Ampliación de Inferencia Estadística

Tercero Grado de Estadística

Universidad de Sevilla

Tema 4: Inferencia Bayesiana en el modelo Poisson

Práctica-R-4

Problema 1.

Supongamos que deseamos estudiar la razón de éxito de operación de trasplantes de corazón en un determinado hospital. En particular, se observa el número de operaciones n y el número de fallecimientos en 30 días desde la operación que se denotará por y. En definitiva, nos interesa la probabilidad de fallecimiento de los pacientes intervenidos. Esta predicción se basa en un modelo que usa información como la condición del paciente antes de la operación, el sexo y la raza. Un modelo estandar consiste en suponer que el número de fallecimientos y sigue una Poisson de media $n \cdot \theta$, y el objetivo es estimar la razón de fallecimientos θ. Teniendo en cuenta que es una aproximación ya que si realizamos n operaciones como máximo habrá n fallecimientos y, como bien se sabe, los posibles valores de una variable aleatorioa con distribución de Poisson es $0 \cup \mathbb{N}$. En este sentido, en la aproximación consideraremos el rango viable para el número de fallecimientos hasta que prácticamente su probabilidad sea cero.

La estimación usual de θ es el estimador de máxima verosimilitud: $\hat{\theta} = y/n$. Desafortunadamente, este estimador puede ser "pobre" cuando el número de fallecimientos y es muy próximo a cero. En este caso, cuando se contabilizan pocos fallecimientos, es preferible usar una estimación Bayesiana que usa un conocimiento a priori acerca del tamaño de la razón de mortalidad.

En concreto, supongamos que el número de fallecimientos observados es z_j y el número de operaciones es o_j para cada diez hospitales ($j = 1, \ldots, 10$) donde z_j es una Poisson con media $o_j \cdot \theta$. Si asignamos a θ una a priori no informativa $g(\theta) \propto 1/\theta$ la distribución a posteriori es una gamma de parámetros $a = \sum z_j$ y $\lambda = \sum o_j$. Supongamos que los datos arrojan un valor de 16 fallecimientos y un total de 15174 operaciones entre los diez hospitales.

Ahora bien, consideremos que para el hospital A se encuentra un solo fallecimiento entre 66 operaciones efectuadas. Es decir este hospital tiene una razón de fallecimientos de 1/66. Supongamos que en el mes próximo se van a realizar $n_0 = 10$ operaciones. Estamos interesados en predicir el número de fallecimientos. Se tiene que el número de fallecimientos sigue, aproximadamente, una distribución de Poisson de parámetro $n_0\theta = 10\theta$ y la distribución a posteriori (es decir, una vez que hemos observado 1 fallecimiento de 66 operaciones) es una gamma de parámetros $16 + 1$ y $15174 + 66$. Por tanto, la distribución predictiva es

$$\begin{aligned} P(Y = k|\mathbf{x}) &= \int_0^{+\infty} e^{-10\theta} \frac{(10\theta)^k}{k!} \frac{(15174+66)^{17}}{\Gamma(17)} \theta^{16} e^{-(15174+66)\theta} d\theta \\ &= \frac{(16+k)!}{16!k!} \left(\frac{15174+66}{10+15174+66}\right)^{17} \left(\frac{10}{10+15174+66}\right)^k \text{ para } k = 0, 1, 2, \ldots \end{aligned}$$

Como se puede comprobar $Y|\mathbf{x} \sim BinNeg(17, \frac{15174+66}{10+15174+66})$. Aquí de nuevo hay que entender el espacio paramétrico Θ y el rango de Y. Tal como se ha definido el posible valor de la intensidad de fallecimientos en realidad se tendría que $\Theta = [0, 1]$. Sin embargo, se ha tomado como $\Theta = [0, +\infty)$ por cuestiones de simplificar los cálculos. Hacerlo de otra manera se saldría de los límites y objetivos del presente curso. En cuanto al rango de Y, en teoría no puede haber más de 10 fallecimientos ya que son las operaciones que se van a realizar pero estos modelos nos permiten aproximarnos bastante bien a los valores de dichas probabilidades. Si desde el principio hubiésemos restringido los posibles valores de fallecimientos tendríamos en realidad una Poisson

truncada, es decir la función de probabilidad hubiese sido

$$P(Y=k|\theta,n_0) = \frac{e^{-(n_0\theta)}(n_0\theta)^k/k!}{\sum_{k=0}^{n_0} e^{-(n_0\theta)}(n_0\theta)^k/k!} = \frac{(n_0\theta)^k/k!}{\sum_{k=0}^{n_0}(n_0\theta)^k/k!} \text{ para } k=0,1,\ldots,n_0.$$

Lo cual hubiese complicado excesivamente el problema. Por ejemplo, simplemente proponer una distribución a priori de Jeffrey nos obligaría a usar métodos de simulación para obtener la distribución a posteriori de θ, y este tópico no entra dentro de los objetivos del curso. En cualquier caso, y por simplicidad, vamos a suponer que la distribución Poisson no está truncada. A continuación hagamos los cálculos necesarios.

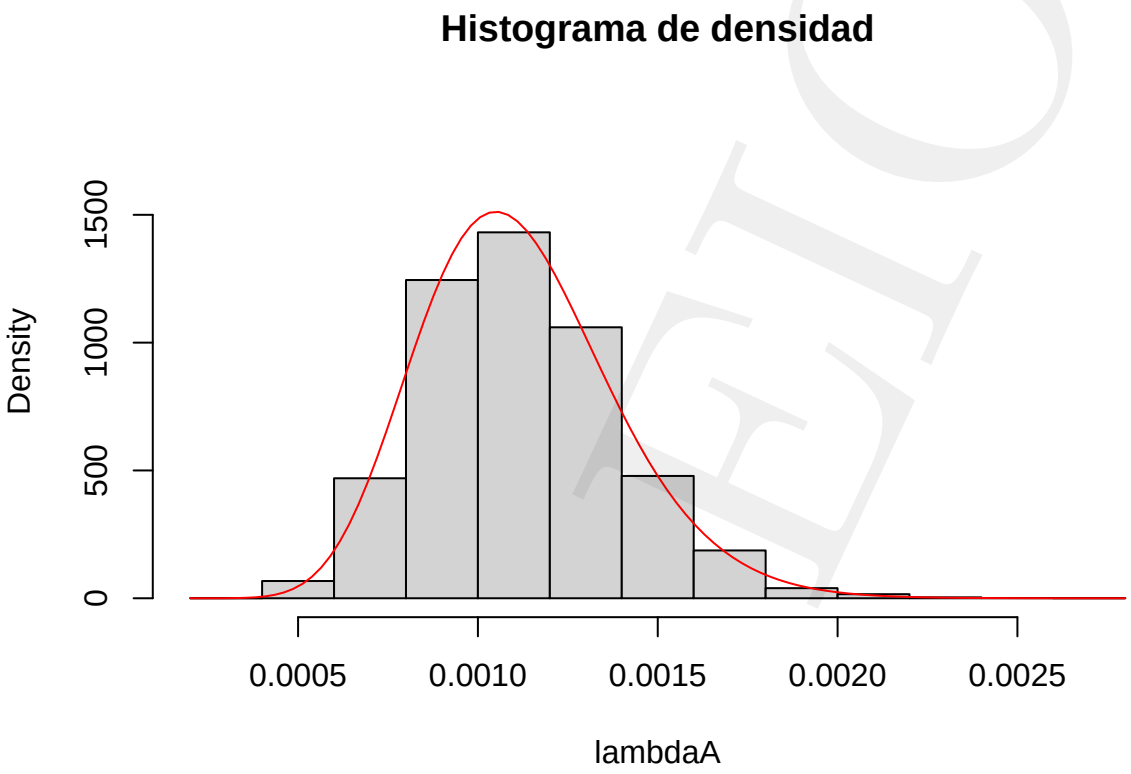
```
alpha=16; beta=15174
yobs=1;ex=66
ncero=10
y=0:10


py=(gamma(alpha+yobs+1)/ gamma(alpha+yobs))*((1/gamma(y+1))*((beta+ex)/(ncero+beta+ex))**(17))*
  ((ncero)/(ncero+beta+ex))**(y)
pp=sum(py)

popi=dnbinom(y,alpha+yobs,(beta+ex)/(ncero+beta+ex))

cbind(y,round(popi,3),round(py/pp,3))
```

```
##         y
##  [1,]  0 0.989 0.999
##  [2,]  1 0.011 0.001
##  [3,]  2 0.000 0.000
##  [4,]  3 0.000 0.000
##  [5,]  4 0.000 0.000
##  [6,]  5 0.000 0.000
##  [7,]  6 0.000 0.000
##  [8,]  7 0.000 0.000
##  [9,]  8 0.000 0.000
## [10,]  9 0.000 0.000
## [11,] 10 0.000 0.000
```

Como puede observarse hay diferencias en las dos columnas cuando se trata de un fallecimiento y eso se debe a las aproximaciones de R. En este caso, y como son todos números naturales se aconseja usar el comando correspondiente de la binomial negativa.

La densidad a posteriori de λ se puede resumir mediante 10000 valores simulados de la gamma. Esta simulación de valores es necesaria para posteriormente construir la región de credibilidad HPDI. Inicialmente veamos como el ajuste entre los valores muestrales y la densidad es buena de manera visual.

```
lambdaA=rgamma(10000, shape=alpha+yobs,rate=beta+ex)
hist(lambdaA,freq=FALSE,main="Histograma de densidad",ylim=c(0,1850))
curve(dgamma(x,alpha+yobs,rate=beta+ex),add=TRUE,col="red",xlab="Histograma de densidad")
```



Para calcular el intervalo de máxima densidad (HPDI) tenemos que recurrir a simular una muestra de la densidad a posteriori y luego usar el paquete `HDInterval`. En este caso tendríamos

```
library(HDInterval)
dens2 <- density(lambdaA)
hdi(dens2, credMass=0.90)
```

```
##        lower        upper
## 0.0006750443 0.0015573525
## attr(,"credMass")
## [1] 0.9
## attr(,"height")
## [1] 369.9722
```

Es decir, estaríamos hablando de una intensidad de fallecimientos de entre 6 y 15 por cada 10000 operaciones con una credibilidad del 90%.