Bayesian Analysis

Homework III

We observe the number of times that an event occurs. The prior information about the expected value of the counts of this event that we have is that the expected value will be between 4 and 6 with high probability. After observing a sample of size 5, $x = 1, 2, 2, 8, 10$, you have to decide whether this data have been generated by a *Poisson* model or by a *Geometric* model using a bayesian hypothesis test. And then, using the Bayesian model averaging approach, plot the posterior predictive distribution for a new future value and compute a point estimate and a 95% credible interval for the prediction. Explain all the steps you take in detail.

Useful information

If X is a random variable which follows a *Poisson* distribution with paratmeter λ, $Poisson(\lambda)$, then its probability function is:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

with $E(X) = \lambda$, $V(X) = \lambda$, $x \in \{0, 1, 2...\}$ and $\lambda \in R_+$.

If X is a random variable which follows a *Geometric* distribution with paratmeter θ, $Geometric(\theta)$, then its probability function is:

$$p(x) = \theta(1-\theta)^x$$

with $E(X) = \frac{1-\theta}{\theta}$, $V(X) = \frac{1-\theta}{\theta^2}$, $x \in \{0, 1, 2...\}$ and $\theta \in [0, 1]$.

SOLUTION

El enunciado nos pide seleccionar uno de los siguientes modelos:

$$M_1 = \{p_1(y|\lambda) = Poisson(\lambda); \lambda \in (0, \infty); \pi_1(\lambda)\}$$
$$M_2 = \{p_2(y|\theta) = Geometric(\theta); \theta \in (0, 1); \pi_2(\theta)\}$$

Primero hay que definir los modelos bayesianos de cada hipótesis/modelo, por ello hay que especificar las distribuciones a priori. El enunciado nos da información a priori respecto al volor esperado que estará con alta probabilidad entre 4 y 6.

Para el modelo M_1 el parámetro λ corresponde precisamente al valor esperado, así que escogeromos una distribución a priori para λ con alta probabilidad de tomar valores entre 4 y 6, como por ejemplo una Gamma de parametros 100 y 20.

Para el modelo M_2, dado que el parámetro $\theta \in (0, 1)$, podemos escoger como distribución a priori una *Beta*. Sabemos que si $y|\theta \sim Geometric(\theta)$ entonces $E(X) = \frac{1-\theta}{\theta}$, de modo que a partir de la igualdad $5 = \frac{1-\theta}{\theta}$ deducimos que el valor esperado the θ deberá estar alrededor de 1/6. Así que se trata de escoger los valores de los parámetros de la distribución *Beta* que den lugar a que el valor esperado de y esté entre los valores 4 y 6 con alta probabilidad. A través de pureba y error escogemos como distribuci?n a priori para θ una $Beta(100, 500)$.

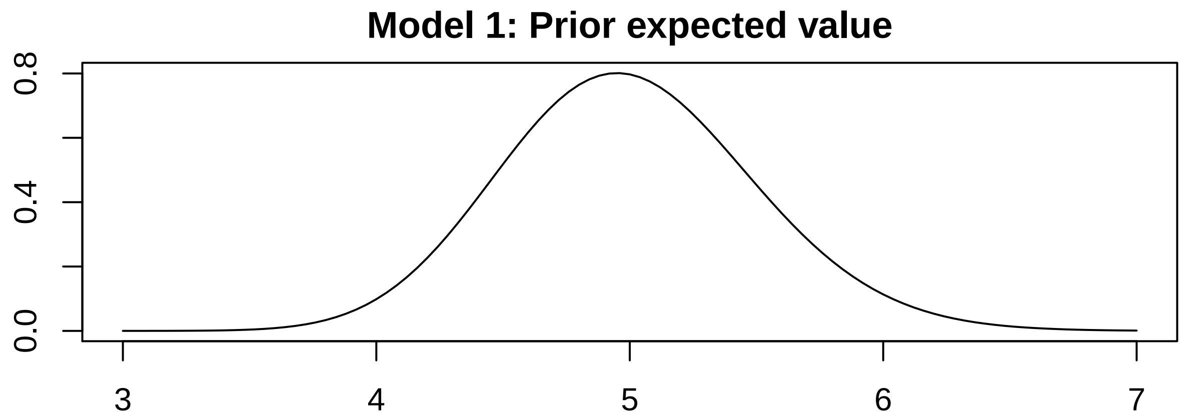
Siempre es muy recomendable dibujar las distribuciones a priori.

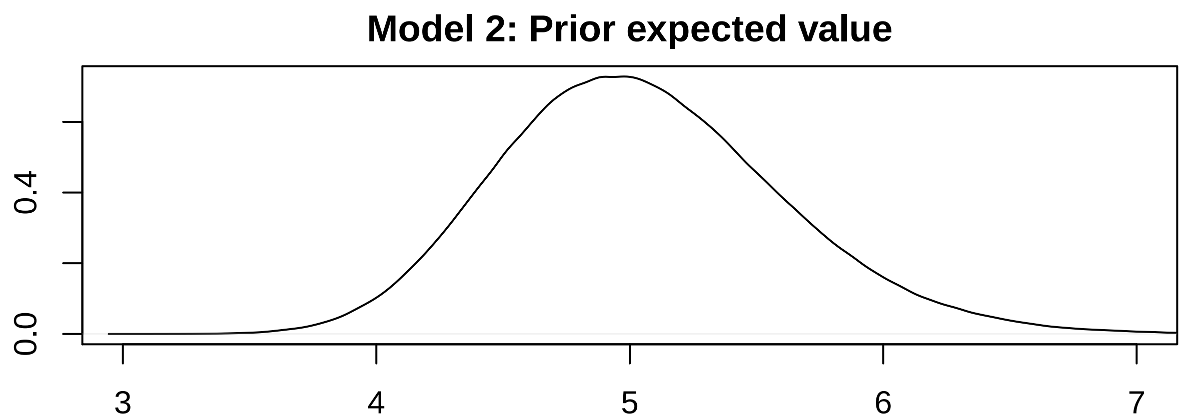
```r
# poisson
a1 <- 100; b1 <- 20

# geometric
a2 <- 100; b2 <- 500


M <- 1000000
th <- rbeta(M,a2,b2)
e.y <- (1-th)/th
#mean(th)

par(mfrow=c(2,1), mar=c(2,2,2,2))
plot(function(x)dgamma(x,a1,b1),xlim=c(3,7),main="Model 1: Prior expected value", ylab="", xlab="")
plot(density(e.y),xlim=c(3,7),main="Model 2: Prior expected value", ylab="", xlab="")
```





El enunciado no especifica que ningún modelo sea mas probable a priori que el otro, así que la probabilidad a priori para cada modelo será 0.5.

En resúmen tenemos que seleccionar uno de los siguientes modelos:

$$M_1 = \{y|\lambda \sim Poisson(y|\lambda); \lambda \sim Gamma(100, 20)\}$$
$$M_2 = \{y|\theta = Geometric(\theta); \theta \sim Beta(100, 500)\}$$

donde $p(M_1) = p(M_2) = 0.5$.

Una vez recogidos los datos:

```r
y <- c(1,2,2,8,10)
n <- length(y)
```

hay que calcular la probabilidad a posteriori para cada hipótesis y escoger aquella con la probabilidad más alta. Así, tenemos que caluclar:

$$P(M_1|y) = \frac{P(M_1)P(y|M_1)}{P(M_1)P(y|M_1) + P(M_2)P(y|M_2)}$$

y

$$P(M_2|y) = \frac{P(M_2)P(y|M_2)}{P(M_1)P(y|M_1) + P(M_2)P(y|M_2)},$$

donde

$$P(y|M_1) = \int_0^\infty Poisson(y|\lambda)Gamma(\lambda|a=100, b=20)d\lambda =$$

$$= \int_0^\infty \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}d\lambda = ... = \frac{1}{\prod_{i=1}^n y_i!}\frac{b^a}{\Gamma(a)}\frac{\Gamma(a+\sum_{i=1}^n y_i)}{(b+n)^{a+\sum_{i=1}^n y_i}}$$

$$P(y|M_2) = \int_0^1 Geometric(y|\theta)Beta(\theta|a=100, b=500)d\theta$$

$$= \int_0^1 \prod_{i=1}^n \theta(1-\theta)^{y_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{(b-1)}d\theta = ... = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+n)\Gamma(b+\sum_{i=1}^n y_i)}{\Gamma(a+b+n+\sum_{i=1}^n y_i)}$$

```r
#H1:poisson(lam); lam ~ Gamma(100,20), p(H1)=0.5
#H2:Geometric(th); th ~ Beta(100,500), p(H2)=0.5

dnbinom <- function(y, a, b) {
 n <- length(y)
 dens <-    a*log(b) - lgamma(a) - sum(lfactorial(y)) +
       lgamma(a+sum(y)) - (a+sum(y))*log(b+n)
 dens <- exp(dens)
 return(dens)
}

dbgeom <- function(y, a, b) {
 n <- length(y)
 dens <-    lgamma(a+b) - lgamma(a) - lgamma(b) +
       lgamma(a+n) + lgamma(b + sum(y)) -
       lgamma(a+b+n+sum(y))
 dens <- exp(dens)
 return(dens)
}

p.y.H1 <- dnbinom(y,a1,b1)
p.y.H2 <- dbgeom(y,a2,b2)

H1.y <- round(p.y.H1/(p.y.H1 + p.y.H2),3)
H2.y <- round(p.y.H2/(p.y.H1 + p.y.H2),3)

#cat(" probabilidad a posteriori M1:poisson(lamb); lamb ~ Gamma(100,20), p(H1)=0.5 es  ",
#    round(p.y.H1/(p.y.H1 + p.y.H2),2) ,"\n",
#"probabilidad a posteriori M2:Geometric(th); th ~ Beta(100,500), p(H2)=0.5   es  ",
#round(p.y.H2/(p.y.H1 + p.y.H2),2),"\n")
```

Después de realizar los cálculos obtenemos que $P(M_1|y) = 0.12$ y que $P(M_2|y) = 0.88$, por lo tanto es más probable que los datos hayan sido generados por un modelo de *Poisson* que no por un modelo *Geometric*.

Finalmente queremos realizar una predicción para un nuevo valor utilizando la técnica del *model averaging*, calculando la predictiva a posteriori como:

$$p(\tilde{y}|y) = p(M_1|y)p_1(\tilde{y}|y) + p(M_2|y)p_2(\tilde{y}|y).$$

Es decir la utilizar como distribución predcitiva a posteriori una mixtura de las predicitivas a posteriori de cada modelo ponderando por su probabilidad a posteriori. Un modo de hacerlo es a través de simulación:

para $m = 1, ..., M$ repetimos

1. simulamos u de una distribución $Uniforme(0, 1)$,
2. si $u < p(M_1|y)$ simulamos $\tilde{y}^{(m)}$ de $p_1(\tilde{y}|y)$
3. si $u > p(M_1|y)$ simulamos $\tilde{y}^{(m)}$ de $p_2(\tilde{y}|y)$.

Utilizando éstas M simulaciones podemos dibujar su distribución, calcular intervalos de probabilidad mediante los percentiles y una estimación puntual mediante una medida de localización como puede ser la media o la mediana a posteriori.

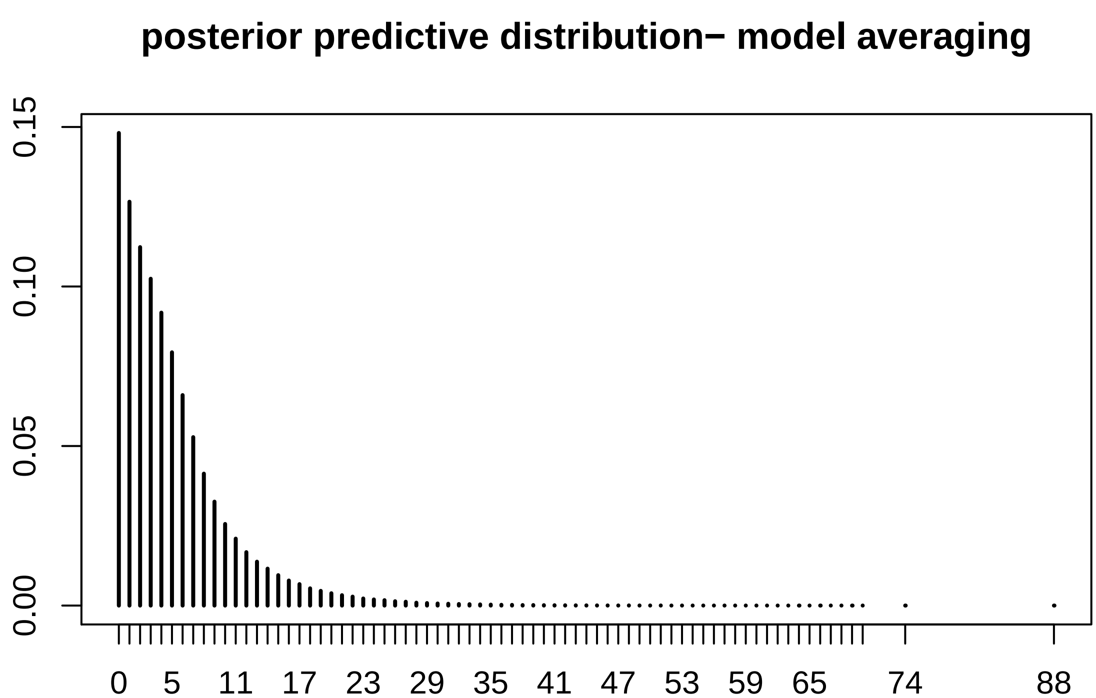
Para simular de las respectivas distribuciones predictivas a posteriori hay que calcular la distribución a posteriori de cada modelo, que al ser congugadas tenemos que $\lambda|y \sim Gamma(a + \sum y, b + n)$ y $\theta|y \sim Beta(a + n, b + \sum y)$.

```
post.pois <- rgamma(M,a1+sum(y), b1+n)
post.geom <- rbeta(M, a2+n, b2+sum(y))

pred.post.pois <- rpois(M,post.pois)
pred.post.geom <- rgeom(M,post.geom)
aux <- (runif(M)<(p.y.H1/(p.y.H1 + p.y.H2)))
TOT <- cbind(pred.post.pois,pred.post.geom,aux,pred.post.pois*aux+pred.post.geom*(1-aux))
```

```
par(mfrow=c(1,1))

plot(table(TOT[,4])/M, ty="h", ylab="", main="posterior predictive distribution- model averaging")
```



```
IC.l <- quantile(TOT[,4], c(0.025))
IC.u <- quantile(TOT[,4], c(0.975))
mp <- round(mean(TOT[,4]),2)
```

Así, un intervalo de credibilidad del 95% será (0,19), y la media a posteriori es 5.01.