Introducción a la estadística Bayesiana con aplicaciones de estimación en áreas pequeñas usando software STAN

Ignacio Alvarez-Castro   Juan José Goyeneche

Instituto de Estadística, Facultad de Ciencias Económicas y Administración, UdelaR.

XV Congreso Latinoamericano de Sociedades de Estadística
9 al 13 de Octubre 2023
Santiago de Cali, Colombia

Probabilidad previa

Objetivo: Realizar afirmaciones probabilísticas sobre θ dados los datos observados y. Es decir obtener la distribución posterior $p(\theta|y)$.

Modelo: El primer paso para lograr nuestro objetivo es establecer un *modelo completo* de probabilidad, es decir, especificar la distribución conjunta $p(y, \theta)$.

La distribución conjunta $p(\theta, y)$ se construye habitualmente como:

- $p(y|\theta)$, un modelo para los datos (Ejemplo: $y_i|\theta \sim Ber(\theta)$)
- $p(\theta)$, una distribución previa (Ejemplo $\theta \sim Unif(0, 1)$)

La distribución previa, $p(\theta)$, expresa la incertidumbre (o la información) existente sobre θ antes de considerar los datos.

Probabilidad de un evento raro

Queremos estimar la prevalencia de una enfermedad *rara* en una ciudad. Para esto se obtiene una muestra de 25 personas para evaluar si estan infectadas o no.

Conocemos que en estudios anteriores, en otras ciudades la prevalencia de esta enfermedad varía entre $(0.05, 0.2)$ y en promedio es 0.10. En base a esto elegimos

Datos: Total de personas infectadas en la muestra	$y \in \{0, 1, \ldots, 25\}$
Parámetro: prevalencia en la población	$\theta \in (0, 1)$

Planteamos el modelo:

$$y \sim binomial(25, \theta) \quad \theta \sim Beta(2, 20)$$

Observamos: $y_0 = 1$

Plan

Previas conjugadas

Conseremos un modelo para los parámetros $p(\theta)$ y un modelo para los datos condicional en los parámetros $p(y|\theta)$, y la la posterior de interés que se obtiene haciendo

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

Si $p(\theta|y)$ pertenece a **la misma familia** de modelos que la previa, decimos que $p(\theta)$ es conjugada para $p(y|\theta)$.

- conveniencia matemática
- interpretabilidad de sus parámetros
- Problema: pueden ser restrictivas para modelos complicados

Problema:

- $y = (y_1, \ldots, y_n)$, con $y_1 \sim N(\mu, \sigma^2)$
- $\mu \sim N(m_0, t_0^2)$
- hacer inferencia para μ asumiendo σ^2 conocido.

Tenemos $p(\mu) \propto e^{-(1/2t_0^2)(\mu^2 - 2m_0\mu)}$, posterior ?

Problema:

- $y = (y_1, \ldots, y_n)$, con $y_1 \sim N(\mu, \sigma^2)$
- $\mu \sim N(m_0, t_0^2)$
- hacer inferencia para μ asumiendo σ^2 conocido.

Tenemos $p(\mu) \propto e^{-(1/2t_0^2)(\mu^2 - 2m_0\mu)}$, posterior ?

$$
\begin{aligned}
p(\mu|y) &\propto \exp -\frac{1}{2\sigma^2}(-2n\mu\bar{y} + n\mu^2) \times \exp^{-\frac{1}{2t_0^2}(\mu^2 - 2m_0\mu)} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{t_0^2}\right)\mu^2 - 2\left(\frac{n\bar{y}}{\sigma^2} + \frac{m_0}{t_0^2}\right)\mu\right]\right\}
\end{aligned}
$$

$$\mu|y \sim N(m_1, t_1^2)$$

Llamamos precisión al inverso de la varianza. Ejemplo: precisión previa de μ es $1/t_0^2$.

- Precisión posterior es la suma de precisión previa más la precisión de la media muestral
$$\frac{1}{t_1^2} = \frac{n}{\sigma^2} + \frac{1}{t_0^2}$$
- Media posterior es un promedio ponderado entre la media previa y la media muestral.
$$m_1 = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{t_0^2}} \bar{y} + \frac{\frac{1}{t_0^2}}{\frac{n}{\sigma^2} + \frac{1}{t_0^2}} m_0$$
- Cuando el tamaño muestral crece, la media posterior tiende a la media muestral

Una familia muy importante de modelos es la *familia exponencial de un parámetro*:

$$p(y_i|\phi) = f(y_i)g(\phi)e^{\phi t(y_i)}$$

Con $y = (y_1, \ldots, y_n)$ iid, y $p(y_i|\phi)$ en la familia exponencial:

$$\begin{aligned} p(y|\phi) &= \prod_i f(y_i)g(\phi)e^{\phi t(y_i)} \\ &\propto g(\phi)^n e^{\phi \sum_i t(y_i)} \end{aligned}$$

Se conocen muchas propiedades para esta familia de modelos. Por ejemplo, $t(y) = \sum_i u(y_i)$ es un estadístico suficiente para ϕ.

Expresemos la distribución *Binomial*(n, θ) como un ejemplo de *familia exponencial de 1 parámetro*.

Comenzando con la expresión usual $p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$, podemos reformularla como

$$p(y|\phi) = \binom{n}{y}\left(\frac{1}{1+e^{\phi}}\right)^n e^{\phi y}$$

con la transformación

$$\phi = \log(\frac{\theta}{1-\theta})$$

Previas conjugadas y familia exponencial

Se puede mostrar que par todos los modelos de la *familia exponenial de 1 parámetro* existe una distribución previa conjugada para ϕ.

En particular, usando como previa

$$p(\phi) \propto g(\phi)^{\eta} e^{\eta\tau\phi}$$

la posterior tiene la forma

$$p(\phi|y) \propto g(\phi)^{n+\eta} e^{(t(y)+\eta\tau)\phi}$$

que tiene la misma forma funcional que $p(\phi)$.

Como queda la previa conjugada de la familia exponencial en el caso Binomial.

$$p(y|\phi) = \binom{n}{y} \left(\frac{1}{1 + e^{\phi}} \right)^{n} e^{\phi y}$$

por lo que

$$p(\phi) \propto \left(\frac{1}{1 + e^{\phi}} \right)^{\eta} e^{\eta \tau \phi}$$

Como queda la previa conjugada de la familia exponencial en el caso Binomial.

$$p(y|\phi) = \binom{n}{y} \left(\frac{1}{1+e^{\phi}} \right)^n e^{\phi y}$$

por lo que

$$p(\phi) \propto \left(\frac{1}{1+e^{\phi}} \right)^{\eta} e^{\eta \tau \phi}$$

expresada en términos del parámetro original θ obtenemos:

$$p(\theta) \propto \theta^{\eta\tau - 1}(1-\theta)^{\eta(1-\tau)-1}$$

es decir una $Beta(\eta\tau, \eta(1-\tau))$

Plan

Previas e información

- Informativas. Incorporan conocimiento de expertos, resultados de investigaciones, resultados ya conocidos.

- Automáticas (No-Informativas). Obtener una previa a partir del modelo para los datos, $p(y|\theta)$, sin incorporar otra información.

- Débilmente informativas.

Previas débilmente informativas

Una previa $p(\theta)$ es *débilmente informativa* cuando:

- descarta valores de θ que no tienen sentido
- es flexible para tener muy bajo efecto en $p(\theta|y)$

La idea es utilizar previas que no contengan información sobre θ excepto para estabilizar la inferencia.

Previas débilmente informativas

Modelo Normal, $y \overset{iid}{\sim} N(\mu, 10)$

Modelo Normal, $y \overset{iid}{\sim} N(\mu, 10)$

- $\mu \sim N(0, 1)$ puede ser restrictiva
- $p(\mu) \propto 1$ asigna masa de probabilidad en regiones irrelevantes
- $\mu \sim N(0, 100)$, $\mu \sim t_3(0, 1)$ pueden ser una alternativa debilmente informativas