

**Especialização em Ciência de Dados - Instituto de Matemática - UFRJ**  
**Aprendizagem Estatística - Segundo Trabalho Aplicado**

Uma empresa de suporte técnico está fazendo um estudo sobre a duração de atendimentos a clientes e, para tal fim, registra o **número de clientes** atendidos, em intervalos com uma hora de duração. Uma amostra aleatória de contagens de atendimentos foi tomada sobre um longo período de tempo, com observações registradas em horas não consecutivas, de tal forma que se pode assumir independência entre as observações. O arquivo `atendimentos.csv` guarda 100 registros dessas contagens. Assuma que as contagens sigam distribuição Poisson( $\theta$ ) e que se adote um enfoque bayesiano para o processo inferencial, com descrição a priori sobre  $\theta$  dada por uma distribuição Gama( $\alpha, \lambda$ ).

- (a) Assuma, a princípio, que  $\alpha = \lambda = 0,001$ . Passaremos a denominar essa especificação de *priori 1*. Determine os seguintes sumários sobre a especificação subjetiva inicial sobre  $\theta$  (*priori1*): valor esperado ( $E[\theta]$ ) e desvio-padrão ( $DP[\theta]$ ). A especificação a priori é bastante informativa ou é vaga? Justifique.
- (b) Assuma que um segundo analista creia, a priori, que, em média, haja em torno de 5 atendimentos por hora. Perguntado sobre a probabilidade de que a média de atendimentos por hora seja superior a 10, ele responde que acredita que essa probabilidade seja próxima de 5%. Especifique valores para os parâmetros  $\alpha$ , e  $\lambda$ , de forma a traduzir essa informação a priori. Passamos a nos referir a essa segunda especificação como *priori 2*. As linhas de código abaixo podem ajudar nessa tarefa.

```
#Elicitação dos parâmetros da priori 2
lambda<-seq(0,5,0.1)\
alpha<-5*lambda\
matriz<-cbind(alpha,lambda,1-pgamma(10,shape=alpha,rate=lambda))
```

- (c) Determine valor esperado ( $E[\theta]$ ) e desvio-padrão ( $DP[\theta]$ ) de  $\theta$  sob a priori 2 e esboce, em um mesmo gráfico, as duas densidades a priori (use as linhas de código abaixo). A priori 2 é mais vaga ou mais informativa que a priori 1?

```
#Gráfico das prioris para uma grade de valores de theta entre 0 e 15
theta<-seq(0,15,0.1)
priori1<-dgamma(theta,shape=0.01, rate=0.01)
priori2<-dgamma(theta,shape=???, rate=???)
plot(theta, priori2)
lines(theta, priori1)
```

Agora, vamos estabelecer dois cenários de disponibilidade de dados: no primeiro cenário, assuma amostra grande, composta pelas 100 contagens de atendimentos disponíveis no arquivo *atendimento.csv*. Para a amostra pequena, tome um subconjunto aleatório contendo 10 dessas observações:

```
amostra.gr<-scan(file="atendimentos.csv")
amostra.pq<-sample(amostra.gr,10)
```

- (d) Esboce, em um mesmo gráfico, a função de verossimilhança, sob as duas amostras (ou seja, sob o cenário de dados abundantes e escassos). Comente o comportamento da função de verossimilhança sob as duas amostras. (Especifique  $n1$ ,  $n2$ ,  $t1$  e  $t2$  adequadamente e use o código abaixo).

```
theta<-seq(0,10,0.01)
vero.gr<-exp(-n1*theta)*theta^t1
vero.pq<-exp(-n2*theta)*theta^t2
plot(theta,vero.gr/max(vero.gr))
lines(theta,vero.pq/max(vero.pq))
```

(e) Passaremos agora a estudar o comportamento da distribuição a posteriori sob as duas especificações a priori, considerando os dois cenários: amostra grande e amostra pequena. Obtenha a distribuição a posteriori :

- sob priori 1 e amostra pequena;
- sob priori 2 e amostra pequena;
- sob priori 1 e amostra grande;
- sob priori 2 e amostra grande.

Esboce gráficos das distribuições a posteriori obtidas sob as quatro especificações acima e teça comentários sobre a variação em seu comportamento. (O código R abaixo pode ajudar na confecção dos gráficos. Cabe a você definir quem é a curva a posteriori, em cada caso).

```
##Preparação de uma janela gráfica particionada em duas linhas e duas colunas, para acomodar os 4 gráficos que  
par(mfrow=c(1,2))  
#Posteriori para theta sob amostra pequena e priori 1:  
plot(theta,poster.1.pq,type="l",ylab="n=10")  
#Posteriori para theta sob amostra pequena e priori 2:  
lines(theta,poster.2.pq,type="l",ylab="n=10",col=2)  
#Posteriori para theta sob amostra grande e priori 1:  
plot(theta,poster.1.gr,type="l",ylab="n=100")  
#Posteriori para theta sob amostra grande e priori 2:  
lines(theta,poster.2.gr,type="l",ylab="n=100",col=2)  
legend("topright",legend=c("Priori 1","Priori 2"),lty=c(1,1),col=c(1,2))
```