Introdução à Estatística Bayesiana - Lista 2

1) Seja

$$\begin{cases} y_1, ..., y_n|\theta \sim LogN(0, \theta), & \text{iid;} \\ \theta \sim IG(3, 8), & . \end{cases}$$

Como auxílio do Maple e sabendo que $\boldsymbol{x} = (1, 1, 2, 2, 4, 5, 4, 4, 6, 7)$.

a) Encontre a distribuição a posterior de θ. Faça o triplot (gráficos da distribuição a priori, a posteriori e função de verossimilhança).

b) Encontre a estimativa a posterior de θ sob perda quadrática, absoluta e zero-um. Compare esses estimadores com o E.M.V..

c) Encontre um intervalo de credibilidade a 95% para θ.

d) Utilize o Openbugs para rodar o modelo e compare os estimadores obtidos no item (b) com os obtidos por simulação.

d) Esboce a expressão da distribuição a posterior preditiva de uma nova observação $\tilde{y}$.

2) Seja $x_1, ..., x_n|\lambda \sim Poisson(\lambda)$ iid, e $\lambda \sim G(\alpha, \beta)$. Suponha que a amostra observada foi $\boldsymbol{x} = (6, 6, 2, 6, 5, 8, 3, 6, 4, 5)$

a) O responsável pela pesquisa admite total desconhecimento sobre λ. Qual das duas priores abaixo é a mais indicada? Justifique.

- Priori I: $\alpha = 1$ e $\beta = 0.001$.
- Priori II: $\alpha = 1$ e $\beta = 1000$.

b) Desenhe o triplot.

c) Encontre o estimador Bayesiano de λ.

d) Escreva um código no Openbugs para o modelo.

e) Compare as estimativas obtidas no Openbugs com aquelas obtidas no item (c).

f) Encontre um I.C.B. a 80% usando Openbugs e calculando pelo Maple.

g) Encontre a distribuição preditiva $P(\tilde{x}|\boldsymbol{x})$ e faça um gráfico.

h) Qual o valor futuro estimado e o erro associado a ele?

3) Em um estudo se deseja saber a proporção de pessoas que aprovam a atual administração municipal. Um cientista político acredita fortemente que *o valor mais provável para a proporção seja de* 5% e variância de 0,01. Uma amostra de tamanho $n = 500$ foi selecionada e observou-se que o número de pessoas que aprovam a atual administração é de 40 indivíduos. Um modelo Bernoulli-Beta foi sugerido.

a) Monte o modelo, especificando a verossimilhança e distribuição a priori. Especifique a distribuição a posterior e faça um gráfico. *Dica: Escolha a distribuição a priori mais coerente com a informação do especialista.*

b) Encontre o estimador Bayesiano para a proporção sob: perda quadrática, perda absoluta, e perda zero-um.

c) Encontre um intervalo de credibilidade a 90% para a proporção.

d) Um modelo alternativo é o modelo Binomial-Kumaraswamy. Se $\theta \sim Kum(a, b)$, então sua f.d.p. é data por $f(\theta) = ab\theta^{a-1}(1 - \theta^a)^{b-1} \times I(0 < \theta < 1)$. Propriedade importante: Se $Z \sim Beta(1, b) \Leftrightarrow W = Z^{1/a} \sim Kum(a, b)$. O modelo proposto é $\sum_{i=1}^{500} X_i \sim Bin(500, \theta)$ e $\theta \sim Kum(2, 9)$. Com base nas informações acima escreva um modelo no `Openbugs` e encontre as estimativas a posteriori para a proporção sob perda quadrática e perda absoluta.

e) Em relação ao ítem (d), encontre uma estimativa para o valor preditivo.

f) Em relação ao ítem (d), também encontre um intervalo de credibildiade a 90%.

4) Um bairro de um certa cidade está localizado próximo a uma fábrica de amianto, o qual pode causar problemas de saúde, dentre eles problemas pulmonares. A secretaria de saúde deciciu fazer um estudo, selecionando 10 residências ao acaso, e em cada residência observou-se a proporção de pessoas com problemas respiratórios (fadiga, dificuldade em dormir, etc). Os dados foram: $\boldsymbol{x}$=(0.1,0.3,0.3,0.2,0.5,0.7,0.1,0.6,0.7,0.2). Um modelo Beta foi sugerido para os dados, com parâmetros α e β. Um estudo a priori mais detalhado sugeriu que $\alpha \sim U(10, 30)$ e $\beta \sim G(100, 3)$.

a) Justifique o uso de Openbugs para esse problema.

b) Usando o Openbugs, reuna evidências sobre a verdadeira distribuição de X, isto é quais os valores esperados de α e β?

c) Analise os gráficos das distribuições marginais de α e β, as estimativas a posteriori sob perda quadrática e sob perda absoluta conincidem? Caso não, justifique.

d) Com base nas estimativas a posteriori, faça um gráfico (maple) da distribuição dos dados aproximada, ela é razoavelmente simétrica?

e) Com base nessa distribuição aproximada para X, se uma residência é selecionada ao acaso, qual a probabilidade da proporção X, de pessoas com problemas, ser superior a 50%? *Use o Maple.*

f) Especifique os intervalos de credibilidade de α e β a 95%, obtidos no Openbugs.

5) Em uma carregamento de frutas, estamos interessados em estimar a proporção θ de frutas estragadas. Para isso extraiu-se uma a.a. de tamanho 20 (com reposição) e verificou-se se cada fruta estava estragada. Seja Y o número de frutas estragadas na amostra, da qual observou-se $y = 8$. Por simplicidade, o comprador do carregamento trabalha somente com *três* possibilidades para θ: $\theta = 0.1$, $\theta = 0.3$ e $\theta = 0.5$. De observações de carregamentos anteriores, tem-se que $P(\theta = 0.1) = 0.5$, $P(\theta = 0.3) = 0.3$ e $P(\theta = 0.5) = 0.2$. Assim, o problema consistirá em decidir sobre as três possibilidades acima. Um estudo foi feito do quanto se perderá quando uma decisão errada for tomada, chegou-se aos seguintes valores:

onde $d_1 : \{\theta = 0.1\}$, $d_2 : \{\theta = 0.3\}$, e $d_2 : \{\theta = 0.5\}$.

a) Modele o problema convenientemente. Um modelo Binomial-Beta pode ser usado?

b) Qual hipótese parece ser a melhor, levando em consideração os dados e a informação a priori?

	d_1	d_2	d_3
$\theta = 0.1$	0	1	3
$\theta = 0.3$	2	0	2
$\theta = 0.5$	3	1	0

c) Qual hipótese parece ser a melhor, levando em consideração os dados, a informação a priori e a perda?

d) Seria viável o teste usual para proporção puramente frequentista para testar as hipóteses sugeridas? Justifique.

e) Considerando os dados, a informação a priori e a perda, repita a análise Bayesiana supondo total ignorância sobre θ. As decisões mudam? *Dica: Use uma priori uniforme.*

6) Um estudo visa estimar a taxa de pessoas obesas na população, para isto uma amostra aleatória de tamanho 100 foi obtida, onde se anotou se o indivíduo tinha IMC>30 (índice de massa corporal) ou não. Dos 100 indivíduos, 15 foram classificados como obsesos. Por outro lado, um especialista julga que a amostra não é representativa, que pode não representar o verdadeiro perfil da população. Em suas pesquisas sobre obesidade, o mesmo especialista sugere que a taxa de obesos é muito maior, algo em torno de 40%, porém esta afirmação também carrega incerteza, o especialista afirma então que a margem de erro em volta do valor de 40% é simétrica e que tem 95% de certeza que a verdadeira taxa não ultrapassa 60%. Proponha um modelo (espeficiando a verossimilhança, priori e posteriori) para combinar os dados amostrais com a opinião do especialista. Em particular, estime a proporção de obesos (perda quadrática, absoluta e 0–1), e a variância a posteriori.

7) Problema: Uma planta é capaz de converter amônia em ácido nítrico. Durante o processo de oxidação existe uma perda A velocidade com que isso acontece está condicionada a três fatores: fluxo de ar na planta, temperatura da água e acidez da planta. Os dados estão organizados no R: `stack.loss` e `stack.x`:

Y=`stack.loss`: 10 vezes o percentual de amônia que escapa durante o processo. Esta é uma medida da eficiência da planta.

X_1=`Air.Flow`: fluxo de ar na planta

X_2=`Water.Temp`: temperatura da água

X_3=`Acid.Conc.`: acidez da planta.

Tipicamente, um modelo de regressão linear múltipla é escrito como

$$\begin{cases} y_i|\mu_i, \sigma^2 \sim f(y_i|\mu_i, \sigma^2), \ i = 1, ..., 21 \\ \mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \\ \beta_k \sim t_{(4)}(0, 100), \ k = 1, 2, 3 \\ \sigma^2 \sim IG(a, b) \end{cases}$$

a) Ajuste um modelo Normal de regressão linear múltipla para o problema. Para os coeficientes de regressão assuma que estão distribuídos de acordo com uma t de Student com média *zero* e variância 100, além disso assuma que a priori σ^2 tem distribuição Gama-Inversa com média 0.5 e variância 10 . Especifique todos os parâmetros do modelo.

b) Encontre as estimativas a posteriori dos coeficientes e da variância do erro.

c) Estabeleça intervalos de credibilidade para os parâmetros.

d) Avalie a possibilidade de nulidade dos parâmetros do problema. Você sugere que alguma variável deva ser retirada do modelo? Se sim, rode o modelo sem esta variável.

e) Análise de resíduos: Faça um gráfico dos resíduos e analise se estão relativamente baixos e se exitem outliers.

f) Rode novamente o modelo, agora supondo que não se tenha nenhum conhecimento a priori sobre σ^2, isto é $P(\sigma^2) \propto 1/\sigma^2$. Pesquise no `Openbugs` como declarar esse tipo de distribuição. Compare as estimativas deste modelo com o modelo acima.

g) Qual a probabilidade a posterior de σ^2 ser superior a 1?

h) Segundo os resíduos, esse modelo é melhor que o anterior? Use o gráfico de resíduos e a soma quadrada de resíduos.

i) Quais os valores previstos para $\boldsymbol{x}_1 = c(50, 56, 70)$, $\boldsymbol{x}_2 = c(20, 22, 23)$ e $\boldsymbol{x}_3 = c(80, 82, 91)$? Especifique o erro associado a essas previsões.

8) Um estudo visa estimar a proporção θ de mulheres que nascerão durante 2015 no estado do Ceará. Para isso sugere-se usar os dados da última PNAD (Pesquisa Nacional por Amostra de Domicílios). Como informação a priori, sugere-se analisar, através de um modelo de regressão simples, a evolução da proporção de nascimentos do sexo feminino ao longo das últimas PNADs (de 1990 até o mais recente disponível), usar o valor estimado para 2015 como sendo a média a priori e como variância a priori, a variância da estimativa obtida no modelo de regressão. *Dica: use alguma rotina de otimização para obter os hiperparâmetros estimados.* Feito isso, responda: **8)** Um estudo visa estimar a proporção θ de mulheres que nascerão durante 2015 no estado do Ceará. Para isso sugere-se usar os dados da última PNAD (Pesquisa Nacional por Amostra de Domicílios). Como informação a priori, sugere-se analisar, através de um modelo de regressão simples, a evolução da proporção de nascimentos do sexo feminino ao longo das últimas PNADs (de 1990 até o mais recente disponível), usar o valor estimado para 2015 como sendo a média a priori e como variância a priori, a variância da estimativa obtida no modelo de regressão. *Dica: use alguma rotina de otimização para obter os hiperparâmetros estimados.* Feito isso, responda:

a) Qual a proporção mais provável de nascimentos do sexo feminino para 2015?

b) Há muita incerteza sobre essa estimativa?

c) Qual a probabilidade de nascerem entre 50% e 60% de homens?

d) O governo irá destinar uma parte do orçamento para um programa de saúde de bebês do sexo feminino. O governo calcula um gasto de cerca de $R\$500,00$ por bebê. Como o orçamento precisa ser aprovado antes e é necessário ter uma estimativa da proporção de nascimentos do sexo feminino para 2015. É necessário decidir entre três hipóteses: $H_1)$ $\theta \leq 50\%$, $H_2)$ $50\% < \theta < 60\%$ e $H_3)$ $60\% \leq \theta \leq 70\%$.

   i) Qual dessas hipóteses é mais provável?

   ii) Caso seja preciso decidir entre H_2 e H_3, levando em consideração a perda (em reais) por se tomar a decisão errada, qual hipótese deve ser considerada?