

Cas pràctic del repte 2. Whatsappejant amb Poisson

Anabel Forte, Gonzalo García-Donato i Joaquín Martínez-Minaya

Índex

1	Introducció: Whatsappejant amb Poisson	2
2	Model de Poisson amb prèvia gamma	2
2.1	Versemblança de Poisson	3
2.2	Distribució a priori: gamma	4
2.3	Distribució a posteriori: gamma	6
2.4	Distribució predictiva: gamma-Poisson	8

1 Introducció: Whatsappejant amb Poisson

WhatsApp és una de les aplicacions de missatgeria instantània més utilitzades i va ser creada per Jan Koum i Brian Acton. Aquesta aplicació permet comunicar-nos amb qualsevol persona des de qualsevol lloc del món i de forma immediata. Avui dia, qui no coneix WhatsApp? L'ús de WhatsApp, juntament amb altres xarxes socials, s'està popularitzant cada vegada més entre els joves d'entre deu i catorze anys.

En aquest cas pràctic, intentarem estimar el nombre de missatges de WhatsApp que rep un adolescent durant una hora en un dia no lectiu. Per a això, es va preguntar a vint-i-cinc adolescents sobre el nombre de missatges que rebien a través d'aquesta aplicació en una hora. A continuació, es mostren els resultats d'aquests recomptes.

```
# set.seed(100)
# data <- data.frame(Wp = rpois(25, lambda = 140))
data <- data.frame(Wp = c(134, 121, 150, 141, 143, 133,
                        145, 135, 141, 141, 137,
                        154, 133, 135, 154, 143, 126,
                        149, 143, 149, 130, 129, 142, 126, 130))
```

2 Model de Poisson amb prèvia gamma

Atès que estem interessats en el nombre de missatges de WhatsApp que rep un adolescent en una hora, assumim que la nostra variable resposta (Y) segueix una distribució de **Poisson** de paràmetre λ , és a dir:

$$Y \mid \lambda \sim Po(\lambda),$$

i la seva distribució de probabilitat, com s'ha vist a la teoria, és:

$$p(Y = y \mid \lambda) = \frac{e^{-\lambda} \lambda^y}{y!},$$

sent els seus dos primers moments exactament iguals.

- $E(Y) = \lambda$
- $\text{Var}(Y) = \lambda$

2.1 Versemblança de Poisson

En l'apartat anterior hem definit la nostra variable d'interès y , amb ella, la distribució de Poisson. No obstant això, quan fem una anàlisi estadística disposem de mesures y , dit d'una altra manera, de realitzacions d'aquesta variable aleatòria. Per extreure informació sobre aquestes mesures i poder fer inferència sobre el paràmetre d'interès λ necessitem calcular la funció de versemblança.

Així, donat un conjunt d'observacions y_i per $i = 1, \dots, 25$, la funció de versemblança Poisson és proporcional a

$$L(\lambda \mid \mathcal{D}) \propto e^{-n\lambda} \lambda^r,$$

on $r = \sum_{i=1}^n y_i$, és a dir, r és el nombre total de missatges de WhatsApp que reben els vint-i-cinc adolescents en una hora. És important adonar-se que estem utilitzant el signe $\propto$ per indicar que les dues quantitats són iguals, a falta d'una constant. En concret, la igualtat, com s'ha vist a la teoria, s'aconsegueix dividint pel productori del factorial de cada y_i .

L'expressió anterior es pot calcular en R amb el codi següent:

- Definim l'expressió de la versemblança.

```
y <- data$Wp
r <- sum(y)
Lpois <- function(y, lambda){
  prod(dpois(y, lambda = lambda))
}
```

- Creem una variable auxiliar per a l'eix de les x o, dit d'una altra manera, el paràmetre del qual depèn la funció de versemblança, que en aquest cas és λ . A continuació, calculem el valor de la versemblança en aquests punts.

```
aux <- seq(100, 165, by = 0.01)
aux2 <- sapply(aux, FUN = Lpois, y=y)
```

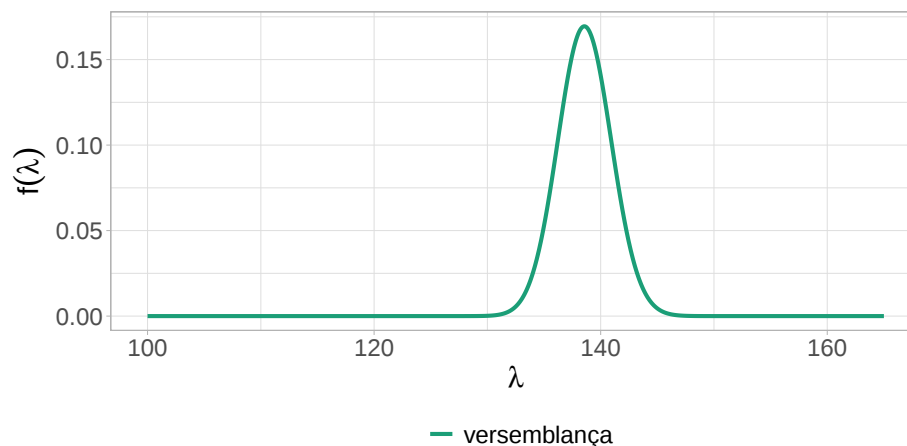
- Ara bé, sabem que la versemblança en si no és una distribució de probabilitat i, per tant, no ha d'integrar 1. En aquest cas, farem que integri 1 per poder comparar-la amb la distribució a priori i la distribució a posteriori.

```
C <- 1 / (0.01 * sum(aux2))
```

- Finalment, la dibuixem utilitzant el paquet de R `ggplot2`.

```
library(ggplot2)
data_vero <- data.frame(x = aux, y = C*aux2, class = "versemblança")

ggplot(data = data_vero) +
  geom_line(aes(x = x, y = y, color = class, linetype = class),
            size = 1.2) +
  scale_linetype_manual(values=c("solid"))+
  scale_color_brewer(palette="Dark2") +
  ylab(expression(f(lambda))) +
  xlab(expression(lambda)) +
  theme_light() +
  theme(axis.text = element_text(size=15),
        axis.title = element_text(size=18),
        legend.text = element_text(size=15)) +
  theme(legend.position = "bottom", legend.title = element_blank())
```



Ja tenim cobert un dels elements necessaris per fer inferència en el context bayesià, la **versemblança**. És el moment de definir la distribució a priori.

2.2 Distribució a priori: gamma

Sabem que λ és un paràmetre que pren valors en l'interval $(0, \infty)$, i atès que la distribució **gamma** és la seva distribució a priori conjugada, la utilitzarem per assignar informació

prèvia sobre el paràmetre λ .

Així, considerem la distribució a priori $\lambda \sim \text{gamma}(a, b)$, amb funció de densitat:

$$\pi(\lambda \mid a, b) = \lambda e^{-b\lambda} \frac{(b\lambda)^a}{\Gamma(a)} \lambda^{a-1}.$$

$\Gamma(\cdot)$ representa la funció gamma, que per a qualsevol enter positiu es defineix com a $\Gamma(a) = (a-1)!$. Els primers dos moments d'aquesta distribució són:

- $E(\lambda) = \frac{a}{b}$
- $Var(\lambda) = \frac{a}{b^2}$

Atès que estem utilitzant la funció gamma com a distribució a priori, els paràmetres a i b seran els encarregats de resumir la **informació prèvia** que tenim sobre λ . Vegem com podem establir aquests dos paràmetres:

1. Podem pensar en un adolescent proper a nosaltres, o fins i tot en nosaltres mateixos. Quants missatges de WhatsApp rebem en una hora en un dia no lectiu?
2. Quina confiança atribuïm a aquesta xifra? Això determina com de gran és l'error que cometem, és a dir, determina la variabilitat de la distribució a priori.
3. Un cop hem fixat el **nombre** i l'**error**, calculem a i b de la mitjana i la variància de la distribució gamma. Per exemple, si pensem que un adolescent rep 128 missatges amb un error de 60 (una desviació estàndard de 4 més o menys), obtindrem els valors de a i b resolent el següent sistema d'equacions:

$$a/b = 128$$

i

$$a/b^2 = 16.$$

Aïllant, $a = 128b$, $128 = 16b$. Per tant, obtenim que $b = 8$ i $a = 128 \cdot 8 = 1024$.

Representem, a continuació, el gràfic anterior juntament amb la distribució a priori.

```
a0 <- 1024
b0 <- 8

data_prior <- data.frame(
  x = aux,
  y = dgamma(aux, shape = a0, rate = b0),
```

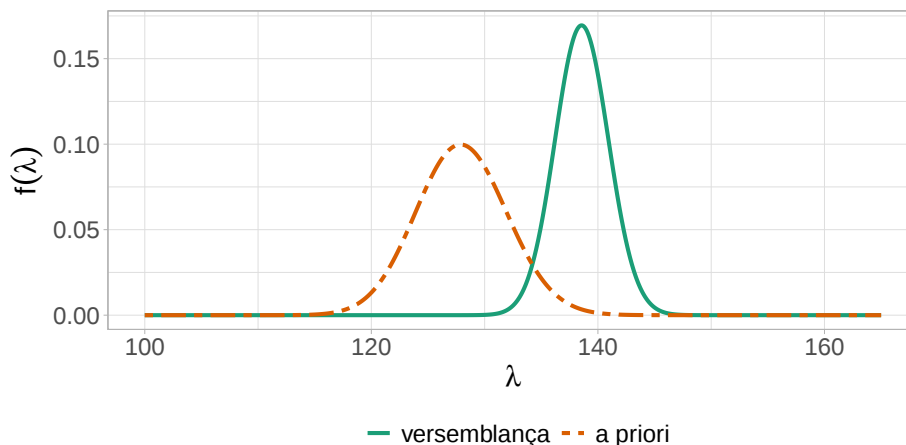
```

class = "a priori")

data_plot <- rbind(data_vero, data_prior)

ggplot(data = data_plot) +
  geom_line(aes(x = x, y = y, color = class, linetype = class),
            size = 1.2) +
  ylab(expression(f(lambda))) +
  xlab(expression(lambda)) +
  theme_light() +
  theme(axis.text = element_text(size=15),
        axis.title = element_text(size=18),
        legend.text = element_text(size=15)) +
  scale_linetype_manual(values=c("solid", "twodash"))+
  scale_color_brewer(palette="Dark2") +
  theme(legend.position = "bottom", legend.title=element_blank())

```



Ja hem definit un dels elements característics de l'estadística bayesiana, aquell que ens permet incorporar informació prèvia sobre els paràmetres en el model, la distribució a priori. Vegem com es calcula la distribució a posteriori.

2.3 Distribució a posteriori: gamma

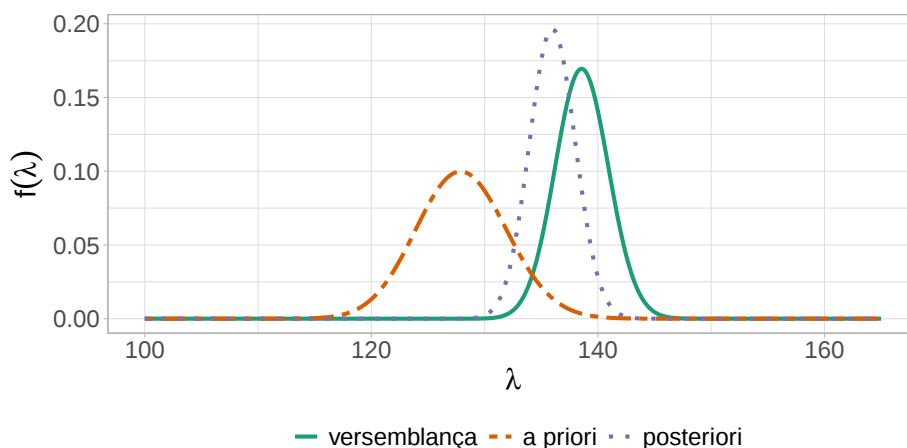
Com que estem fent una anàlisi en què la distribució a priori és conjugada, la distribució a posteriori serà de la mateixa naturalesa que la distribució a priori, és a dir, serà una distribució **gamma** amb paràmetres $a + r$ i $b + N$.

Gràficament, la podem representar de la manera següent:

```
aP <- a0 + r
bP <- b0 + length(y)

data_posterior <- data.frame(
  x = aux,
  y = dgamma(aux, shape = aP, rate = bP ),
  class = "posteriori")

data_plot <- rbind(data_vero, data_prior, data_posterior)
ggplot(data = data_plot) +
  geom_line(aes(x = x, y = y, color = class, linetype = class),
    size = 1.2) +
  ylab(expression(f(lambda))) +
  xlab(expression(lambda)) +
  theme_light() +
  theme(axis.text = element_text(size=15),
    axis.title = element_text(size=18),
    legend.text = element_text(size=15)) +
  scale_linetype_manual(values=c("solid", "twodash", "dotted"))+
  scale_color_brewer(palette="Dark2") +
  theme(legend.position = "bottom", legend.title=element_blank())
```



Podem observar que la distribució a posteriori es troba localitzada entre la distribució a priori i la versemblança. Si tinguéssim més dades, la distribució a posteriori s'acostaria

més a la versemblança, i si utilitzéssim una distribució a priori més restrictiva, aquesta distribució a posteriori s'acostaria més a la distribució a priori.

Un cop tenim dibuixada la distribució a posteriori podem calcular la seva mitjana:

```
aP/bP
```

```
## [1] 136
```

o un interval de credibilitat al 95%.

```
qgamma(0.025, aP, bP)
```

```
## [1] 132.0499
```

```
qgamma(0.975, aP, bP)
```

```
## [1] 140.0075
```

Fins ara hem fet inferència sobre el paràmetre del model, λ . Ara bé, ens podríem preguntar, tot escollint un adolescent a l'altzar, quants missatges de WhatsApp rebrà en un dia no lectiu? Per a això, hem d'obtenir la distribució predictiva.

2.4 Distribució predictiva: gamma-Poisson

Altres distribucions importants en el context bayesià, com ja sabem, són les distribucions predictives a priori i a posteriori, $p(y)$ y $p(y | \mathcal{D})$. La distribució predictiva a priori es calcula integrant sobre les distribucions a priori dels paràmetres:

$$p(Y = y) = \int_{\Theta} p(y | \lambda) p(\lambda) d\lambda,$$

mentre que la distribució predictiva a posteriori es calcula integrant sobre les distribucions a posteriori dels paràmetres:

$$p(Y = y | \mathcal{D}) = \int_{\Theta} p(y | \lambda) p(\lambda | \mathcal{D}) d\lambda.$$

En el cas particular de gamma-Poisson, la distribució predictiva a posteriori del total de recomptes en n noves observacions (r) és una distribució coneguda com a **gamma-Poisson**.

La gamma-Poisson és una distribució de probabilitat per a una variable discreta r amb suport i els nombres naturals, incloent-hi el 0. A més, depèn de tres paràmetres: a , b i N .

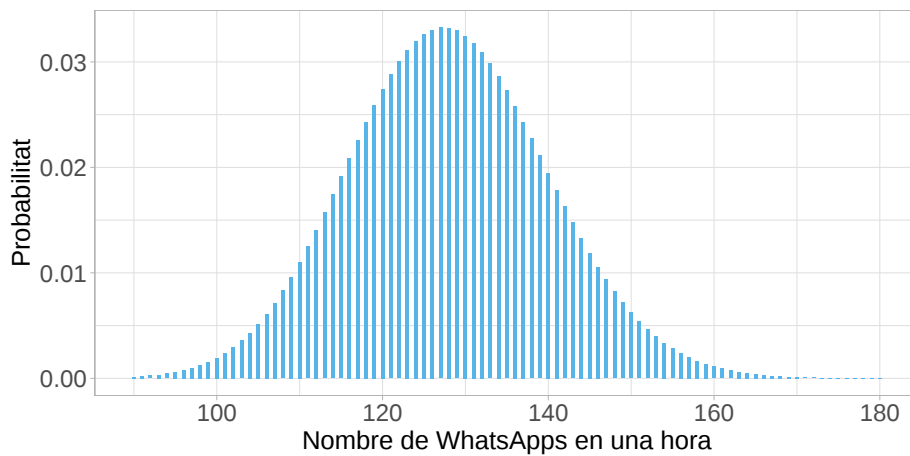
La seva funció de probabilitat és:

$$p(r | a, b, N) = \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + r)}{r!} \frac{N^r}{(b + N)^{a+r}}.$$

En el cas en què estiguem interessats en l'obtenció de la distribució predictiva a priori, els paràmetres de la distribució gamma-Poisson vindran donats pels paràmetres de la distribució gamma a priori. Si, per contra, estem interessats en l'obtenció de la distribució predictiva a posteriori, els paràmetres de la distribució gamma-Poisson vindran donats per la distribució gamma a posteriori.

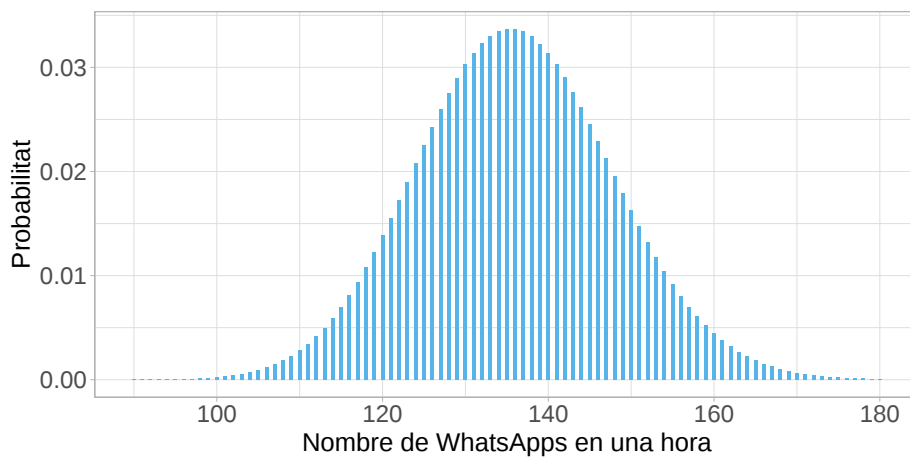
Així, la forma de la distribució predictiva a priori del nombre de missatges de WhatsApp que un adolescent rep en una hora queda de la manera següent:

```
dgammapois <- function(r, a, b, n){  
  C <- a * log(b) - lgamma(a)  
  C2 <- lgamma(a + r) + log(n) * r - lgamma(r + 1) - log(b + n) * (a + r)  
  exp(C+C2)  
}  
  
ggplot(data = data.frame(  
  x = 90:180,  
  y = dgammapois(90:180, a = a0, b = b0, n = 1)),  
  aes(x = x, y = y)) +  
  geom_bar(stat="identity", fill = "#56B4E9", width = 0.4) +  
  theme_light() +  
  xlab("Nombre de WhatsApps en una hora") +  
  ylab("Probabilitat") +  
  theme(axis.text = element_text(size=15),  
        axis.title = element_text(size=16),  
        legend.text = element_text(size=15))
```



Mentre que la distribució predictiva a posteriori és:

```
ggplot(data = data.frame(
  x = 90:180,
  y = dgammapois(90:180, a = aP, b = bP, n = 1)),
  aes(x = x, y = y)) +
  geom_bar(stat="identity", fill = "#56B4E9", width = 0.4) +
  theme_light() +
  xlab("Nombre de WhatsApps en una hora") +
  ylab("Probabilitat") +
  theme(axis.text = element_text(size=15),
        axis.title = element_text(size=16),
        legend.text = element_text(size=15))
```



Podem observar que la distribució predictiva es calcula integrant per a tots els possibles valors de λ . Per aquest motiu és diferent del fet d'utilitzar una distribució de Poisson usant

$E(\lambda)$ o $E(\lambda | i)$. En el gràfic següent podem veure les diferències.

```
data_pred_prior <- data.frame(
  x = seq(90, 180, 2),
  y = dgammapois(seq(90, 180, 2), a = a0, b = b0, n = 1),
  class = "pred_prior")
data_pred_post <- data.frame(
  x = seq(90, 180, 2),
  y = dgammapois(seq(90, 180, 2), a = aP, b = bP, n = 1),
  class = "pred_post")
data_pred_pois <- data.frame(
  x = seq(90, 180, 2),
  y = dpois(seq(90, 180, 2), aP / bP),
  class = "poisson")
data_pred_plot <- rbind(data_pred_prior, data_pred_post, data_pred_pois)

ggplot(data = data_pred_plot, aes(x = x, y = y, fill = class)) +
  geom_bar(stat="identity", position=position_dodge()) +
  xlab("Nombre de WhatsApps en una hora") +
  ylab("Probabilitat") +
  theme_light() +
  theme(axis.text = element_text(size=15),
        axis.title = element_text(size=16),
        legend.text = element_text(size=15)) +
  scale_fill_brewer(palette="Dark2") +
  theme(legend.position = "bottom", legend.title=element_blank())
```

