ y la probabilidad posterior de H_0

$$P(H_0|y) \approx \frac{1}{1 + 1/7.203413} = 0.8780995.$$

En este caso particular, bajo el cociente de verosimilitudes generalizado el p-valor conlleva a rechazar la hipótesis nula, pero usando la probabilidad posterior elegiríamos la hipótesis nula, y por lo tanto las conclusiones serían opuestas. ■

Estas situaciones que se han ilustrado con los ejemplos pueden observarse ya sea por un efecto del tamaño de muestra, por una hipótesis nula precisa contra una alternativa muy

difusa, o por la probabilidad a priori asignada a las hipótesis. En este sentido, el enfoque Bayesiano penaliza las probabilidades iniciales difusas.

Criterio de Información Bayesiano

Existen muchas otras propuestas para evaluar hipótesis o modelos, aquí se plantea una a manera de ilustración.

Para realizar la comparación de estos modelos es usual que se consideren otras medidas de bondad de ajuste, tal que como en el caso del factor de Bayes, puedan usarse para comparar modelos y que ayuden en la toma de decisiones. Uno de los criterios muy utilizados en estadística Bayesiana es el Criterio de Información Bayesiano (BIC, por su nombre en inglés *Bayesian Information Criterion*).

Definición 1.3. Criterio de Información Bayesiana (BIC). Sea $\mathbf{X}$ una muestra aleatoria de tamañ0 n. Suponga que se tienen dos posibles modelos,

$$f_1(\mathbf{X}|\theta_1,\ldots,\theta_{m_1}) \quad \text{y} \quad f_2(\mathbf{X}|\theta_1,\ldots,\theta_{m_2}),$$

cada una parametrizado por m_1 y m_2 parámetros, $\theta_1,\ldots,\theta_{m_1}$ y $\theta_1,\ldots,\theta_{m_2}$, respectivamente, los cuales pueden tener elementos en común. El BIC se define como:

$$BIC = -2\ln\left(\frac{L_1(\theta_1,\ldots,\theta_{m_1}|\mathbf{X})}{L_2(\theta_1,\ldots,\theta_{m_2}|\mathbf{X})}\right) + (m_1 - m_2)\ln(n).$$

Ejemplo 1.8. Suponga que, dada una muestra aleatoria de tamaño n, se tienen las siguientes hipótesis para un fenómeno de interés:

$$H_0 : X_i \sim Gama(\alpha,\beta) \quad \text{vs.} \quad H_a : X_i \sim Exp(\theta).$$

Bajo H_0 se tienen $m_0 = 2$ parámetros, y la verosimilitud es:

$$\begin{aligned} L(\alpha,\beta|\mathbf{X}) &= \prod_{i=1}^{n} Gama(x_i|\alpha,\beta) \\ &= \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i\beta} \\ &= \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} e^{-\sum_{i=1}^{n} x_i\beta}. \end{aligned}$$

Bajo H_a se tiene $m_1 = 1$ parámetro, y la verosimilitud es:

$$\begin{aligned} L(\theta|\mathbf{X}) &= \prod_{i=1}^{n} Exp(x_i|\theta) \\ &= \prod_{i=1}^{n} \theta e^{-x_i\theta} \\ &= \theta^n e^{-\sum_{i=1}^{n} x_i\theta}. \end{aligned}$$

Considere el caso particular $n = 100$, $\bar{x} = 0.479$, $\prod_{i=1}^{100} x_i = 2.898146e-44$ y las hipótesis: $H_0 : X_i \sim Gama(2, 4)$ vs. $H_a : X_i \sim Exp(0.5)$.

El BIC se calcula como:

$$\begin{aligned} BIC &= -2\ln\left(\frac{\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} e^{-\sum_{i=1}^{n} x_i\beta}}{\theta^n e^{-\sum_{i=1}^{n} x_i\theta}}\right) + (m_0 - m_1)\ln(n). \\ &= -2\ln\left(\frac{4^{200} \times (2.898146e-44) \times \exp(-191.6568)}{(0.5^{100})\exp(-23.95711)}\right) + \ln(100) \\ &= -152.6433. \end{aligned}$$

■

En algunos contextos se sugiere que si el $BIC \leq 2$ la evidencia que favorece al primer modelo es muy débil, mientras que si el $BIC > 10$ la evidencia de ello es contundente.