ME 705A - Inferência Bayesiana
Segundo semestre de 2013
Prova III
Data: 27/11/2013

Nome: _____ RA: _____

Leia atentamente as instruções abaixo:

- **Coloque seu nome completo e RA em todas as folhas que você recebeu, inclusive nesta.**
- **Utilize somente um dos lados de cada folha.**
- **Leia atentamente cada uma das questões.**
- **Enuncie, claramente, todos os resultados que você utilizar.**
- **Justifique, adequadamente, seus desenvolvimentos, sem, no entanto, escrever excessivamente.**
- **O(a) aluno(a) só poderá sair da sala após as 10h30, mesmo que já tenha finalizado a prova. Após a saída do(a) primeiro(a) aluno(a) não será permitido a entrada de nenhum(a) outro(a) aluno(a).**
- **Não é permitdo empréstimo de material.**
- **Não serão dirimidas dúvidas de quaisquer natureza, após os 20 minutos iniciais.**
- **Resolva a prova, preferencialmente, à caneta, e procure ser organizado(a). Se fizer à lápis, destaque, à caneta, sua resposta.**
- **Utilize somente um lado de cada folha de resolução. Além disso, inicie a resolução da questão seguinte em outra folha e siga a ordem dos itens, em cada questão.**
- **Contestações a respeito da nota, só serão consideradas se estiverem por escrito.**
- **A nota do aluno(a) será $\frac{NP}{NT} \times 10$, em que NP é o número de pontos obtidos na prova e NT é o numero total de pontos da prova.**

- **Os resultados numéricos finais devem ser apresentados com duas casas decimais, apenas.**

- **A prova terá duração de 120 minutos, das 10h às 12h, improrrogáveis.**

Faça uma excelente Prova!!

1. Seja uma amostra aleatória de tamanho n de $X|(r,\lambda) \sim \text{gama}(r,\lambda), r > 0, \lambda > 0$. Responda os itens abaixo:

   a) Considere r conhecido e suponha que desejamos testar $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$, com $\lambda_0 > 0$, conhecido. Suponha que:

   $$p(\lambda) = \alpha \mathbb{1}_{\{\lambda_0\}}(\lambda) + (1-\alpha) p_1(\lambda) \mathbb{1}_{\Theta_1}(\lambda),$$

   em que $\alpha \in (0,1)$ é conhecido, $\Theta_1 = (0,\infty) - \{\lambda_0\}$ e $p_1(\lambda) = e^{-1/\lambda}$. Obtenha o fator de Bayes para testar as hipóteses acima (150 pontos).

   b) Considere $r = 1$, $n = 10$, $\lambda_0 = 1$ e $\overline{x} = 1,3$. Qual sua conclusão à respeito das hipóteses, através do fator de Bayes? Justifique, adequadamente, sua resposta (50 pontos).

   c) Considere agora que r é desconhecido. Admita que: $p(r,\lambda) = p(r)p(\lambda)$, em que

   $$p(r) \propto r^{\alpha-1} e^{-r/\beta} \mathbb{1}_{(0,\infty)}(r) \text{ e } p(\lambda) \propto \lambda^{-\phi-1} e^{-\gamma/\lambda} \mathbb{1}_{(0,\infty)}(\lambda),$$

   como $(\alpha, \beta, \phi, \gamma)'$, conhecidos. Encontre as duas distribuições condicionais completas, indentificando cada uma delas, sempre que ela corresponder à uma distribuição "conhecida" (100 pontos).

   d) Os dados analisados foram extraídos do censo do IBGE de 2000, e correspondem à renda média mensal (em reais) do chefe ou chefes do domicílio das 27 unidades da federação. O objetivo é estimar a renda média mensal das 27 unidades da federação e para isso o modelo gama em questão foi utilizado. Considerou-se os dois parâmetros desconhecidos e as prioris apresentadas no item c), com $\alpha = 0,01, \beta = 0,01^{-1}$, $\phi = \gamma = 0,01$. Os resultados a seguir, Tabelas 1 e 2, são oriundos de uma amostra válida de 1000 valores, para cada posteriori, obtidas através de um algoritmo MCMC implementado no programa WinBUGS. Assuma que a convergência (portanto, a amostra válida) foi obtida considerando-se valores apropriados para o burn-in, espaçamento e número total de valores simulados. Também foi ajustado um modelo exponencial, com as mesmas quantidades (priori, burn-in, espaçamento e número total de valores simulados). Com base nos resultados apresentados, qual dos dois modelos você escolheria para analisar os dados? Justifique, da forma mais completa possivel, sua conclusão (100 pontos).

Tabela 1: Estimativas bayesianas para os parâmetros do modelo gama: Questão 1

Parâmetro	EAP	DPAP	$IC_B(95\%)$	HPD(95%)
r	8,76	2,42	[4,71 ; 14,06]	[4,47 ; 13,59]
λ	81,54	24,44	[46,00 ; 139,21]	[44,38 ; 130,30]

Tabela 2: Estatística de comparação de modelos: Questão 1

	Modelo	
Estatística	gama	exponencial
Deviance	367,7	405,5
p_D	2,1	1,1
DIC	369,7	406,6

2. Os resultados abaixo se referem à uma análise bayesiana através de um algoritmo MCMC (utilizando o programa WinBUGS), de um conjunto de dados relativo à mortalidade de embriões de Biomphalaria Glabrata (hospedeiro da equistossomose), submetidos à um determinado tipo de extrato vegetal. Mais especificamente, observou-se do total de embriões (m_i) submetidos à determinada dose do extrato vegetal (x_i), quantos morreram (y_i), veja os dados na Tabela 3. Para isso, adotou-se o seguinte modelo (doravante modelo 1) $Y_i|(\beta_0, \beta_1) \overset{ind.}{\sim}$ Binomial$(m_i, p_i), i = 1, 2, ..., 7$, em que:

$$\text{logito}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1(x_i - \overline{x}), \overline{x} = \frac{1}{7}\sum_{i=1}^{7} x_i.$$

Além disso, considerou-se que $\beta_i \sim N(0, 200), i = 0, 1$, mutuamente independentes. O modelo 2 corresponde ao modelo 1 com $\beta_1 = 0$. Os resultados a seguir, Figuras 1 e 2 e Tabelas 4 e 5, são oriundos de uma amostra válida de 1000 valores, para cada posteriori, em cada modelo. Assuma que a convergência (portanto, a amostra válida) foi obtida considerando-se valores adequados para o burn-in, espaçamento e número total de valores simulados, para os dois modelos. Responda os itens abaixo:

a) Descreva, da forma mais completa possível, o comportamento das distribuições à posteriori. Você escolheria o EAP como estimativa pontual? Ainda com base no comportamento das posterioris, a diferença quase nula entre os IC'_Bs e os $HPD's$ é esperada? Justifique, adequadamente, todos os seus comentários (200 pontos).

b) Com base nos resultados, da forma mais completa possível, o que você pode falar sobre a significância dos parâmetros do modelo 1? A probabilidade de mortalidade de embriões aumenta, significativamente, com o aumento em uma unidade do extrato vegetal? De quanto é esse aumento? Justifique, adequadamente, todos os seus comentários. (200 pontos).

c) Qual dos dois modelos você escolheria para analisar os dados? O modelo 1 se ajustou bem aos dados? Justifique, da forma mais completa possivel, suas afirmações. (200 pontos).

Tabela 3: Dados de mortalidade dos embriões

m_i	y_i	x_i
50	4	0
50	5	15
50	14	20
50	29	25
50	38	30
50	41	35
50	47	40

Tabela 4: Estimativas bayesianas para os parâmetros do modelo 1: Questão 2

	EAP	DPAP	$IC_B(95\%)$	HPD(95%)
β_0	-0,11	0,14	[-0,40 ; 0,16]	[-0,37 ; 0,17]
β_1	0,16	0,02	[0,13 ; 0,19]	[0,13 ; 0,19]

Tabela 5: Estatística de comparação de modelos: Questão 2

	Modelo	
Estatística	1	2
Deviance	38,1	199,3
p_D	2,0	0,9
DIC	40,1	200,2

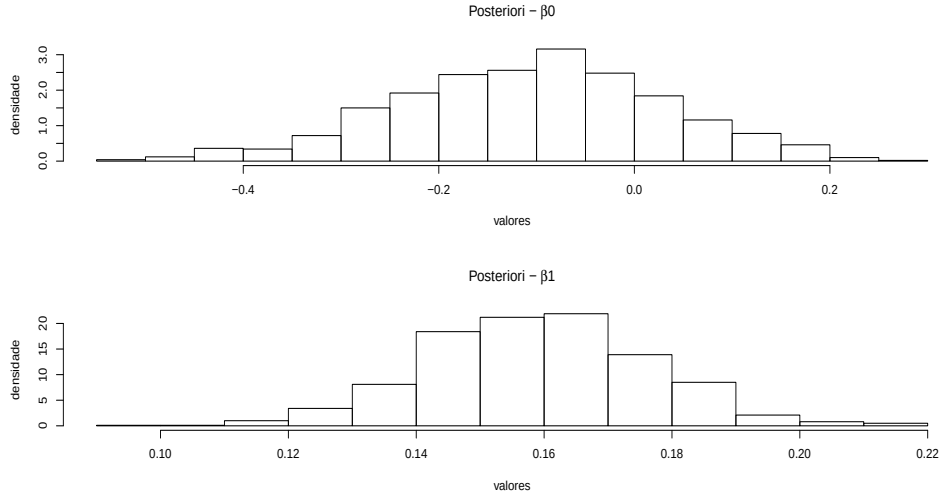

Figura 1: Histograma das posterioris (amostra válida) para cada parâmetro do modelo 1, utilizando um conjunto de cadeias: Questão 2

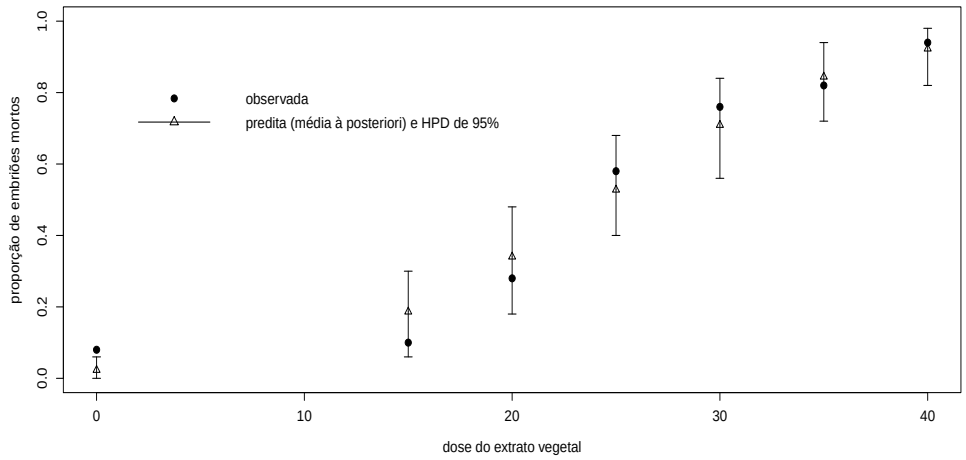

Figura 2: Proporções, observadas e preditas pelo modelo 1, de embriões mortos em função da dose do extrato vegetal

3. Os dados analisados se referem à uma pesquisa realizada na Austrália que teve, entre outros objetivos, comparar a assiduidade de estudantes (representada pelo número de faltas durante um certo período) entre duas etnias (A: aborígene e N: não aborígene), respectivamente grupo 1 e grupo 2. Para realizar esta comparação, o seguinte o modelo foi considerado:

$$\begin{aligned} Y_{ij}|\boldsymbol{\beta} &\overset{ind.}{\sim} \text{Poisson}(\mu_i) \\ \ln \mu_i &= \mu + \alpha_2, \alpha_1 = 0, \boldsymbol{\beta} = (\mu, \alpha_2)', \end{aligned}$$

em que $i = 1, 2; j = 1, 2, ..., 69$, se $i = 1; j = 1, 2, ..., 77$, se $i = 2$. Além disso, considerou-se que $\mu \sim N(0, 1000)$ e $\alpha_2 \sim N(0, 1000)$, mutuamente independentes. Os resultados a seguir, Figuras de 3 à 9 e Tabela 6, são oriundos de uma amostra válida de 1000 valores (com exceção dos gráficos de diagnóstico da convergência do algoritmo MCMC, os quais foram construídos com toda a amostra de uma ou das três cadeias), para cada posteriori, obtidas através de um algoritmo MCMC implementado no programa WinBUGS. Considerou-se os seguintes valores para a obtenção da amostra válida: burn-in = 10001, espaçamento = 5 e número total de valores simulados = 60000. Responda os itens abaixo:

a) Observando os gráficos de diagnóstico, o que você pode concluir sobre a convergência do algoritmo implementado? Você concorda com as quantidades adotadas (número de valores simulados, burn-in e espaçamento), para se ter amostras aleatórias (aproximadamente não correlacionadas) das distribuições à posteriori de interesse? Justifique, adequadamente, suas respostas (200 pontos).

b) Observando os resultados relativos às estimativas, incluindo as médias de cada grupo, o que você conclui sobre a assiduidade dos grupos? Qual grupo de estudantes é mais assíduo? Qual é a razão estimada entre o número médio de faltas do grupo menos assíduo e o mais assíduo? Justifique, adequadamente, suas respostas (200 pontos).

c) Observando os resultados relativos ao ajuste do modelo, Figuras 8 e 9, o que você conclui sobre a qualidade do ajuste dele? Você utilizaria o modelo em questão para analisar os dados ? Porquê? Caso você tenha concluído por não utilizar o modelo em questão, proponha um modelo alternativo. Justifique, adequadamente, suas respostas (250 pontos).

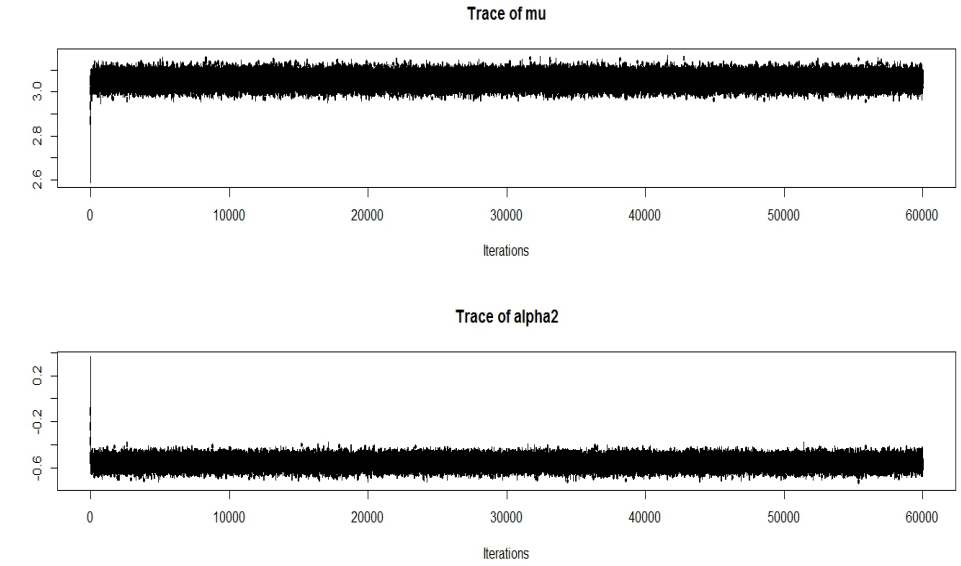

Figura 3: Gráficos de trajetórias (Traceplots) para os três conjuntos de cadeias geradas (as cadeias se diferenciam pelo tipo de linha: sólida, tracejada e pontilhada: Questão 3)

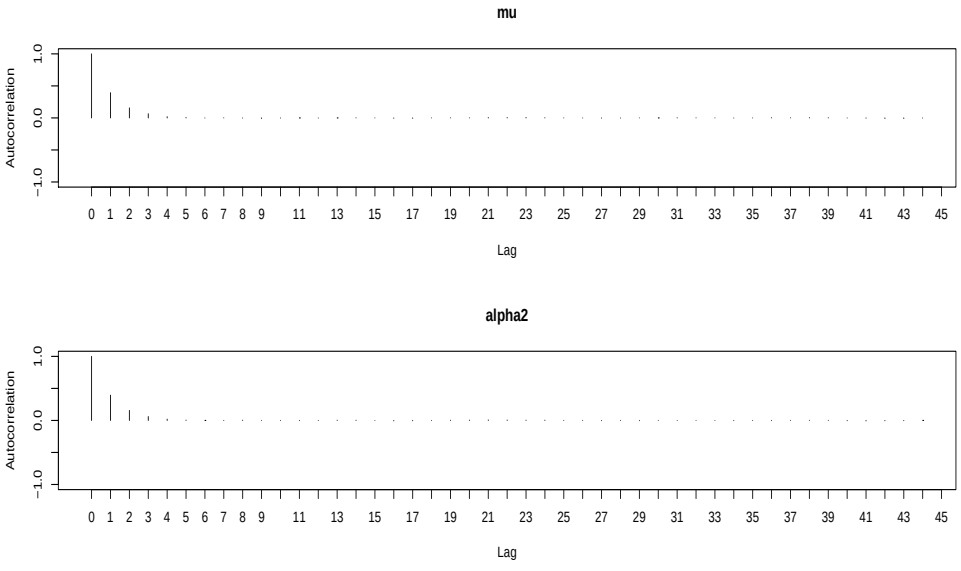

Figura 4: Autocorrelações para um dos conjuntos de cadeias: Questão 3

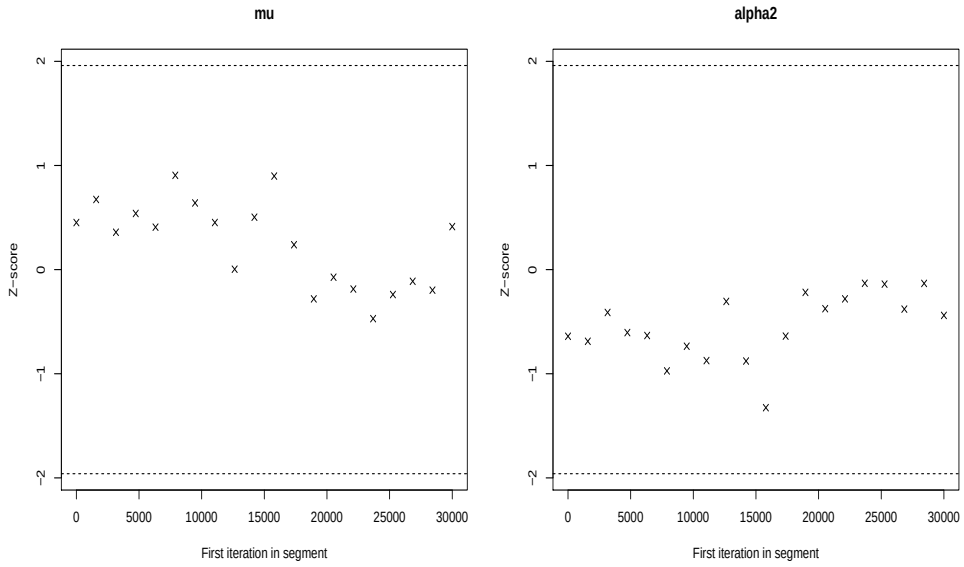

Figura 5: Gráfico da estatística de Geweke para um dos conjuntos de cadeias: Questão 3

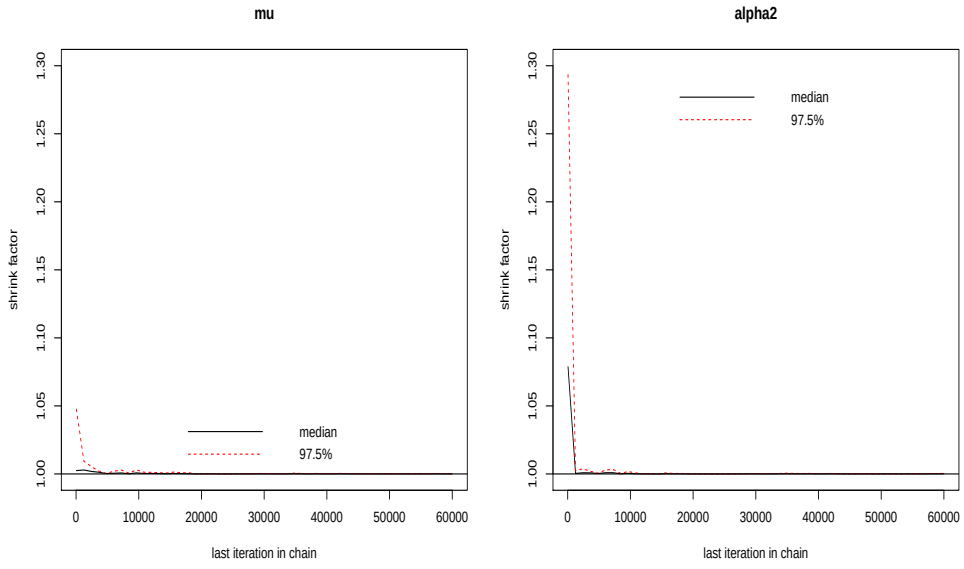

Figura 6: Gráfico da estatística de Gelman-Rubin considerando os três conjuntos de cadeias: Questão 3

Tabela 6: Estimativas bayesianas para os parâmetros do modelo: Questão 3

	EAP	DPAP	$IC_B(95\%)$	HPD(95%)
μ	3,05	0,03	[3,00 ; 3,11]	[3,00 ; 3,10]
α_2	-0,56	0,04	[-0,64 ;-0,48]	[-0,64 ; -0,48]
μ_1	21,22	0,55	[20,16 ; 22,32]	[20,17 ; 22,33]
μ_2	12,18	0,40	[11,42 ; 12,99]	[11,41 ; 12,96]

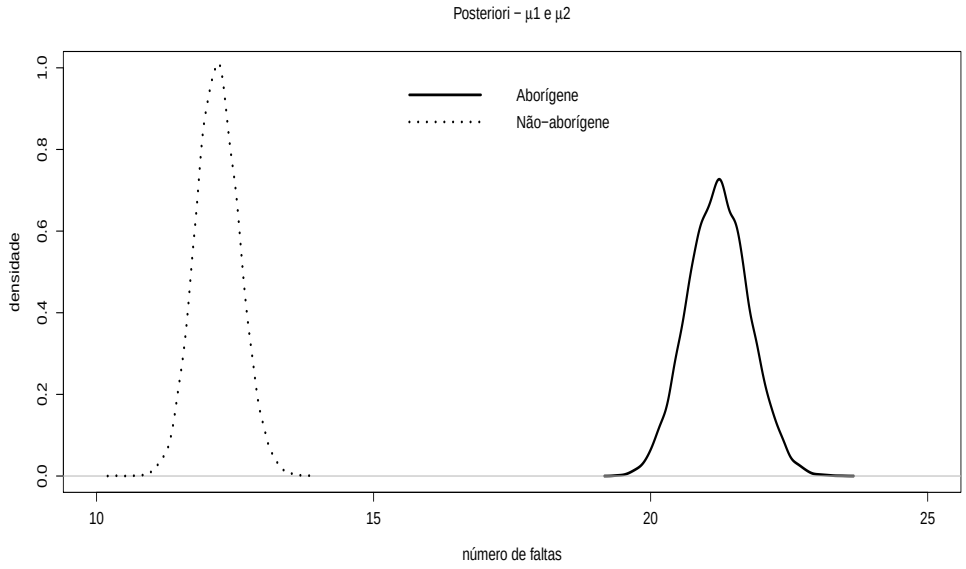

Figura 7: Posterioris das médias de cada grupo: Questão 3

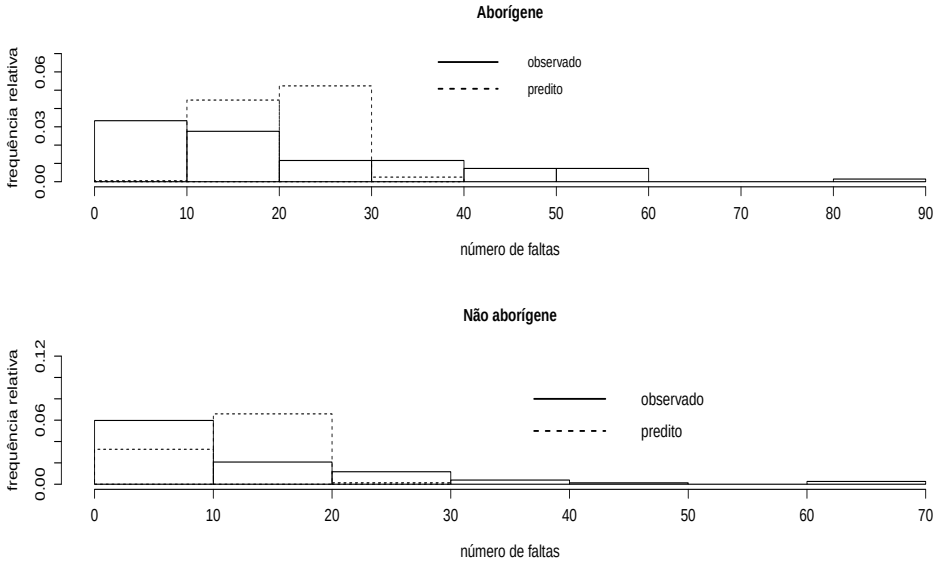

Figura 8: Valores observados e distribuições preditivas (linha sólida - valores observados ; linha tracejada - valores preditos) (gráficos sobrepostos): Questão 3

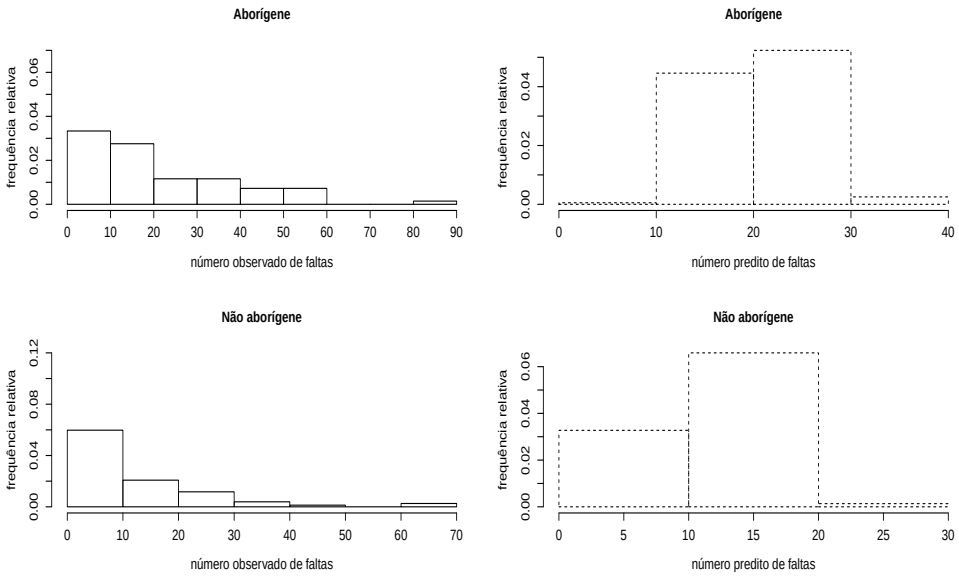

Figura 9: Valores observados e distribuições preditivas (gráficos separados): Questão 3

Formulário

1. Se $X|(r,\theta) \sim \text{gama}(r,\theta), r>0, \theta>0$, então $p(x|r,\theta) = \dfrac{1}{\theta^r \Gamma(r)} e^{-\frac{x}{\theta}} x^{r-1} \mathbb{1}_{(0,\infty)}(x)$. Se r=1, então $X|\theta \sim exp(\theta)$.

2. Se $X|(a,b) \sim \text{IG}(a,b), a>0, b>0$, então $p(x|a,b) = \dfrac{b^a}{\Gamma(a)} e^{-\frac{b}{x}} x^{-a-1} \mathbb{1}_{(0,\infty)}(x)$.

Teste de Hipóteses Bayesianos

- Para as hipóteses $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0, \theta \in \Theta_\theta$, temos

  Priori $p(\lambda) = \alpha \mathbb{1}_{\{\lambda_0\}}(\lambda) + (1-\alpha) p_1(\lambda) \mathbb{1}_{\Theta_1}(\lambda), \Theta_1 = \Theta - \{\lambda_0\}$, em que $p_1(.)$ é uma distribuição de probabilidade em Θ_1. Então o fator de Bayes é dado por:

  $B(\boldsymbol{x}) = \dfrac{p_1(\boldsymbol{x})}{p(\boldsymbol{x}|\lambda_0)}$, $p_1(\boldsymbol{x}) = \displaystyle\int_{\Theta_1} p(\boldsymbol{x}|\lambda) p_1(\lambda) d\lambda$ e $p(\boldsymbol{x}|\lambda_0)$ é a verossimilhança avaliada em λ_0.
- Para o Fator de Bayes:

Valor	Evidência a favor de H_1
< 1	Contra
$[1;3)$	Leve
$[3;10)$	Moderada
$[10;30)$	Forte
$[30;100)$	Muito forte
≥ 100	Decisiva