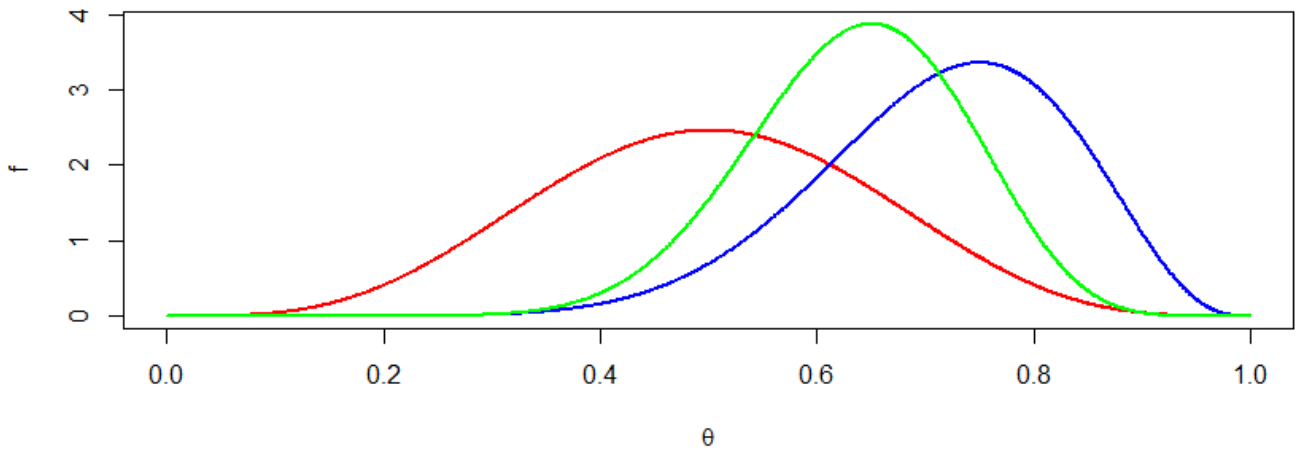
# Distribuciones conjugadas: cuando la inferencia bayesiana es fácil




Mike Wiper

Departamento de Estadística

Universidad Carlos III de Madrid

# Objetivo

Examinar las situaciones cuando se pueden hacer los cálculos necesarias para la inferencia bayesiana de forma sencilla.

# Tirando monedas otra vez

En problemas donde tiramos monedas (binomial, geométrica, binomial negativa, ...) la función de verosimilitud es:

$$l(\theta|\text{datos}) = c\theta^{\text{cruces}}(1-\theta)^{\text{caras}}$$

donde $c$ is un constante determinado por el diseño del experimento.

# Tirando monedas otra vez

En problemas donde tiramos monedas (binomial, geométrica, binomial negativa, ...) la función de verosimilitud es:

$$l(\theta|\text{datos}) = c\theta^{\text{cruces}}(1-\theta)^{\text{caras}}$$

donde $c$ is un constante determinado por el diseño del experimento.

Si empleamos una distribución a priori Beta$(\alpha, \beta)$ para $\theta$:

$$f(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

# Tirando monedas otra vez

En problemas donde tiramos monedas (binomial, geométrica, binomial negativa, ...) la función de verosimilitud es:

$$l(\theta|\text{datos}) = c\theta^{\text{cruces}}(1-\theta)^{\text{caras}}$$

donde $c$ is un constante determinado por el diseño del experimento.

Si empleamos una distribución a priori Beta$(\alpha, \beta)$ para $\theta$:

$$f(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

entonces la distribución a posteriori es también una beta:

$$f(\theta|\text{datos}) \quad \propto \quad \theta^{\alpha+\text{cruces}-1}(1-\theta)^{\beta+\text{caras}-1}$$

# Tirando monedas otra vez

En problemas donde tiramos monedas (binomial, geométrica, binomial negativa, ...) la función de verosimilitud es:

$$l(\theta|\text{datos}) = c\theta^{\text{cruces}}(1-\theta)^{\text{caras}}$$

donde $c$ is un constante determinado por el diseño del experimento.

Si empleamos una distribución a priori Beta$(\alpha, \beta)$ para $\theta$:

$$f(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

entonces la distribución a posteriori es también una beta:

$$\begin{aligned} f(\theta|\text{datos}) &\propto \theta^{\alpha+\text{cruces}-1}(1-\theta)^{\beta+\text{caras}-1} \\ \theta|\text{datos} &\sim \text{Beta}(\alpha+\text{cruces}, \beta+\text{caras}) \end{aligned}$$

La distribución a priori beta es **conjugada** al distribución de muestreo binomial (o geométrica o binomial negativa).

# Predicción: la distribución beta-binomial

Dada una distribución Beta$(\alpha, \beta)$ para $\theta$, si queremos predicir el número de cruces, $X$, en $n$ tiradas más, tenemos:

$$\begin{aligned} P(X = x) &= \int_0^1 P(X = x|\theta) f(\theta)\, d\theta \\ &= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta \\ &= \binom{n}{x} \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, d\theta \\ &= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)} \quad \text{para } x \in 0, 1, ..., n. \end{aligned}$$

Esta distribución es la *distribución beta-binomial* con parámetros $n, \alpha, \beta$. Se tiene:

$$E[X] = \frac{n\alpha}{\alpha+\beta}, \qquad V[X] = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

# Ejemplo

Supongamos una distribución Beta(5,5) a priori para $\theta = P(\text{cruz})$.

Si observamos 9 cruces y 3 caras en 12 tiradas, ¿cuál es la distribución a posteriori de $\theta$?

# Ejemplo

Supongamos una distribución Beta(5,5) a priori para $\theta = P(\text{cruz})$.

Si observamos 9 cruces y 3 caras en 12 tiradas, ¿cuál es la distribución a posteriori de $\theta$?

$$\theta|\text{data} \sim \text{Beta}(9+5, 3+5) = \text{Beta}(14, 8).$$

¿Cuál es la media a posteriori de $\theta$?

# Ejemplo

Supongamos una distribución Beta(5,5) a priori para $\theta = P(\text{cruz})$.

Si observamos 9 cruces y 3 caras en 12 tiradas, ¿cuál es la distribución a posteriori de $\theta$?

$$\theta|\text{data} \sim \text{Beta}(9+5, 3+5) = \text{Beta}(14, 8).$$

¿Cuál es la media a posteriori de $\theta$?

$$E[\theta|\text{datos}] = \frac{14}{14+8} = \frac{7}{11}.$$

¿Cuál es la distribución del número de cruces en 10 tiradas más de la moneda?

# Ejemplo

Supongamos una distribución Beta(5,5) a priori para $\theta = P(\text{cruz})$.

Si observamos 9 cruces y 3 caras en 12 tiradas, ¿cuál es la distribución a posteriori de $\theta$?

$$\theta|\text{data} \sim \text{Beta}(9+5, 3+5) = \text{Beta}(14, 8).$$

¿Cuál es la media a posteriori de $\theta$?

$$E[\theta|\text{datos}] = \frac{14}{14+8} = \frac{7}{11}.$$

¿Cuál es la distribución del número de cruces en 10 tiradas más de la moneda?

$$X|\theta, \text{datos} \sim \text{Binomial}(10, \theta), \qquad X|\text{datos} \sim \text{Beta-binomial}(10, 14, 8).$$

# Ventajas de distribuciones conjugadas

- Facilidad del cálculo:

# Ventajas de distribuciones conjugadas

- Facilidad del cálculo:
  Para computar la distribución a posteriori, la distribución predictiva etc. sólo tenemos que cambiar los valores de los parámetros.
- Fácil interpretación de la información representada por la distribución a priori:

# Ventajas de distribuciones conjugadas

- Facilidad del cálculo:
  Para computar la distribución a posteriori, la distribución predictiva etc. sólo tenemos que cambiar los valores de los parámetros.
- Fácil interpretación de la información representada por la distribución a priori:
  En nuestro ejemplo, dada la a priori Beta($\alpha, \beta$), la distribución a posteriori es Beta($\alpha$+cruces,$\beta$+caras).

# Ventajas de distribuciones conjugadas

- Facilidad del cálculo:
  Para computar la distribución a posteriori, la distribución predictiva etc. sólo tenemos que cambiar los valores de los parámetros.
- Fácil interpretación de la información representada por la distribución a priori:
  En nuestro ejemplo, dada la a priori Beta($\alpha, \beta$), la distribución a posteriori es Beta($\alpha$+cruces,$\beta$+caras).
  Luego, la información representado por la a priori equivale a ver una muestra de $\alpha + \beta$ tiradas de la moneda con $\alpha$ cruces y $\beta$ caras.

# Ventajas de distribuciones conjugadas

- Facilidad del cálculo:
  Para computar la distribución a posteriori, la distribución predictiva etc. sólo tenemos que cambiar los valores de los parámetros.
- Fácil interpretación de la información representada por la distribución a priori:
  En nuestro ejemplo, dada la a priori Beta($\alpha, \beta$), la distribución a posteriori es Beta($\alpha$+cruces,$\beta$+caras).
  Luego, la información representado por la a priori equivale a ver una muestra de $\alpha + \beta$ tiradas de la moneda con $\alpha$ cruces y $\beta$ caras.

$$
\begin{aligned}
E[\theta|\text{data}] &= \frac{\alpha + \text{cruces}}{\alpha + \beta + \text{cruces} + \text{caras}} \\
&= \frac{\alpha + \beta}{\alpha + \beta + \text{tiradas}} \times \frac{\alpha}{\alpha + \beta} + \frac{\text{tiradas}}{\alpha + \beta + \text{tiradas}} \times \frac{\text{cruces}}{\text{tiradas}} \\
&= wE[\theta] + (1 - w)\hat{\theta} \quad \text{donde } w = \frac{\alpha + \beta}{\alpha + \beta + \text{tiradas}}.
\end{aligned}
$$

# Derivación de una distribución a priori "objetiva"

Dejando $\alpha, \beta \to 0$, tenemos una distribución a priori que equivale a haber observado 0 cruces y 0 caras.

Esta distribución se llama la *distribución de Haldane*:

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

# Derivación de una distribución a priori "objetiva"

Dejando $\alpha, \beta \to 0$, tenemos una distribución a priori que equivale a haber observado 0 cruces y 0 caras.

Esta distribución se llama la *distribución de Haldane*:

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

Es una distribución *impropia* porque $\int_0^1 f(\theta)\, d\theta = \infty$.

¿Importa?

# Derivación de una distribución a priori "objetiva"

Dejando $\alpha, \beta \to 0$, tenemos una distribución a priori que equivale a haber observado 0 cruces y 0 caras.

Esta distribución se llama la *distribución de Haldane*:

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

Es una distribución *impropia* porque $\int_0^1 f(\theta)\, d\theta = \infty$.

¿Importa?

Dada la muestra, tenemos $\theta|\text{datos} \sim \text{Beta}(\text{cruces}, \text{caras})$ con media a posteriori $E[\theta|\text{datos}] = \hat{\theta}$, el EMV. Por lo general, ningún problema.

# Derivación de una distribución a priori "objetiva"

Dejando $\alpha, \beta \to 0$, tenemos una distribución a priori que equivale a haber observado 0 cruces y 0 caras.

Esta distribución se llama la *distribución de Haldane*:

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

Es una distribución *impropia* porque $\int_0^1 f(\theta)\, d\theta = \infty$.

¿Importa?

Dada la muestra, tenemos $\theta|\text{datos} \sim \text{Beta}(\text{cruces}, \text{caras})$ con media a posteriori $E[\theta|\text{datos}] = \hat{\theta}$, el EMV. Por lo general, ningún problema.

Pero si sólo observamos cruces (o caras), la distribución a posteriori es también impropia. ¡Un gran problema!

# Mixturas de distribuciones conjugadas

De vez en cuando, una única distribución conjugada no representa bien las opiniones del experto. No obstante, podemos utilizar una mixtura

$$f(\theta) = \sum_{i=1}^{k} w_i f(\theta|\alpha_i, \beta_i)$$

para representar cualquier distribución continua.

Luego, la distribución a posteriori sigue siendo una mixtura de distribuciones conjugadas:

$$f(\theta|\text{datos}) = \sum_{i=1}^{k} w_i^* f(\theta|\alpha_i^*, \beta_i^*).$$

# Ejemplo

Volvemos al ejemplo y supongamos que en lugar de la a priori anterior, utilizamos una mixtura:

$$f(\theta) = 0{,}6\text{Beta}(5,5) + 0{,}4\text{Beta}(3,1).$$

Ahora, la distribución a posteriori es:

$$\begin{aligned} f(\theta|\text{datos}) &\propto \left(0{,}6\frac{1}{B(5,5)}\theta^{5-1}(1-\theta)^{5-1} + 0{,}4\frac{1}{B(3,1)}\theta^{3-1}(1-\theta)^{1-1}\right)\theta^{9}(1-\theta)^{3} \\ &\propto \frac{0{,}6}{B(5,5)}\theta^{14-1}(1-\theta)^{8-1} + \frac{0{,}4}{B(3,1)}\theta^{12-1}(1-\theta)^{4-1} \\ &\propto \frac{0{,}6B(14,8)}{B(5,5)}\frac{1}{B(14,8)}\theta^{14-1}(1-\theta)^{8-1} + \\ &\quad \frac{0{,}4B(12,4)}{B(3,1)}\frac{1}{B(12,4)}\theta^{11-1}(1-\theta)^{4-1} \\ &= w^{*}\text{Beta}(14,8) + (1-w^{*})\text{Beta}(12,4) \text{ donde } w^{*} = \frac{\frac{0{,}6B(14,8)}{B(5,5)}}{\frac{0{,}6B(14,8)}{B(5,5)} + \frac{0{,}4B(12,4)}{B(3,1)}}. \end{aligned}$$

# ¿Siempre existe una distribución a priori conjugada?

Supongamos que tomamos una muestra de la distribución Cauchy con parámetro de localización $\theta$ y parámetro de escala 1. Luego, la verosimilitud es: CD

$$l(\theta|\text{data}) \propto \prod_{i=1}^{n} \frac{1}{(1+(y_i-\theta)^2)}$$

y no existe ninguna distribución a priori conjugada en este caso.

# ¿Cuando existe una distribución a priori conjugada?

Si la distribución del muestreo es una *familia exponencial*, tenemos:

$$f(y|\theta) = h(y)g(\theta)\exp\left(\eta(\theta)T(y)\right).$$

Dada una muestra de tamaño $n$, la verosimilitud es

$$l(\theta|\text{datos}) = \prod_{i=1}^{n} h(y_i)g(\theta)\exp\left(\eta(\theta)T(y_i)\right) \propto g(\theta)^n \exp\left(n\bar{T}(\text{y})\eta(\theta)\right)$$

donde $\bar{T}(\text{y}) = \frac{1}{n}\sum_{i=1}^{n} T(y_i)$ es un *estadístico suficiente*.

# ¿Cuando existe una distribución a priori conjugada?

Luego, si utilizamos una distribución a priori

$$f(\theta) \propto g(\theta)^a \exp\left(b\eta(\theta)\right),$$

entonces la distribución a posteriori es

$$f(\theta|\text{datos}) \propto g(\theta)^{a^*} \exp\left(b^*\eta(\theta)\right),$$

donde $a^* = a + n$ y $b^* = b + n\bar{T}$.

La distribución tiene la misma forma que la a priori, sólo cambiando los parámetros. ¡La distribución a priori es conjugada!

# Sucesos raros

Supongamos que vamos a observar el número de sucesos raros, $X$, en un periodo de tiempo $t$ donde el número medio en un periodo de tiempo 1 es igual a $\theta$. Luego $X|\theta \sim \text{Poisson}(t\theta)$ y entonces

$$P(X = x|\theta) = \frac{(t\theta)^x e^{-t\theta}}{x!}.$$

¿Es una familia exponencial?

# Sucesos raros

Supongamos que vamos a observar el número de sucesos raros, $X$, en un periodo de tiempo $t$ donde el número medio en un periodo de tiempo 1 es igual a $\theta$. Luego $X|\theta \sim \text{Poisson}(t\theta)$ y entonces

$$P(X = x|\theta) = \frac{(t\theta)^x e^{-t\theta}}{x!}.$$

¿Es una familia exponencial?

$$P(X = x|\theta) = \underbrace{\frac{t^x}{x!}}_{h(x)} \underbrace{e^{-t\theta}}_{g(\theta)} \exp\left( \underbrace{x}_{T(x)} \underbrace{\log\theta}_{\eta(\theta)} \right).$$

¿Cómo sería una distribución a priori conjugada?

# Sucesos raros

Supongamos que vamos a observar el número de sucesos raros, $X$, en un periodo de tiempo $t$ donde el número medio en un periodo de tiempo 1 es igual a $\theta$. Luego $X|\theta \sim \text{Poisson}(t\theta)$ y entonces

$$P(X = x|\theta) = \frac{(t\theta)^x e^{-t\theta}}{x!}.$$

¿Es una familia exponencial?

$$P(X = x|\theta) = \underbrace{\frac{t^x}{x!}}_{h(x)} \underbrace{e^{-t\theta}}_{g(\theta)} \exp\left( \underbrace{x}_{T(x)} \underbrace{\log\theta}_{\eta(\theta)} \right).$$

¿Cómo sería una distribución a priori conjugada?

$$f(\theta) \propto \left(e^{-t\theta}\right)^a \exp\left(b \log\theta\right).$$

¿Reconocemos esta distribución?

# Sucesos raros

$$\begin{aligned} f(\theta) &\propto \left(e^{-t\theta}\right)^a \exp\left(b \log \theta\right) \\ &\propto \theta^b e^{-at\theta} \\ &\propto \theta^{\alpha-1} e^{-\beta\theta} \quad \text{donde } \alpha = b+1, \beta = at. \end{aligned}$$

¿Reconocemos esta distribución ahora?

# Sucesos raros

$$\begin{aligned} f(\theta) &\propto \left(e^{-t\theta}\right)^a \exp\left(b \log \theta\right) \\ &\propto \theta^b e^{-at\theta} \\ &\propto \theta^{\alpha-1} e^{-\beta\theta} \quad \text{donde } \alpha = b+1, \beta = at. \end{aligned}$$

¿Reconocemos esta distribución ahora?

Es una distribución gamma:

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$$

para $\theta > 0$.

# La distribución a posteriori

Dada la distribución a priori $\theta \sim \text{Gamma}(\alpha, \beta)$, la distribución a posteriori es

$$\begin{aligned} f(\theta|\text{datos}) &\propto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\frac{(t\theta)^x e^{-t\theta}}{x!} \\ &\propto \theta^{\alpha+x-1}e^{-(\beta+t)\theta} \\ &= \frac{(\beta+t)^{\alpha+x}}{\Gamma(\alpha+x)}\theta^{\alpha+x-1}e^{-(\beta+t)\theta} \\ \theta|\text{datos} &\sim \text{Gamma}(\alpha+x, \beta+t). \end{aligned}$$

# La media a posteriori

La media a posteriori es

$$\begin{aligned} E[\theta|\text{datos}] &= \frac{\alpha + x}{\beta + t} \\ &= \frac{\beta}{\beta + t}\frac{\alpha}{\beta} + \frac{t}{\beta + t}\frac{x}{t} \\ &= wE[\theta] + (1 - w)\hat{\theta} \quad \text{donde } \hat{\theta} = \frac{x}{n} \text{ es el EMV.} \end{aligned}$$

La media a posteriori es una media ponderada de la media a priori y el EMV donde el peso de la media a priori es $w = \beta/(\beta + t)$.

# Interpretación de los parámetros y a priori "objetiva"

La información en la a priori equivale a la información cuando observemos $\alpha$ sucesos en un periodo de tiempo $\beta$.

Cuando dejamos $\alpha, \beta \to 0$, la distribución a priori es impropia:

$$f(\theta) \propto \frac{1}{\theta}$$

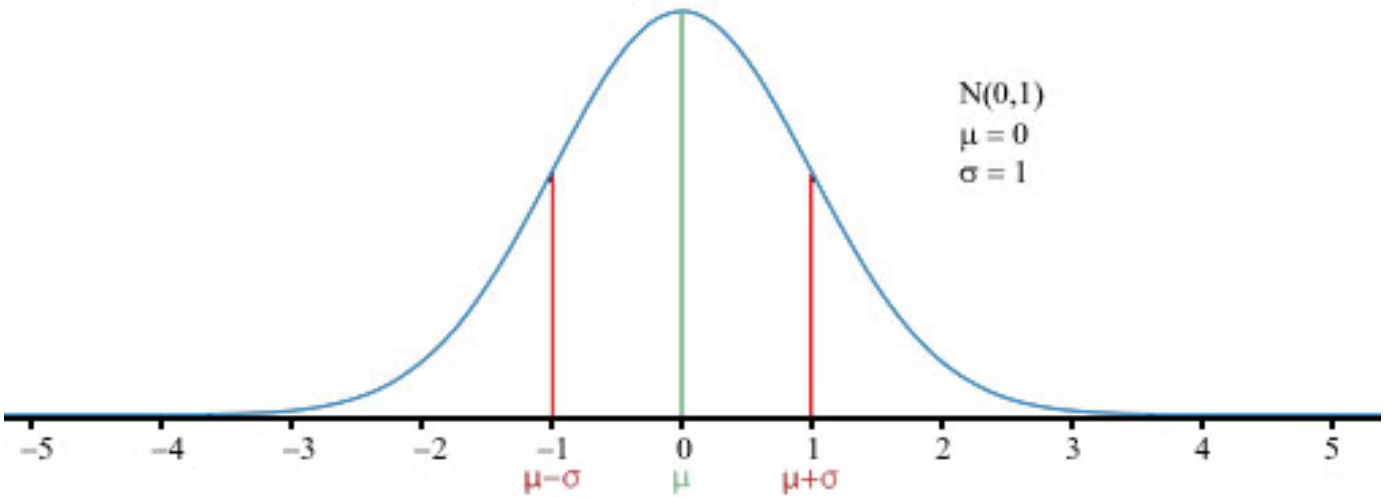
y la distribución a posteriori es

$$\theta|\text{datos} \sim \text{Gamma}(x, t)$$

con media $E[\theta|\text{datos}] = \frac{x}{t} = \hat{\theta}$.

# Resumen y siguiente sesión

En esta clase, hemos introducido las distribuciones a priori conjugadas.




En la siguiente sesión, exploramos inferencia para modelos asociados con la distribución normal.

# Apéndice: la distribución Cauchy

Una variable continua, $Y$, tiene una distribución Cauchy con parámetro de localización $\theta \in \mathbb{R}$ y parámetro de escala $\psi > 0$ si

$$f(y) = \frac{1}{\psi\pi\left(1 + \left(\frac{y-\theta}{\psi}\right)^2\right)} \quad \text{para } y \in \mathbb{R}.$$

La variable no tiene ni media ni varianza.