Revista Colombiana de Estadística
No.31 y No. 32

ANALISIS DE BAYES EMPIRICO MEDIANTE UN EJEMPLO

VÍCTOR HUGO PRIETO, CONSTANZA QUINTERO, ISMAEL RODRÍGUEZ

Profesores Asociados
Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia
Santafé de Bogotá, Colombia

§ 1. INTRODUCCIÓN

El objeto de este artículo es ilustrar en forma elemental el desarrollo de la metodología de Bayes Empírico, sin pretender comparar las estimaciones que aquí se obtengan con las que se pueden obtener a través del método clásico o del método de Bayes. En general, el método de Bayes se plantea como una metodología alterna al método clásico. Dicho método se fundamenta en el conocimiento de una distribución sobre el parámetro θ, $\pi(\theta)$, denominada distribución a priori. Tal distribución se supone completamente conocida, no tiene parámetros desconocidos e involucra probabilidades subjetivas.

El método de Bayes Empírico depende también de la existencia de una distribución a priori $\pi(\theta)$ a la cual se le puede dar una interpretación frecuencial y se puede estimar usando observaciones apropiadas. Así el método de Bayes Empírico puede ser esencialmente no Bayesiano, en el sentido de no involucrar probabilidades subjetivas.

El Bayesiano empírico está de acuerdo con el método de Bayes, pero no especifica valores de los parámetros de la distribución a priori $\pi(\theta)$, estima tales parámetros a

partir de datos auxiliares (pasados o presentes) para la construción de las reglas de Bayes.

En términos de distribuciones se trata de determinar una distribución a posteriori

$$f(\theta|\boldsymbol{x})\alpha f(\boldsymbol{x}|\theta)\pi(\theta) \tag{1}$$

donde $f(\boldsymbol{x} \mid \theta)$ es la función de verosimilitud evaluada en θ; $\pi(\theta)$ es la distribución a priori, en el caso de Bayes Empírico los parámetros desconocidos de $\pi(\theta)$ que se estimarán para poder trabajar la ecuación (1) de la misma manera que el modelo de Bayes.

Se puede enfatizar entonces que la diferencia fundamental entre el método de Bayes y el método de Bayes Empírico radica en que el método de Bayes supone el conocimiento total de la distribución a priori $\pi(\theta)$ y el método de Bayes Empírico estima la distribución a priori $\pi(\theta)$.

Para ilustrar este método se utilizan los datos del experimento de promedio de colesterol en ratas para *Actividad hipocolesterolémica en ratas del fruto solamun melongena* (*Berenjena*) en el cual se utilizaron 24 ratas macho Fisher 344 en un diseño de bloques al azar. Los datos de colesterol se redistribuyeron en 4 grupos de seis ratas cada uno. y se desea estimar puntual y por intervalos el promedio de colesterol.

El grupo uno corresponde a las ratas a las que se les administró la dieta normal durante todo el tiempo de estudio.

El grupo dos corresponde a las ratas a las que se les administró la dieta grasa durante todo el tiempo de estudio.

El grupo tres corresponde a las ratas a las que se les administró la dieta grasa

durante los primeros 45 días al cabo de los cuales se les empezó a administrar el medicamento PRAVACOL manteniendo la dieta grasa.

El grupo cuatro corresponde a las ratas a las que se les administró la dieta grasa durante los primeros 45 días al cabo de los cuales se les empezó a administrar el extracto seco de Berenjena, manteniendo la dieta grasa.

Para cada grupo se consideró la medida de colesterol al inicio del experimento, a los 45 días y a los 90 días.

§ 2. METODOLOGÍA

Se consideran las variables aleatorias Y_i, $i = 1, 2, 3, 4$ correspondientes al promedio de las observaciones en cada grupo de ratas.

Se supone

$$Y_i \sim n(\theta_i, \sigma^2), \ i = 1, 2, 3, 4 \tag{1.1}$$

σ^2 conocida y θ_i el verdadero promedio de colesterol en el grupo i.

Se supone además que

$$\Theta_i \sim n(\mu, \tau^2), \ i = 1, 2, 3, 4, \tag{1.2}$$

es decir, se supone una información a priori sobre el parámetro Θ_i, a través de la distribución a priori sobre Θ_i con μ y τ^2 desconocidos. Es esta una diferencia con la metodología de Bayes en la cual μ y τ^2 se suponen conocidos.

La distribución a posteriori de $\Theta \mid Y_i$ es normal de media

$$\mu(Y_i) = \left[\frac{\sigma^2}{\sigma^2 + \tau^2}\right]\mu + \left[\frac{\tau^2}{\tau^2 + \sigma^2}\right]Y_i \tag{1.3}$$

y varianza

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

donde

$$f(\theta_i \mid y_i) \qquad \text{resulta de} \qquad \frac{\Pi(\theta_i)f(y_i \mid \theta_i)}{\int \Pi(\theta_i)f(y_i \mid \theta_i)d\theta_i}$$

2.1 Estimación puntual

Se conoce que el estimador a posteriori de Bayes para θ_i es la media de la distribución a posteriori, es decir,

$$\hat{\theta}_{iB} = \left[\frac{\sigma^2}{\sigma^2+\tau^2}\right]\mu + \left[\frac{\tau^2}{\sigma^2+\tau^2}\right]Y_i \tag{1.4}$$

Al asumir μ y τ^2 desconocidos surge el modelo de Bayes empírico, por tanto es necesario estimar μ y τ^2 a partir de los datos actuales o pasados. G. Casella [1992] propone un estimador insesgado para μ y un estimador insesgado para $\frac{\sigma^2}{\sigma^2+\tau^2}$ a saber:

$$\hat{\mu} = \overline{Y}; \qquad \widehat{\frac{\sigma^2}{\sigma^2+\tau^2}} = \frac{(p-3)\sigma^2}{\sum(Y_i-\bar{Y})^2} \tag{1.5}$$

Resulta entonces que el estimador de Bayes empírico de θ_i , $\delta_i^E(Y)$ es dado por :

$$\delta_i^E = \hat{\Theta}_i = \left[\frac{(p-3)\sigma^2}{\sum(Y_i-\bar{Y})^2}\right]\overline{Y} + \left[1-\frac{(p-3)\sigma^2}{\sum(Y_i-\bar{Y})^2}\right]Y_i \tag{1.6}$$

con V_i^E dado por

$$\sigma^2\left(1-\frac{p-1}{p}\hat{B}\right) + \frac{2}{p-3}\hat{B}^2(Y_i-\bar{Y})^2 \ , \ \text{con } \hat{B} = \frac{(p-3)\sigma^2}{\sum(Y_i-\bar{Y})^2}$$

Por lo tanto se tienen las siguientes estimaciones para θ_i con la varianza respectiva.

Estimaciones del parámetro y la varianza en la primera fecha

Parámetro estimado	Varianza estimada
$\hat{\Theta}_1 = \delta_1^E = 45.049013$	$V_1^E = 1.626821$
$\hat{\Theta}_2 = \delta_2^E = 44.140370$	$V_2^E = 1.374653$
$\hat{\Theta}_3 = \delta_3^E = 43.950524$	$V_3^E = 1.360434$
$\hat{\Theta}_4 = \delta_4^E = 42.229771$	$V_4^E = 1.837984$

En estos cálculos se estimó σ^2 a partir de la expresión

$$\hat{\sigma}^2 = \frac{1}{n(n-1)p}\sum_{i=1}^{p}\sum_{j=1}^{n}(y_i^j - \bar{y}_i)^2 = 1.64582 \qquad \hat{B} = \frac{(p-3)\hat{\sigma}^2}{\sum(Y_i - \bar{Y})^2} = 0.232947$$

dada en Berger (1985).

Estimaciones del parámetro y la varianza en la segunda fecha

Parámetro estimado	Varianza estimada
$\hat{\Theta}_1 = \delta_1^E = 49.680323$	$V_1^E = 2.606246$
$\hat{\Theta}_2 = \delta_2^E = 49.718864$	$V_2^E = 2.963490$
$\hat{\Theta}_3 = \delta_3^E = 53.192743$	$V_3^E = 3.184456$
$\hat{\Theta}_4 = \delta_4^E = 51.473841$	$V_4^E = 2.482155$

$$\hat{\sigma}^2 = 3.065521 \qquad \hat{B} = 0.334606$$

Estimaciones del parámetro y la varianza en la tercera fecha

Parámetro estimado	Varianza estimada
$\hat{\Theta}_1 = \delta_1^E = 50.025352$	$V_1^E = 2.046888$
$\hat{\Theta}_2 = \delta_2^E = 43.140176$	$V_2^E = 1.93304045$
$\hat{\Theta}_3 = \delta_3^E = 45.113633$	$V_3^E = 1.906027$
$\hat{\Theta}_4 = \delta_4^E = 42.677509$	$V_4^E = 1.946309$
$\hat{\sigma}^2 = 1.984134$	$\hat{B} = 0.0525526$

2.2. Regiones de confiabilidad

Para estimar θ_i a partir de una región de credibilidad del $100(1-\alpha)\%$ se puede proceder de acuerdo al método sugerido por Maritz, con una ligera modificación. En efecto, se considera, como se ha supuesto en la estimación puntual, Y_i con distribución $f(Y_i \mid \theta_i)$, $n(\theta_i, \sigma^2)$, con σ^2 conocida, Θ_i con distribución a priori $\pi(\Theta_i)$, $n(\mu, \tau^2)$, μ, τ^2 desconocidos. Como en el caso de estimación puntual se considera la distribución a posteriori $f(\Theta_i \mid Y_i) \sim n(\mu(Y_i), \tau^{*2})$ con

$$\mu(Y_i) = (\tau^2 + \sigma^2)^{-1}(\tau^2 Y_i + \sigma^2 \mu)$$

$$\tau^{*2} = \left(\frac{1}{\tau^2 + \sigma^2}\right)\tau^2\sigma^2$$

así,

$$\mu(Y_i) = \frac{\tau^2}{\tau^2 + \sigma^2} Y_i + \frac{\sigma^2}{\tau^2 + \sigma^2}\mu$$

Los límites del intervalo se pueden determinar a partir de la expresión

$$\hat{\lambda}^*(Y_i, \mu, \tau^2, \alpha) = \delta_i^E \pm Z(\alpha/2)\sqrt{V_i E}$$

Con los datos se tiene la región de confiabilidad del 95% para cada $\Theta_i \quad i = 1, 2, 3, 4$ en cada fecha

INTERVALO DE CONFIANZA

$\hat{\theta}_i$	Primera fecha	segunda fecha	Tercera fecha
$\hat{\theta}_1$	(42.549094, 47.548932)	(46.516124, 52.844522)	(47.22119, 52.829512)
$\hat{\theta}_2$	(41.842356, 46.438384)	(46.344765, 53.092963)	(40.415113, 45.865239)
$\hat{\theta}_3$	(41.664426, 46.236622)	(49.695114, 56.690372)	(42.407678, 47.819588)
$\hat{\theta}_4$	(39.572555, 44.886987)	(48.385889, 54.561793)	(39.94311, 45.411908)

COMENTARIOS

1. Entre la primera y la segunda fecha aumenta el promedio de colesterol en todos los grupos, inclusive en el grupo I que no tuvo dieta grasa. Además este aumento es muy similar en los grupos I y II, y en los grupos III y IV.

2. En el grupo I no hay cambio entre la segunda y la tercera fecha; mientras que en los tres grupos restantes se nota un cambio similar, concluyéndose además que aparentemente no hay diferencia entre el efecto de Pravacol y Berenjena.

3. Vale la pena observar que en el grupo II, en el cual se administró dieta grasa durante todo el tiempo del estudio, disminuye el promedio de colesterol entre la segunda y la tercera fecha.

BIBLIOGRAFÍA

Berger J.O.,(1985), *Statistical Decision Theory and Bayesian Analysis Second edition*, Springer Verlang.

Casella G.,(1992), *An introduction to empirical Bayes Data Analysis* **39 N^0 2**, The American Statistician, 83–87.

Morris C., (1983), *Parametric Empirical Bayes Inference: Theory and Applications, (with discussion)* **78, N^0 381**, Journal of the American Statistical Association, 47–65.

Medina A. Mejía G., (1993), *Actividad Hipocolesterolémica en ratas del fruto de "Solamun Melongena" (Berenjena) (Trabajo de Grado)*..