Ejercicios de clase: Datos Poisson

Sergio Marrero Marrero

Lic. en Economía. Cuarto Curso

Asignatura: Métodos Bayesianos

Tutor: Francisco José Vázquez Polo

Universidad Las Palmas de Gran Canaria(ULPGC)

25 de marzo de 2017

1. Ejercicio 1

Supongamos que deseamos hacer inferencia sobre el parámetro $\lambda > 0$, de una población exponencial:

$$f(x \mid \lambda) = \lambda \cdot e^{-\lambda x}, \quad x \geq 0$$

donde $\mathbb{E}(X) = \frac{1}{\lambda}$ y $\mathbb{V}(X) = \frac{1}{\lambda^2}$. Para ello obtenemos una muestra aleatoria simple de tamaño n, $x_1, ..., x_n$. Se pide:

1. Obtener la función de verosimilitud y el estimador de máxima verosimilitud de λ.
2. Para una a priori para λ del tipo Gamma $\mathbb{G}(\alpha, \beta)$, obtener la distribución a posteriori de λ. Calcular la media y moda a posterior y comparar con el estimador de máxima verosimilitud.

Apartado 1

a) La verosimilitud de la muestra de tamaño n será:

$$f(\underline{x} \mid \lambda) = \prod_{i=1}^{n} f(x_i \mid p) = \lambda^n e^{\sum_{i=1}^{n} -\lambda x_i} = \lambda^n e^{-\lambda \bar{x} n}$$

b) Aplicamos logaritmos:

$$l(\underline{x} \mid \lambda) = log(f(\underline{x} \mid \lambda) = n[\ln(\lambda) - \lambda\overline{x}\ln(e)] = n[\ln(\lambda) - \lambda\overline{x})]$$

igualamos a cero y despejamos λ:

$$\frac{dl(\underline{x} \mid \lambda)}{d\lambda} = 0 \rightarrow \frac{n}{\lambda} - n\overline{x} = 0 \rightarrow \lambda = \frac{1}{\overline{x}}$$

Apartado 2

a) Aplicando el teormea de bayes:

$$\pi(\lambda \mid \underline{x}) \propto \pi(\lambda)f(\underline{x} \mid \lambda) = \lambda^{\alpha-1}e^{-\beta\lambda} \cdot \lambda^{n}e^{-\lambda\overline{x}n} = \lambda^{n+\alpha-1} \cdot e^{-(\overline{x}n+\beta)\lambda}$$

Con lo cual:

$$\pi(\lambda \mid \underline{x}) \equiv Gamma(a,b) \qquad \text{donde} \qquad a = n+\alpha \quad \text{y} \quad b = \overline{x}n+\beta$$

b) Se pide calcular la media y la moda y luego comparar con el EMV. Usando la media y moda de la distribución gamma obtenemos:

$$\text{Moda} = \frac{a-1}{b} = \frac{n+\alpha-1}{\overline{x}n+\beta} \qquad \text{Media} = \frac{a}{b} = \frac{n+\alpha}{\overline{x}n+\beta} \qquad \text{EMV} = \frac{1}{\overline{x}}$$

Para estudiar los casos límites exHacemos las siguientes transformaciones:

$$\text{Moda} = \frac{n}{\overline{x}n+\beta} + \frac{\alpha-1}{\overline{x}n+\beta} \qquad \text{Media} = \frac{n}{\overline{x}n+\beta} + \frac{\alpha}{\overline{x}n+\beta}$$

Si la muestra tendiera a infinito $n \rightarrow \infty$ entonces:

$$\text{Moda} = \frac{1}{\overline{x}} \qquad \text{Media} = \frac{1}{\overline{x}} \qquad \text{EMV} = \frac{1}{\overline{x}}$$

Media y moda a posteriori coincidirían con EMV.

2. Ejercicio 2

La siguiente tabla contiene el número de días que funciona un máquina entre fallos sucesivos:

$$\begin{bmatrix} 8 & 10 & 11 & 4 & 5 & 2 & 5 & 8 & 5 & 7 \\ 1 & 6 & 1 & 6 & 4 & 1 & 1 & 1 & 6 & 1 \\ 1 & 30 & 7 & 91 & 2 & 1 & 87 & 47 & 10 & 9 \\ 10 & 60 & 16 & 38 & & & & & & \end{bmatrix}$$

El controlador de la máquina asume un modelo de Poisson para estos datos con una media de fallos al mes de $30\,\theta$. Su juicio sobre el parámetro $(30\,\theta)$ antes de tomar los datos es que vale 6, y se sorprendería de valores menores de 1 y mayores que 10. Interpretar esta información a priori (más bien pobre o difusa) en términos de una distribución Gamma y obtener su distribución a posteriori. Se pide:

1. Obtener la probabilidad a posteriori de que $\theta > \frac{6}{30}$.
2. Comparar ahora con una información a priori no informativa del tipo $\pi(\theta) \propto \frac{1}{\theta}$. Comentar.

Apartado 1

a) La primera tarea será calcular el número de errores mensual. Para ello se ha diseñado el siguiente algoritmo(Matlab).

```

lista = [8, 10, 11, 4, 5, 2, 5, 8, 5, 7, 1,...
          6, 1, 6, 4, 1, 1, 1, 6, 1, 1, 30, ...
          7, 91, 2, 1, 87, 47, 10, 9, 10, 60, 16, 38];
lista_acumulada = cumsum(lista); %suma acumulada de la lista
dias_total = lista_acumulada(end); %numero total de dias
meses_total = round(dias_total / 30); %numero total de meses
lista_de_fallos_en_cada_mes = zeros(1,meses_total);
for i = 1:numel(lista_de_fallos_en_cada_mes)
    %fallos cometidos en un mes
    fallos_en_mes = lista_acumulada <= 30;
    lista_de_fallos_en_cada_mes(i) = sum(fallos_en_mes);
    %Reducimos la lista con los datos usados restamos 30
    lista_acumulada(1:sum(fallos_en_mes)) = [];
    lista_acumulada = lista_acumulada - 30;
end
if sum(lista_de_fallos_en_cada_mes) == numel(lista) %Comprobacion
    disp('Las fallos por mes han sido:')
    disp(lista_de_fallos_en_cada_mes)
end
```

El resultado viene en la siguiente lista:

$$\begin{bmatrix} 3 & 6 & 9 & 3 & 2 & 0 & 0 & 3 & 0 \\ 0 & 1 & 1 & 3 & 0 & 1 & 1 & 1 & \end{bmatrix}$$

b) Modelamos la creencia del experto usando una Gamma(α, β). Asumimos un grado de confianza del 95 %:

$$\begin{cases} 30\theta = 6 \quad \rightarrow \quad \text{Media} = \frac{\alpha}{\beta} = 6 \quad \rightarrow \quad \alpha = 6\beta \\ Prob\{1 < \theta < 10\} > 5\,\% \quad \rightarrow \quad \alpha = 1.5, \beta = 9 \quad \text{se obtiene} \quad Prob\{1 < \theta < 10\} = 0.96 \end{cases} \tag{1}$$

Con lo cual, es aceptable modelar la opinión del experto con una Gamma(1.5, 9).

c) La a posteriori vendría dado conjugando la verosimilitud con la a priori a través del teorema de bayes, tal que:

$$\pi(\theta \mid \underline{x}) \propto \pi(\theta) f(\underline{x} \mid \theta) = \theta^{n\overline{x}+\alpha-1} \cdot e^{-(n+\beta)\theta}.$$

Se pide calcular la probabilidad a posteriori de que $\theta > \frac{6}{30}$. Introduciendo en el bayes los datos transformados y la a priori obtendremos:

$$prob\{(30\theta)\} > 6 = 0$$

Apartado 2)

a) A continuación la a priori que se utilizará será no informativa del tipo: $\pi(\theta) \propto \frac{1}{\theta}$. Esta no informativa la modelamos como Gamma$(\alpha \longrightarrow 0, \beta \longrightarrow 0)$.

Introduciendo estos datos en el Bayes con los siguientes parámetros Gamma(0.2,0.2)(el firstBayes no acepta valores más pequeños) podemos ver la siguiente comparativa de las dos prioris.

	$\pi(\theta_1)$	$\pi(\theta_2)$
media	1.9884	2.3243
moda	1.9302	2.27
mediana	1.9689	2.3062
HDI(95 %)	[1.344 2.665]	[1.65, 3.02]
Prob[$30\theta > 6$]	0	0

La siguiente figura muestra el triplot de ambas distribuciones:

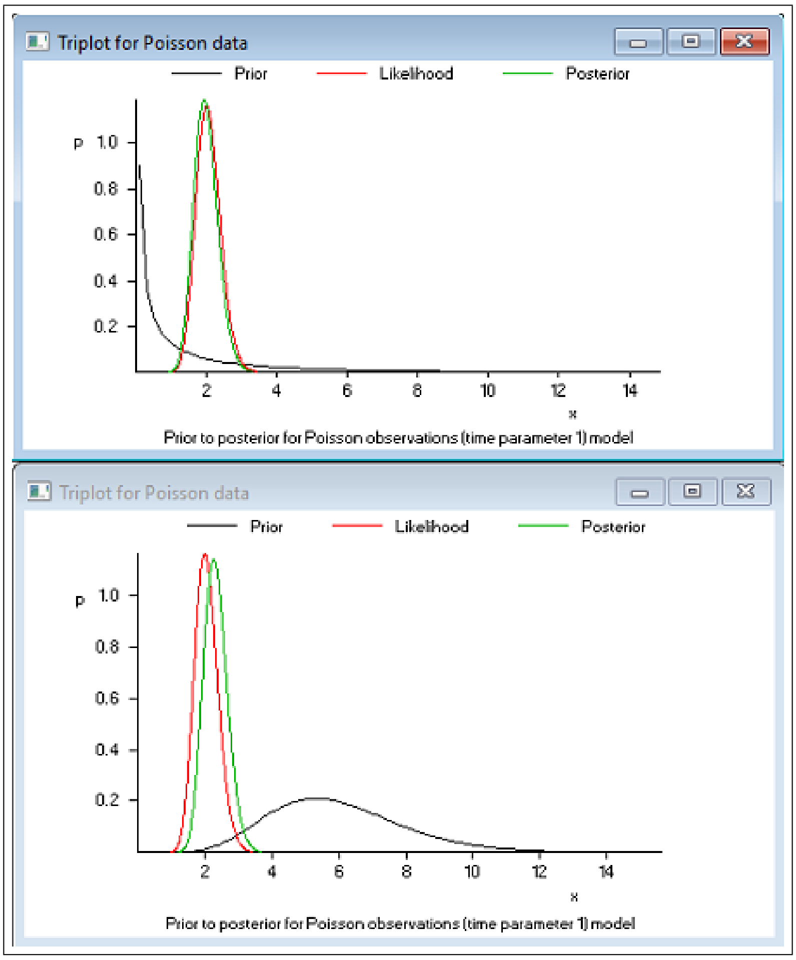

Figura 1: Triplot de ambas a prioris

Se observa que el análisis es robusto, ya que existe cierta independencia del a priori utilizado.

3. Ejercicio 3

Una compañía aseguradora asume que el número de reclamaciones recibidas anualmente por cada cliente de una determinada cartera tiene una distribución de Poisson de parámetro λ y que este número es independiente de un año a otro. Supongamos que se observan T años de un asegurado elegido al azar de dicha cartera. Estamos interesados en el número medio de reclamaciones λ. Consideramos una densidad a priori del tipo:

$$\pi(\lambda) = K \cdot (1+\lambda) \cdot e^{-\lambda}, \quad \lambda > 0$$

1. Calcular K y la media a priori de λ. Calcular también la probabilidad a priori $\Pr(\lambda > 4)$.
2. Supongamos que los últimos 10 años el asegurado presenta las siguientes reclamaciones: $1, 0, 1, 2, 3, 0, 0, 1, 1, 2$. Obtener la densidad a posteriori, media, mediana y moda a posteriori y un intervalo bayesiano a posteriori del 99 %.
3. Suponiendo que la compañía tiene 500 asegurados en esta cartera y que destina 100 euros para cada reclamación recibida. Obtener un intervalo al 95 % de previsión de costes para el siguiente año.
4. Contrastar la hipótesis de que el número medio de reclamaciones que recibe la compañía no es mayor de 3.
5. Realizar el mismo análisis de los apartados anteriores pero con una a priori tipo Gamma, conociendo que la compañía estima como más frecuente 2 reclamaciones al año y que no más de 5 reclamaciones son asumibles en un año.

Apartado 1

a) Teniendo en cuenta la a priori representa una mixtura y que además, la integral de cada una de ellos con sus constantes correspondientes debe integrar la unidad, se obtiene que:

$$K[\frac{\Gamma(\alpha_1)}{\beta_2^{\alpha_1}} + \frac{\Gamma(\alpha_2)}{\beta_2^{\alpha_2}}] = 1 \rightarrow K = \frac{1}{\frac{\Gamma(\alpha_1)}{\beta_2^{\alpha_1}} + \frac{\Gamma(\alpha_2)}{\beta_2^{\alpha_2}}}$$

Como para este caso concreto $\alpha_1 = 1, \beta_1 = 1, \alpha_2 = 2, \beta_2 = 1$ queda que:

$$K = \frac{1}{\frac{1}{1} + \frac{1}{1}} \rightarrow K = \frac{1}{2}$$

Por otro lado, los coeficientes de la mixtura serán forzosamente: $\epsilon_1 = \epsilon_2 = \frac{1}{2}$.
Para la media a priori usamos el valor esperado de la distribución Gamma.

$$E[\lambda] = \epsilon_1 E[\lambda_1] + \epsilon_2 E[\lambda_2] = \epsilon_1 \frac{\alpha_1}{\beta_1} + \epsilon_2 \frac{\alpha_2}{\beta_2} = \frac{1}{2}1 + \frac{1}{2}2 = \frac{3}{2}$$

Con el FirstBayes se obtienen los mismo resultados.

b) Con el FirstBayes calculamos la probabilidad a priori de que $\lambda > 4$. $Prob\{\lambda > 4\} = 0.055$

2 Introduciendo los datos dados, y con la información de la a priori, obtenemos que la densidad a posteriori sigue una mixtura tal que $0.478Ga(11, 12) + 0.522Ga(11, 13)$. En la siguiente tabla se visualizan los estadísticos a posteriori pedidos. La figura 2 muestra el triplot de este caso.

	$\pi(\lambda)$	$\pi(\lambda \mid \underline{x})$
media	1.5	1.1383
moda	0.014	1.1078
mediana	1.14	1.0465
HDI(99 %)	[0, 5.98]	[0.43, 2.07]

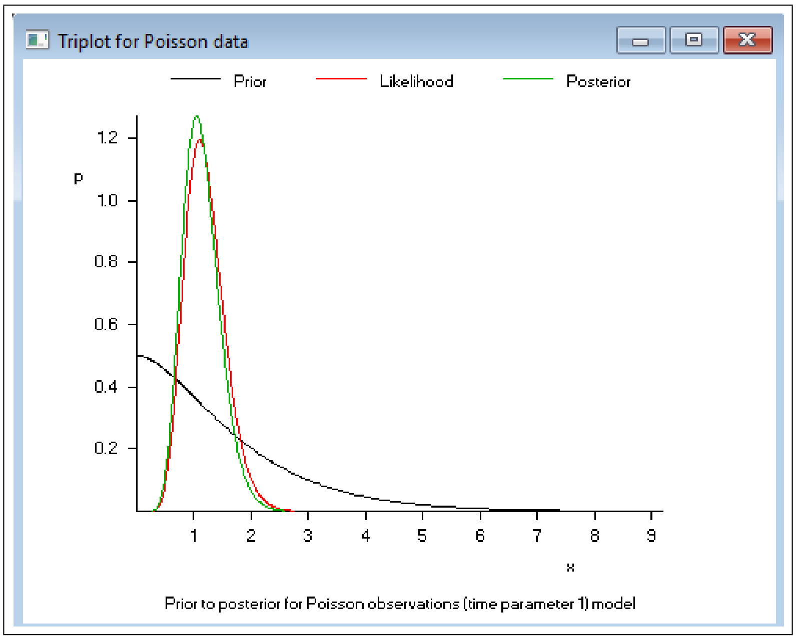

Figura 2: Triplot de ambas a prioris

Apartado 3

a) La predictiva para un sólo cliente y un sólo año, muestra que el HDI(95 %) se encuentra en el intervalo $[0, 3]$. Esto significa que hay un 95 % de probabilidades de que al año siguiente un cliente haga entre ninguna y tres reclamaciones. Como por cada reclamación se deberá destinar 100 € y en total hay 500 clientes, entonces el intervalo bayesiano en términos de previsión de costes será de $[0, 150.000€]$

Apartado 4 Contrastar la hipótesis de que el número medio de reclamaciones que recibe la compañía no es mayor a 3.

$$\begin{cases} H_0 : \lambda > 3 \\ H_1 : \lambda \leq 3 \end{cases} \tag{2}$$

La siguiente tabla muestra los resultados obtenidos al realizar el test de hipótesis para las a priori y la a posteriori:

	$\pi(\lambda)$	$\pi(\lambda \mid \underline{x})$
$Prob\{H_0\}$	0.12	0
$Prob\{H_1\}$	0.88	1

Los *odd ratio* a priori y a posteriori son:

$$oddratio_{priori} = \frac{\pi_0}{\pi_1} = \frac{0.12}{0.88} = 0.14 \qquad oddratio_{posteriori} = \frac{p_0}{p_1} = \frac{0}{1} = 0$$

$$B_{01} = \frac{\frac{p_0}{p_1}}{\frac{\pi_0}{\pi_1}} = \frac{0}{0.14} = 0$$

A priori se observa que H_1 es 7 veces más probable que H_0. A posteriori se obserba que H_1 es infinitamente más probable que H_0. Se observa a través del B_{01} que sólo los datos dictan que la H_1 es infinitamente más probable que la H_0.

Apartado 5

a) Modelamos la información de la compañía con una Gamma(α, β). Asumiendo que se considera suficiente la probabilidad de que un cliente tenga más de cinco reclamaciones sea infererior al 1 % se tiene que:

$$\begin{cases} Moda[\lambda] = 6 \quad \rightarrow \quad \text{Moda} = \frac{\alpha-1}{\beta} = 2 \quad \rightarrow \quad \alpha = 2\beta + 1 \\ Prob\{\lambda > 5\} < 1\,\% \quad \rightarrow \quad \alpha = 7, \beta = 3 \quad \text{se obtiene} \quad Prob\{\lambda > 5\} = 0.7\,\% \end{cases} \tag{3}$$

Con lo cual, es aceptable modelar la opinión del experto con una Gamma(7, 3).

b) Contrastar la hipótesis de que el número medio de reclamaciones que recibe la compañía no es mayor a 3.

$$\begin{cases} H_0 : \lambda > 3 \\ H_1 : \lambda \leq 3 \end{cases} \tag{4}$$

La siguiente tabla muestra los resultados obtenidos al realizar el test de hipótesis para las a priori y la a posteriori:

	$\pi(\lambda)$	$\pi(\lambda \mid \underline{x})$
$Prob\{H_0\}$	0.2	0
$Prob\{H_1\}$	0.8	1

Los *odd ratio* a priori y a posteriori son:

$$oddratio_{priori} = \frac{\pi_0}{\pi_1} = \frac{0.2}{0.8} = 0.25 \qquad oddratio_{posteriori} = \frac{p_0}{p_1} = \frac{0}{1} = 0$$

$$B_{01} = \frac{\frac{p_0}{p_1}}{\frac{\pi_0}{\pi_1}} = \frac{0}{0.25} = 0$$

A priori se observa que H_1 es 4 veces más probable que H_0. A posteriori se observa que H_1 es infinitamente más probable que H_0. Se observa a través del B_{01} que sólo los datos dictan que la H_1 es infinitamente más probable que la H_0.

Se deduce un análisis bayesiano robusto, pues se observa independencia de los datos.

4. Ejercicio 4

Una empresa desea contratar publicidad en una web y para ello investiga el número de visitas diarias que recibe dicha página web. Asumimos que el número de visitias sigue una distribución de Poisson de parámetro λ. Para la estimación de λ usamos las observaciones de n días, $x_1, ..., x_n$ y el juicio a priori no informativo dado por una densidad tipo Jeffreys:

$$\pi(\lambda) \propto \lambda^{-1}, \qquad \lambda > 0$$

1. Obtener la densidad a posteriori e identificar con el caso conjugado para datos de Poisson.

2. Para $n = 10$, se han observado los datos $350, 300, 270, 100, 700, 450, 300, 350, 325, 615$. Calcular media, moda y mediana a posteriori para el número medio de visitas a la web. Obtener un intervalo bayesiano a posteriori al 95 %. Obtener las probabilidades a posteriori para el test de hipótesis $H_0 : \lambda \leq 300, H_1 : 300 < \lambda \leq 500$

3. Supongamos ahora que estamos interesados en el test de hipótesis $H_0 : \lambda = 300, H_1 : \lambda = 400$. Para un caso no informativo, calcular el factor Bayes y comparar con el caso anterior.

Apartado 1 Combinando la a priori con la verosimilitud se obtiene:

$$\pi(\lambda \mid \underline{x}) \propto \pi(\lambda) f(\underline{x} \mid \lambda) = \lambda^{n\overline{x}-1} \cdot e^{-n\lambda}$$

Apartado 2

a) A continuación la a priori que se utilizará será no informativa del tipo: $\pi(\theta) \propto \frac{1}{\theta}$. Esta no informativa la modelamos como Gamma($\alpha \longrightarrow 0, \beta \longrightarrow 0$).

Calcularemos esto de dos formas diferentes y compararemos los resultados. Por un lado introduciremos en el FirstBayes una Gamma(0.2,0.2)(números menores no son aceptados). Por otro lado, utilizaremos la expresión de la densidad a posteriori exacta, la que hemos obtenido en el apartado anterior. En la siguiente tabla podemos ver los estadísticos comparados.

	$\pi(\theta)_{exacta}$	$\pi(\theta)_{aprox}$
media	376	368.65
moda	375.9	368.55
mediana	375.97	368.61
HDI(95 %)	[364.01 388.05]	[356.89, 380.46]

b) Realizamos el siguiente test de hipótesis

$$\begin{cases} H_0 : \lambda \leq 300 \\ H_1 : 300 < \lambda \leq 500 \end{cases} \tag{5}$$

La siguiente tabla muestra las prioridades a priori y posteriori:

	$\pi(\lambda)$	$\pi(\lambda \mid \underline{x})$
$Prob\{H_0\}$	1	0
$Prob\{H_1\}$	0	1
odd ratio	∞	0
B_{01}	0	

A priori se observa que H_0 es ∞ veces más probable que H_0. A posteriori se observa que H_1 es infinitamente más probable que H_0. Se observa a través del B_{01} que sólo los datos dictan que la H_1 es infinitamente más probable que la H_0.

Apartado 3 Realizamos ahora el siguiente test de hipótesis

$$\begin{cases} H_0 : \lambda = 300 & \rightarrow \pi_0 = \frac{1}{2} \\ H_1 : \lambda = 400 & \rightarrow \pi_1 = \frac{1}{2} \end{cases} \tag{6}$$

Con lo cual

$$\frac{p_0}{p_1} = \frac{\pi_0}{\pi_1} B_{01} \qquad \longrightarrow \qquad \frac{p_0}{p_1} = B_{01}$$

$$B_{01} = \frac{p_0}{p_1} = \frac{f(\underline{x} \mid \theta_0)}{f(\underline{x} \mid \theta_1} = \frac{\lambda_0^{n\overline{x}} e^{-n\lambda_0}}{\lambda_1^{n\overline{x}} e^{-n\lambda_1}} = (\frac{\lambda_0}{\lambda_1})^{n\overline{x}} e^{n(\lambda_1 - \lambda_1)} = [(\frac{\lambda_0}{\lambda_1})^{\overline{x}} e^{(\lambda_1 - \lambda_1)}]^n.$$

Solucionamos esta ecuación para los valores $\lambda_0 = 300, \lambda_1 = 400, n = 10, \overline{x} = 376$ se obtiene que $B_{01} = 0$. De la misma forma que en el apartado anterioror, la hipótesis H_1 es infinitamente más probable que la hipótesis H_0