PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

INTERROGACIÓN 2 - PARTE II
Métodos Bayesianos - EYP2807 - EYP280I

Profesora	:	Ana María Araneda
Ayudante	:	Josefa Silva
Fecha publicación	:	25 de mayo de 2023, 22:00 hrs.
Fecha entrega	:	31 de mayo de 2023, 12:00 hrs.

La nota de la Interrogación 2 corresponde a:

$$\frac{1}{2}\,(\text{Nota Parte I } + \text{ Nota Parte II})$$

1. **[20 %]** En el verano de 2008, en Estados Unidos, se produjeron grandes inundaciones en la región medio-oeste del país, en particular en el poblado de Coralville, en el estado de Iowa. Los esfuerzos por recuperar la ciudad tuvieros escasos efectos, por lo que, un año después, se propuso un impuesto adicional, de 1 centavo por dólar, para costear medidas preventivas frente a potenciales inundaciones futuras. Poco antes de las votaciones, una muestra obtenida de manera telefónica entre los votantes registrados entregó un 40 % a favor de la aplicación del nuevo impuesto, entre las 327 personas que respondieron a la encuesta. En base a estos datos, se desea demostrar que la proporción de personas que aprobará la medida es menor a 50 %.

   a) Establezca condiciones para que las observaciones obtenidas sean representativas de lo que ocurrirá realmente en las votaciones.

   b) Plantee las hipótesis necesarias desde el punto de vista clásico y obtenga el valor-p. ¿Qué puede concluir?

   c) Proponga un modelo Bayesiano para las observaciones, que le permita evaluar la conjetura. Asuma que no tiene conocimiento previo sobre la proporción de interés.

   d) Utilice las observaciones disponibles para evaluar la conjetura desde el punto de vista Bayesiano, y compare con el resultado obtenido desde el punto de vista clásico. Utilice una función de pérdida 0-1 generalizada, justificando brevemente los valores de c_1 y c_2 propuestos. Utilice la decisión óptima bajo dichos valores.

2. **[15 %]** Considere un problema de test de hipótesis para decidir entre dos modelos candidatos, M_0 y M_1, como generadores de los datos, $\boldsymbol{y} = (y_1, \dots, y_n)$, y el Factor de Bayes asociado:

$$\text{BF} = \frac{p(\boldsymbol{y}|M_0)}{p(\boldsymbol{y}|M_1)}.$$

   Se han propuesto en la literatura una serie de aproximaciones numéricas para obtener Factores de Bayes. En particular, Tierney y Kadane (1986) propusieron aproximar $p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ por una densidad Normal de media $\tilde{\boldsymbol{\theta}}$ y matriz de covarianzas $\tilde{\Sigma} = (-D^2 l(\tilde{\boldsymbol{\theta}}))^{-1}$, donde $\tilde{\boldsymbol{\theta}}$ corresponde a la moda de la distribución a posteriori de $\boldsymbol{\theta}$ y $D^2 l(\tilde{\boldsymbol{\theta}})$ corresponde a la matriz de segundas derivadas de:

$$l(\boldsymbol{\theta}) = \log \{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\},$$

evaluada en $\tilde{\boldsymbol{\theta}}$. Con esto, Tierney y Kadane muestran que $p(\boldsymbol{y}|M_i)$ puede ser aproximado como:

$$p(\boldsymbol{y}|M_i) \approx (2\pi)^{d/2}|\tilde{\Sigma}|^{1/2} p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}})p(\tilde{\boldsymbol{\theta}}), \tag{1}$$

donde d corresponde a la dimensión de $\boldsymbol{\theta}$.

Considere ahora un problema de regresión lineal por el origen, donde se desea testear la significancia de la pendiente. En ese caso, el modelo M_0 corresponde a:

$$y_1, \ldots, y_n|(\theta = 0) \overset{iid}{\sim} \text{Normal}(0, \sigma^2),$$

con $\sigma > 0$ conocido, y el modelo M_1 corresponde a:

$$y_1, \ldots, y_n|\theta \sim \text{Normal}(\theta x_i, \sigma^2),$$

donde $x_1, \ldots, x_n$ corresponden a valores de un predictor dado, $\sigma > 0$ conocido, $y_1, \ldots, y_n$ independientes, y distribución a priori Normal$(0, \tau_\theta^2)$ para θ.

Utilice el resultado de Tierney y Kadane en (1) para encontrar una expresión para el Factor de Bayes en este problema.

3. **[15 %]** En un estudio se desea predecir la altura de los árboles de cierta familia, en términos de la circunferencia de su tronco. Para esto, la bióloga a cargo decide recolectar mediciones de estas variables, en una muestra de árboles de la familia de interés. Los datos se encuentran en el archivo `arboles.csv`. Se plantea ajustar una regresión por el origen.

   a) Obtenga un gráfico de dispersión de los datos, ajuste y represente la recta de mínimos cuadrados ordinarios (sin intercepto), y realice el test-t para determinar la significancia de la circunferencia del tronco como predictor de la altura de un árbol. Concluya (desde el punto de vista clásico).

   b) Analice el problema desde el punto de vista Bayesiano. Para ello, utilice el resultado para el Factor de Bayes que encontró en el problema anterior, considerando $\sigma = 250$ y $\tau_\theta = 100$, ambas cantidades expresadas en pulgadas. Compare con el resultado obtenido bajo el enfoque clásico y concluya en términos del problema.