Analise de dados, escolha de prioris e intervalos (regiões) de credibilidade

Prof. Caio Azevedo

Dados reais: estimação do número médio de acidentes

- Descrição: número de acidentes (com algum tipo de trauma para as pessoas envolvidas) em 92 dias durante o ano de 1961, medidos em algumas regiões da Suécia.
- Considerou-se apenas 43 dias, correspondendo àqueles em que não havia limite de velocidade.
- Vamos assumir que

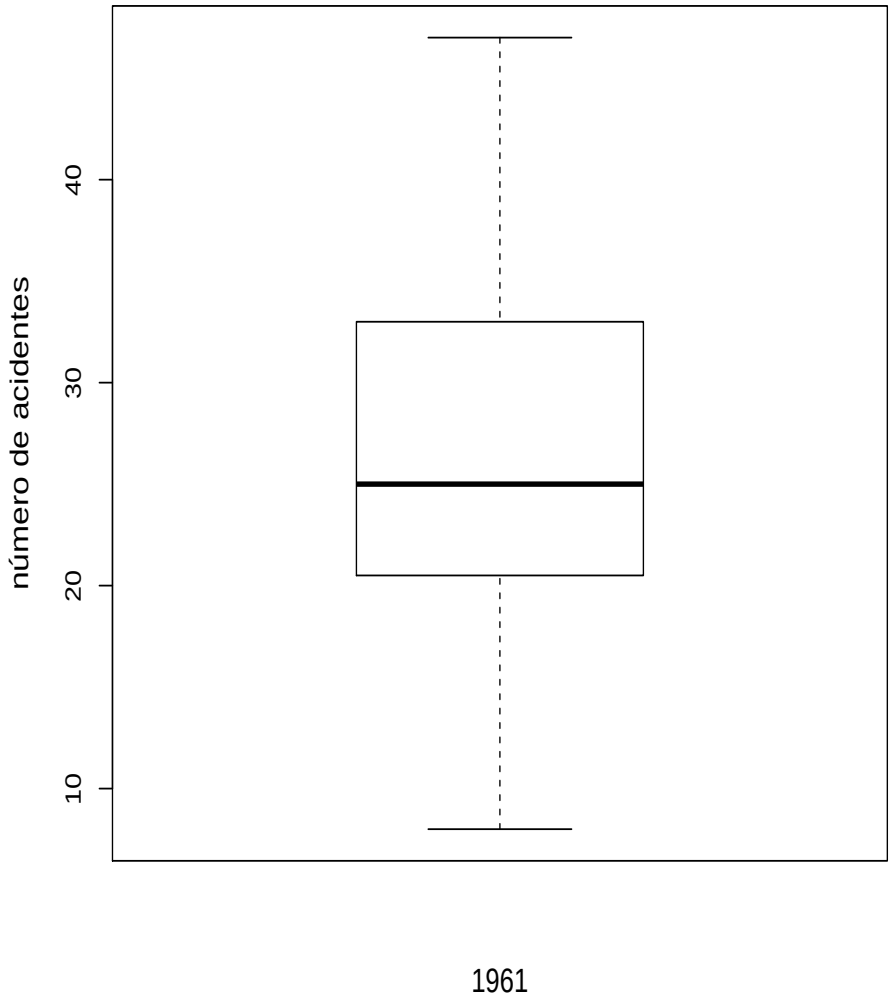
  $X_i|\lambda \overset{i.i.d.}{\sim} Poisson(\lambda), i = 1, ..., 43.$

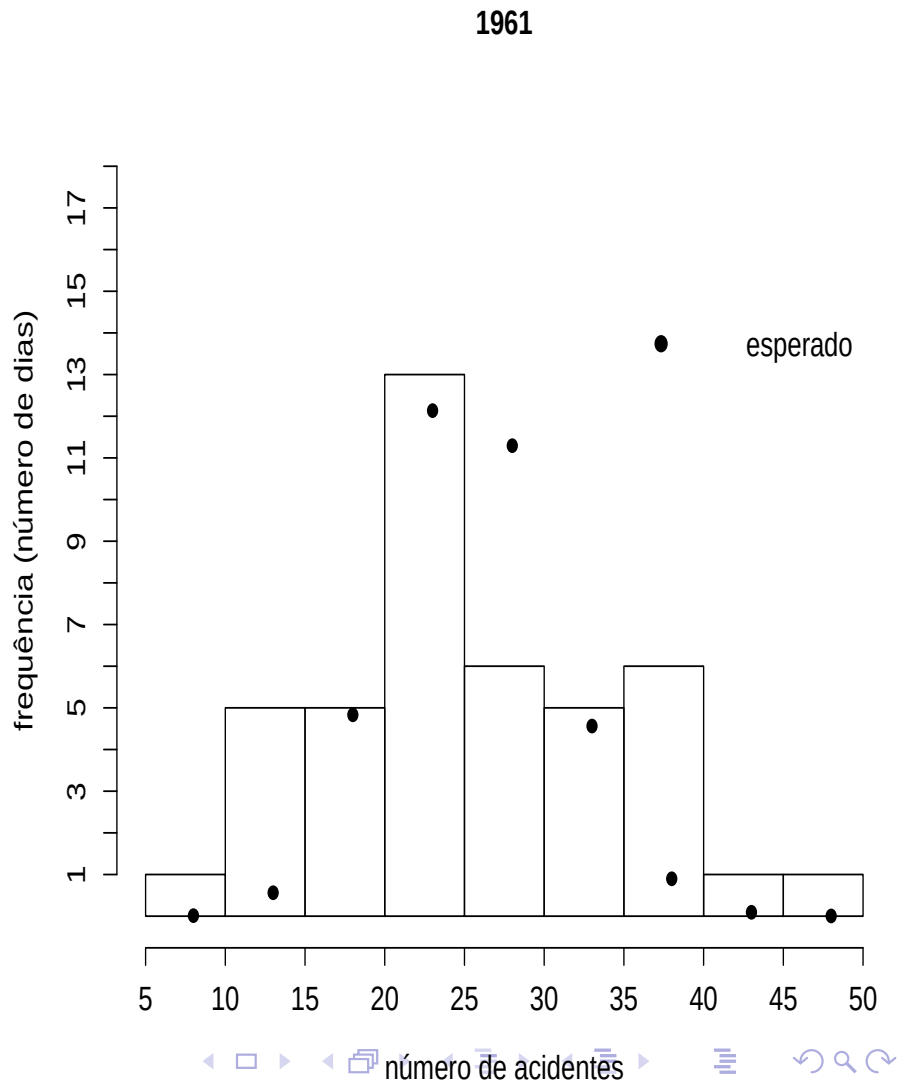
  que representa o número de acidentes observados no i-ésimo dia.
- Objetivo : estimar λ (pontual e intervalarmente).

Análise descritiva: medidas resumo

média	var.	dp	cv(%)	1^o quartil	med.	3^o quartil
26,05	82,66	9,09	34,91	8,00	25,00	47,00

Boxplot e histograma





Inferência frequentista: estimação por MV

- Sabemos, sob as suposições anteriores, que:

$$\widehat{\lambda} \approx N\left(\lambda, \frac{\lambda}{n}\right)$$

para n suficientemente grande, em que $\widehat{\lambda} = \overline{X}$ (estimador de máxima verossimilhança).

- Assim, $EP_A(\widehat{\lambda}) = \sqrt{\frac{\lambda}{n}}$.
- Além disso, $IC_A(\lambda; \gamma\%) = \left[\overline{X} - \widehat{EP}_A(\widehat{\lambda}) z_{\frac{1-\gamma}{2}}; \overline{X} + \widehat{EP}_A(\widehat{\lambda}) z_{\frac{1-\gamma}{2}}\right]$

em que $P(Z > z_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2}$, $Z \sim N(0,1)$ e $\widehat{EP}_A(\widehat{\lambda}) = \sqrt{\frac{\widehat{\lambda}}{n}}$.

Inferência frequentista: estimação por MV (cont.)

Estimativa (MV)	EP_A	IC(λ,95%)
26,05	0,78	[24,52;27,57]

Inferência bayesiana

- Qual priori utilizar?
- Consideraremos três possibilidades
  - Não informativa: $p^{NI}(\theta) \propto \mathbb{1}_{(0,\infty)}(\theta)$.
  - Priori de Jeffreys (aproximadamente não informativa):
    $p^{J}(\theta) \propto \lambda^{-1/2}\mathbb{1}_{(0,\infty)}(\theta)$.
  - Família conjugada (gama(a,b^{-1})): $p^{FC}(\theta) \propto \lambda^{a-1}e^{-\lambda b}\mathbb{1}_{(0,\infty)}(\theta)$.
- As duas primeiras prioris são casos limite da terceira. Com efeito:
  $p^{NI}(\theta)$ é obtida fazendo-se a=1 e $b \to 0$, e
  $p^{J}(\theta)$ é obtida fazendo-se a=1/2 e $b \to 0$

Distribuições preditivas

- Seja $p(.)$ uma priori e $p(.|\mathbf{x})$ a respectiva posteriori, para uma dada verossimilhança $p(.|\theta)$, $\mathbf{x} = (x_1, ...x_n)$
- Seja x_{n+1} uma observação de $X_{n+1}|\theta$ (uma variável aleatória cujo valor observado não foi utilizado para a construção da posteriori). Ou seja, $X_{n+1}|\theta$ é uma (nova) variável aleatória a ser sorteada.
- Em nosso caso, seria um outro dia em que se registraria o número de acidentes.

Distribuições preditivas (cont.)

- Distribuição preditiva à priori

$$p^*(x_{n+1}|\mathbf{x}) = \int_\Theta p(x_{n+1}|\theta, \mathbf{x}) p(\theta) d\theta$$

- Distribuição preditiva à posteriori

$$p(x_{n+1}|\mathbf{x}) = \int_\Theta p(x_{n+1}|\theta, \mathbf{x}) p(\theta|\mathbf{x}) d\theta$$

Distribuições preditivas (cont.)

- Se as observações forem condicionamente independentes (em θ), então $p(x_{n+1}|\theta, \mathbf{x}) = p(x_{n+1}|\theta)$ e, assim:
  - Distribuição preditiva à priori

$$p^*(x_{n+1}|\mathbf{x}) = p(x_{n+1}) = \int_\Theta p(x_{n+1}|\theta)p(\theta)d\theta$$

  - Distribuição preditiva à posteriori

$$p(x_{n+1}|\mathbf{x}) = \int_\Theta p(x_{n+1}|\theta)p(\theta|\mathbf{x})d\theta$$

- Objetivo: comparar as prioris em termos de qualidade de reprodutibilidade dos dados observados através das distribuições preditivas à posteriori.

Distribuições à posteriori

- Para cada uma das prioris consideradas anteriormente, temos as seguintes posteriores

$$
\begin{aligned}
p^{FC}(\theta|\mathbf{x}) &\propto e^{-(n+b)\lambda}\lambda^{n\bar{x}+a-1}\mathbb{1}_{(0,\infty)}(\lambda) \\
p^{NI}(\theta|\mathbf{x}) &\propto e^{-n\lambda}\lambda^{n\bar{x}+1-1}\mathbb{1}_{(0,\infty)}(\lambda) \\
p^{J}(\theta|\mathbf{x}) &\propto e^{-n\lambda}\lambda^{n\bar{x}+1/2-1}\mathbb{1}_{(0,\infty)}(\lambda)
\end{aligned}
$$

- Ou seja, respectivamente

$\lambda|\mathbf{x} \sim gama(n\bar{x}+a,(n+b)^{-1})$; $\lambda|\mathbf{x} \sim gama(n\bar{x}+1,n^{-1})$;

$\lambda|\mathbf{x} \sim gama(n\bar{x}+1/2,n^{-1})$

Distribuições preditivas à posteriori no exemplo

- Para a posteriori (FC), temos (considere que $x_{n+1} \equiv x$, $a^* = (n\overline{x} + a)$ e $b^* = (n + b)$):

$$
\begin{aligned}
p^{FC}(x_{n+1}|\mathbf{x}) &= \int_0^\infty p(x_{n+1}|\lambda) p^{FC}(\theta|\mathbf{x}) d\theta \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} \mathbb{1}_{\mathcal{N}}(x) \frac{(b^*)^{a^*}}{\Gamma(a^*)} \lambda^{a^*-1} e^{-\lambda b^*} d\theta \\
&= \frac{(b^*)^{a^*}}{\Gamma(a^*)x!} \mathbb{1}_{\mathcal{N}}(x) \int_0^\infty \lambda^{a^*+x-1} e^{-\lambda(b^*+1)} d\theta \\
&= \frac{\Gamma(a^*+x)}{\Gamma(a^*)x!} \frac{(b^*)^{a^*}}{(b^*+1)^{a^*+x}} \mathbb{1}_{\mathcal{N}}(x) \\
&= \frac{\Gamma(a^*+x)}{\Gamma(a^*)x!} \left(\frac{b^*}{b^*+1}\right)^{a^*} \left(\frac{1}{b^*+1}\right)^{x} \mathbb{1}_{\mathcal{N}}(x)
\end{aligned}
$$

Distribuições preditivas à posteriori no exemplo (cont.)

- Se $a \in \{0, 1, ...\}$, então $a^* = (a + n\overline{x}) \in \{0, 1, ...\}$ e :

$$p^{FC}(x_{n+1}|\mathbf{x}) = \begin{pmatrix} x + a^* - 1 \\ x \end{pmatrix} \left(\frac{b^*}{b^* + 1}\right)^{a^*} \left(\frac{1}{b^* + 1}\right)^{x} \mathbb{1}_{\mathcal{N}}(x)$$

logo $X_{n+1}|\mathbf{x} \sim BN(a^*, \theta), \theta = \frac{b^*}{b^*+1}$.

- No geral,

$$p^{FC}(x_{n+1}|\mathbf{x}) = \frac{\Gamma(a^* + x)}{\Gamma(a^*)x!} \left(\frac{b^*}{b^* + 1}\right)^{a^*} \left(\frac{1}{b^* + 1}\right)^{x} \mathbb{1}_{\mathcal{N}}(x)$$

Distribuições preditivas à posteriori no exemplo (cont.)

- OBS: mesmo que *a* não seja um número natural (desde que seja positivo), podemos utilizar as funções no R, relativas à distribuição binomial negativa, para calcular probabilidades envolvendo a distribuição preditiva.
- Funções: *dnbinom, pnbinom, qnbinom, rnbinom*.

Distribuições preditivas à posteriori no exemplo (cont.)

- Para as outras prioris, temos

$$
\begin{aligned}
p^{NI}(x_{n+1}|\mathbf{x}) &= \frac{\Gamma(n\overline{x}+x+1)}{\Gamma(n\overline{x}+1)x!}\left(\frac{n}{n+1}\right)^{n\overline{x}+1}\left(\frac{1}{n+1}\right)^{x}\mathbb{1}_{\mathcal{N}}(x) \\
&= \begin{pmatrix} x+n\overline{x} \\ x \end{pmatrix}\left(\frac{n}{n+1}\right)^{n\overline{x}+1}\left(\frac{1}{n+1}\right)^{x}\mathbb{1}_{\mathcal{N}}(x) \\
p^{J}(x_{n+1}|\mathbf{x}) &= \frac{\Gamma(n\overline{x}+x+1/2)}{\Gamma(n\overline{x}+1/2)x!}\left(\frac{n}{n+1}\right)^{n\overline{x}+1/2}\left(\frac{1}{n+1}\right)^{x}\mathbb{1}_{\mathcal{N}}(x)
\end{aligned}
$$

- Sob a priori NI, temos que $X_{n+1}|\mathbf{x} \sim BN\left(n\overline{x}+1, \frac{n}{n+1}\right)$.

Hiperparâmetros da priori conjugada

- Como obter os hiperparâmetros (a,b) da priori FC?
- Podemos utilizar os dados (inferência bayesiana empírica) para obter os hiperparâmetros.
- Em nosso caso, note que

$$
\begin{aligned}
p(\mathbf{x}|\lambda)p(\lambda) &= \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_{i=1}^{n} x_i!}\frac{b^a}{\Gamma(a)}e^{-b\lambda}\lambda^{a-1}\mathbb{1}_{(0,\infty)}(\lambda) \\
&= \frac{e^{-(n+b)\lambda}\lambda^{(n\bar{x}+a)-1}b^a}{\prod_{i=1}^{n} x_i!\Gamma(a)}\mathbb{1}_{(0,\infty)}(\lambda)
\end{aligned}
$$

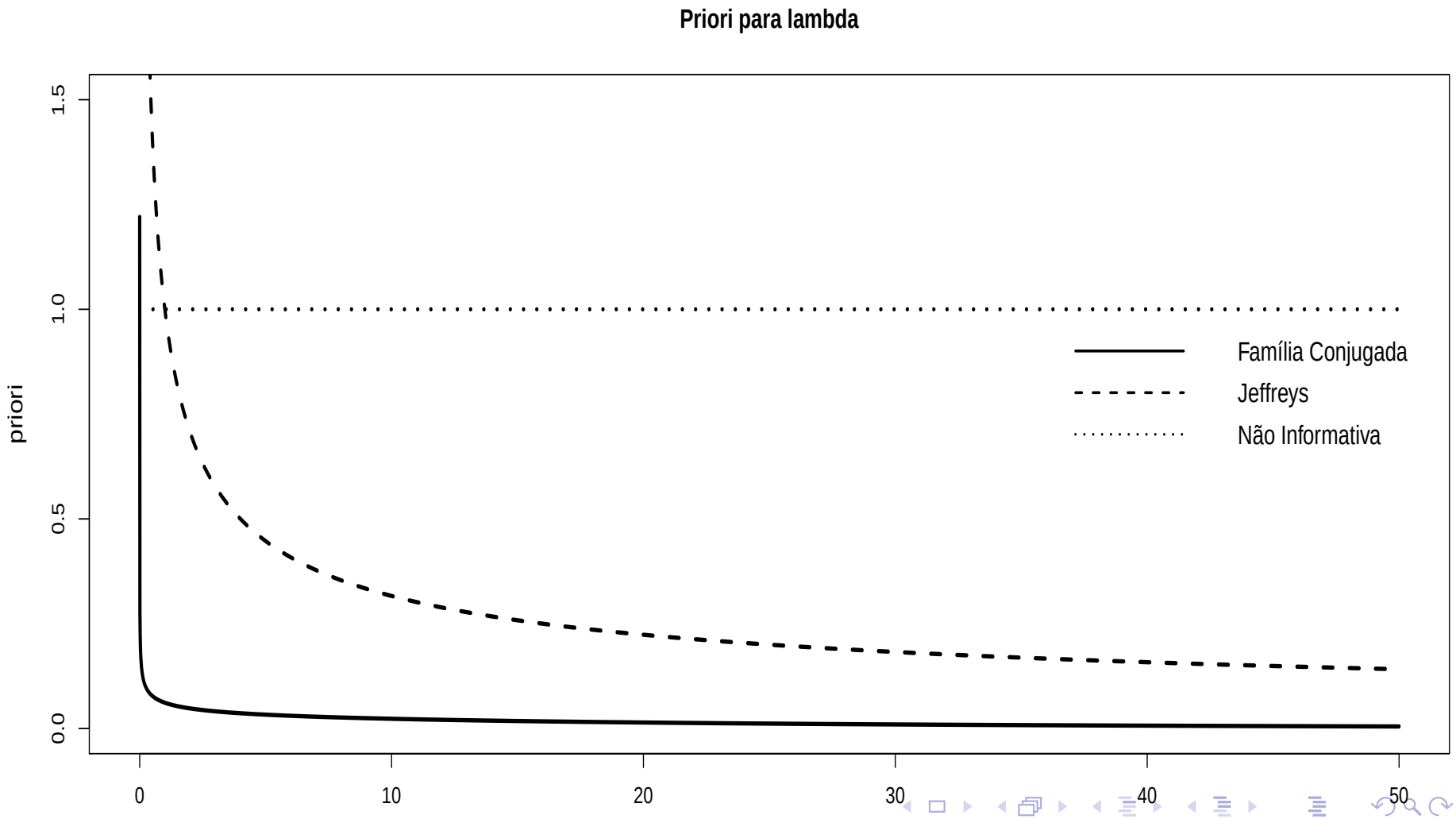
Hiperparâmetros da priori conjugada (cont.)

- Assim, temos

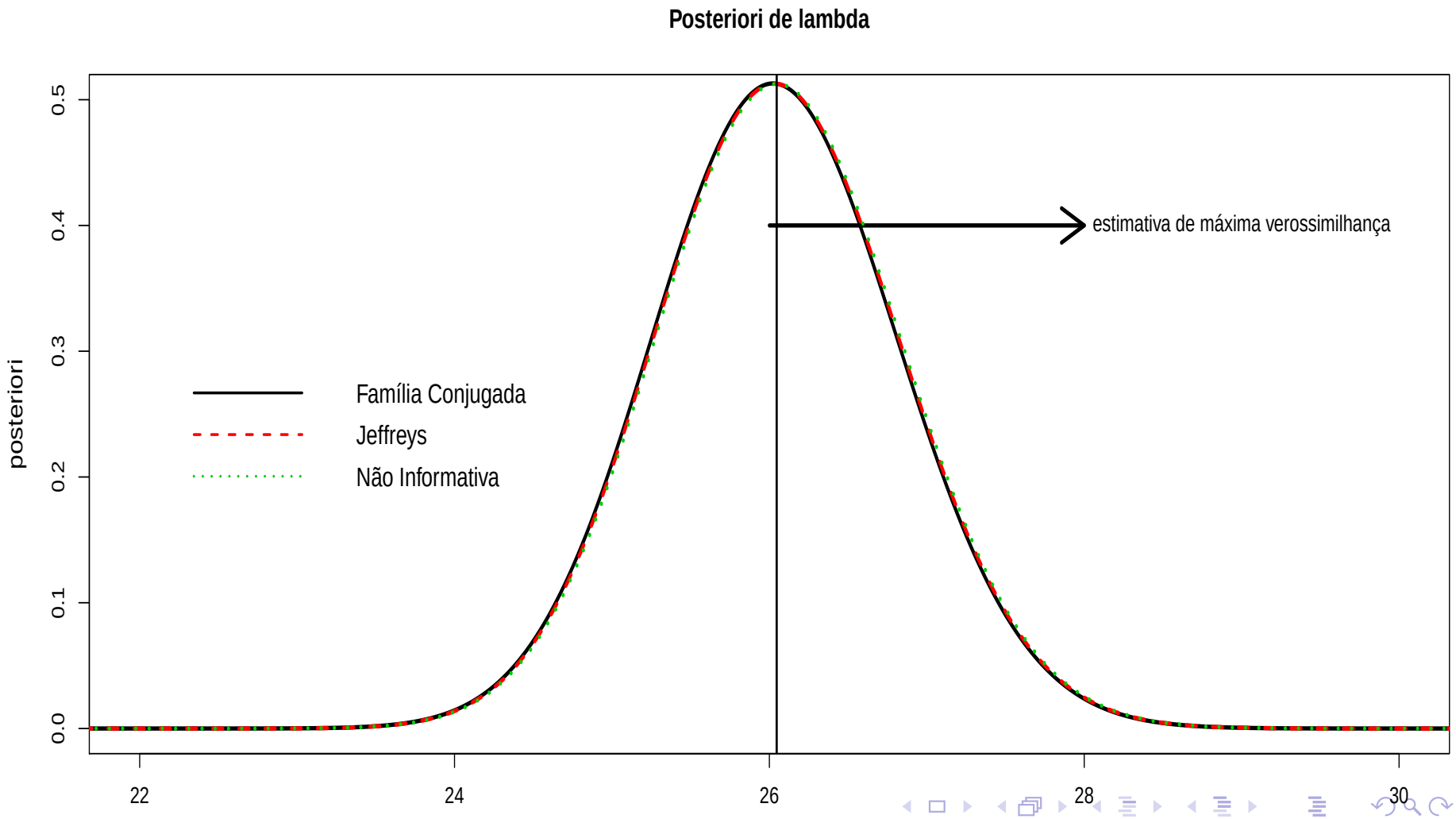
$$p(\mathbf{x}|a,b) = \int_0^\infty p(\mathbf{x}|\lambda)p(\lambda)d\lambda = \frac{\Gamma(n\overline{x}+a)b^a}{(n+b)^{n\overline{x}+a}\Gamma(a)\prod_{i=1}^n x_i!}$$

- Portanto, eliminou-se λ da função acima, originando uma espécie de verossimilhança para os hiperparâmetros. Portanto, podemos obter as estimativas de MV para (a,b) e usá-las na priori.
- No entanto, determinaremos os hiperparâmetros através da relação $\mathcal{E}(\lambda) = \frac{a}{b}$ e $\mathcal{V}(\lambda) = \frac{a}{b^2}$.
- Fixando-se $\mathcal{E}(\lambda) = 26,05$ (média amostral) e $\mathcal{V}(\lambda) = 1000$, obtem-se $a \approx 0,6784$ e $b \approx 0,02605$.

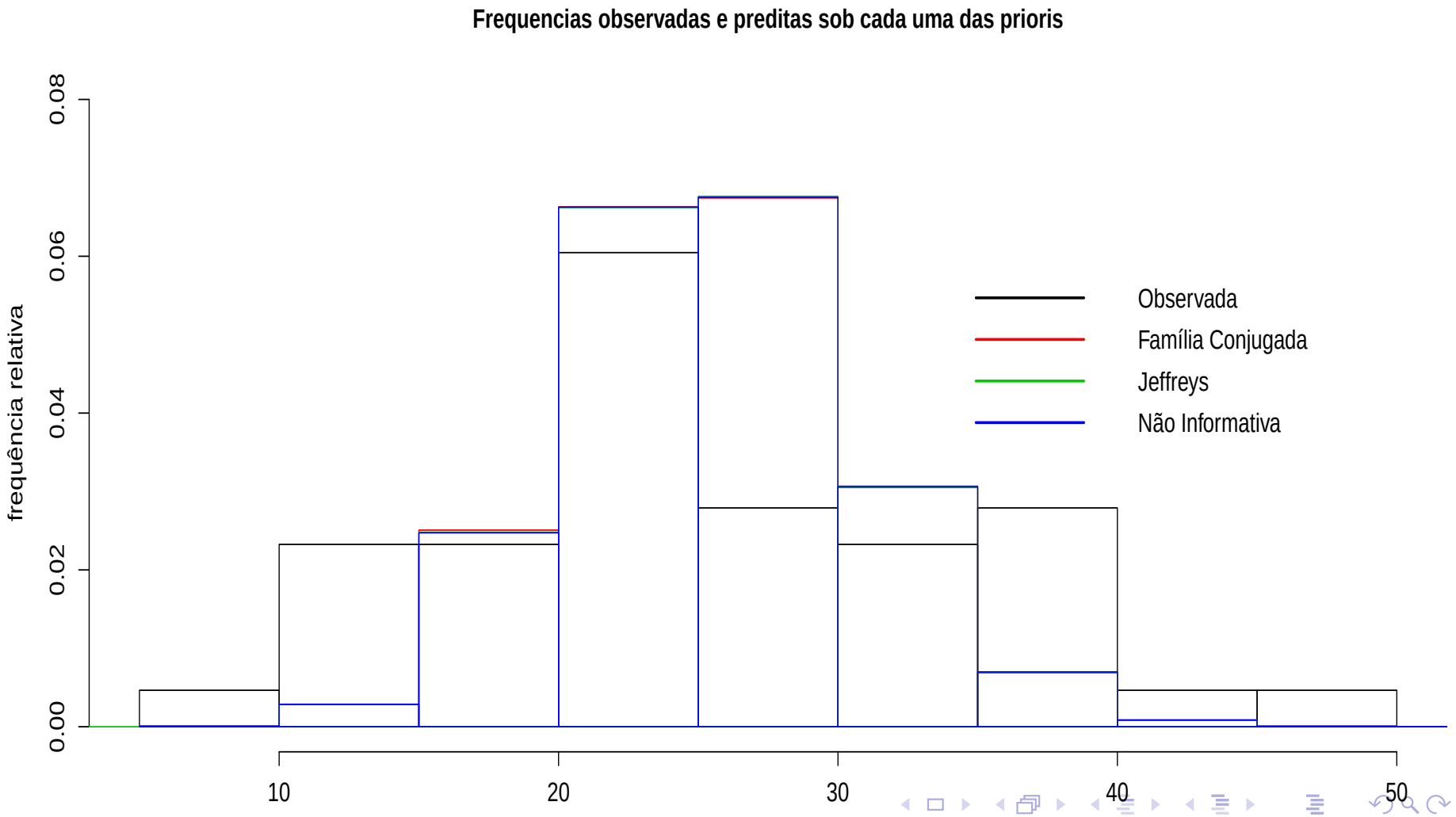
Comparação entre as prioris



Comparação entre as posterioris



Comparação entre as distribuições preditivas



Comparação entre as distribuições preditivas (cont.)

- Quantitativamente, podemos comparar as distribuições preditivas (prioris) através das frequências observadas e preditas.
- Frequência observada (FO_{ij}): número de dias em que ocorreu de i (exclusive) à j (inclusive) acidentes.
- Frequência predita: $FP_{ij} = nP(i < X_{n+1} \leq j|\mathbf{x})$.
- Para a primeira categoria considera-se igualdade também para o valor i.
- Estatística: $\sum_{i<j} |FO_{ij} - FP_{IJ}|$.

Comparação entre as distribuições preditivas (cont.)

Priori	Estatística
Família Conjugada	18,97
Jeffreys	18,99
Não-informativa	19,01

As prioris, sob o critério acima, apresentaram, praticamente, o mesmo desempenho. Vamos considerar, doravante,a priori correspondente à família conjugada.

Estimativa pontual e desvio padrão à posteriori

Estimativa (EAP)	DAP
26,05	0,78

Pergunta: como construir estimativas intervalares para λ?

Intervalos (regiões) de credibilidade

- Uma região $R(\mathbf{x}) \in \Theta$ é ditar ser uma região de credibilidade (RC_B) γ para θ se

$$P(\theta \in R(\mathbf{x})|\mathbf{x}) = \int_{R(\mathbf{x})} p(\theta|\mathbf{x})d\theta \geq \gamma$$

a desigualdade acima é considerada para contemplar o caso em que θ é discreto. No caso contínuo, em geral, trabalha-se com a igualdade.

- Particularmente, se θ é um escalar, então, em geral, $R(\mathbf{x}) = [R_1(\mathbf{x}), R_2(\mathbf{x})], R_1(\mathbf{x}) \leq R_2(\mathbf{x}), \forall x \in \mathbf{X}(\Omega)$.

Intervalos (regiões) de credibilidade (cont.)

- Neste caso,

$$P(\theta \in R(\mathbf{x})|\mathbf{x}) = \int_{R_1(\mathbf{x})}^{R_2(\mathbf{x})} p(\theta|\mathbf{x})d\theta \geq \gamma$$

e $R(\mathbf{x})$ passa a ser chamado de intervalo de credibilidade (IC_B) γ para θ.

- Uma medida de precisão frequentista para o IC_B é o comprimento esperado (frequentista), ou seja

$$\mathcal{E}_{\mathbf{X}|\theta}(R(\mathbf{X})) = \mathcal{E}_{\mathbf{X}|\theta}[R_2(\mathbf{x}) - R_1(\mathbf{x})].$$

Intervalos (regiões) de credibilidade (cont.)

- Contudo, busca-se obter intervalos com o menor comprimento (ou volume), sem se tomar a esperança. Ou seja, avaliando-os em termos da amostra observada.
- Estes intervalos (regiões) correspondem àqueles de maior densidade à posteriori, em inglês, *highest posterior density* (HPD).

Intervalos (regiões) de credibilidade HPD

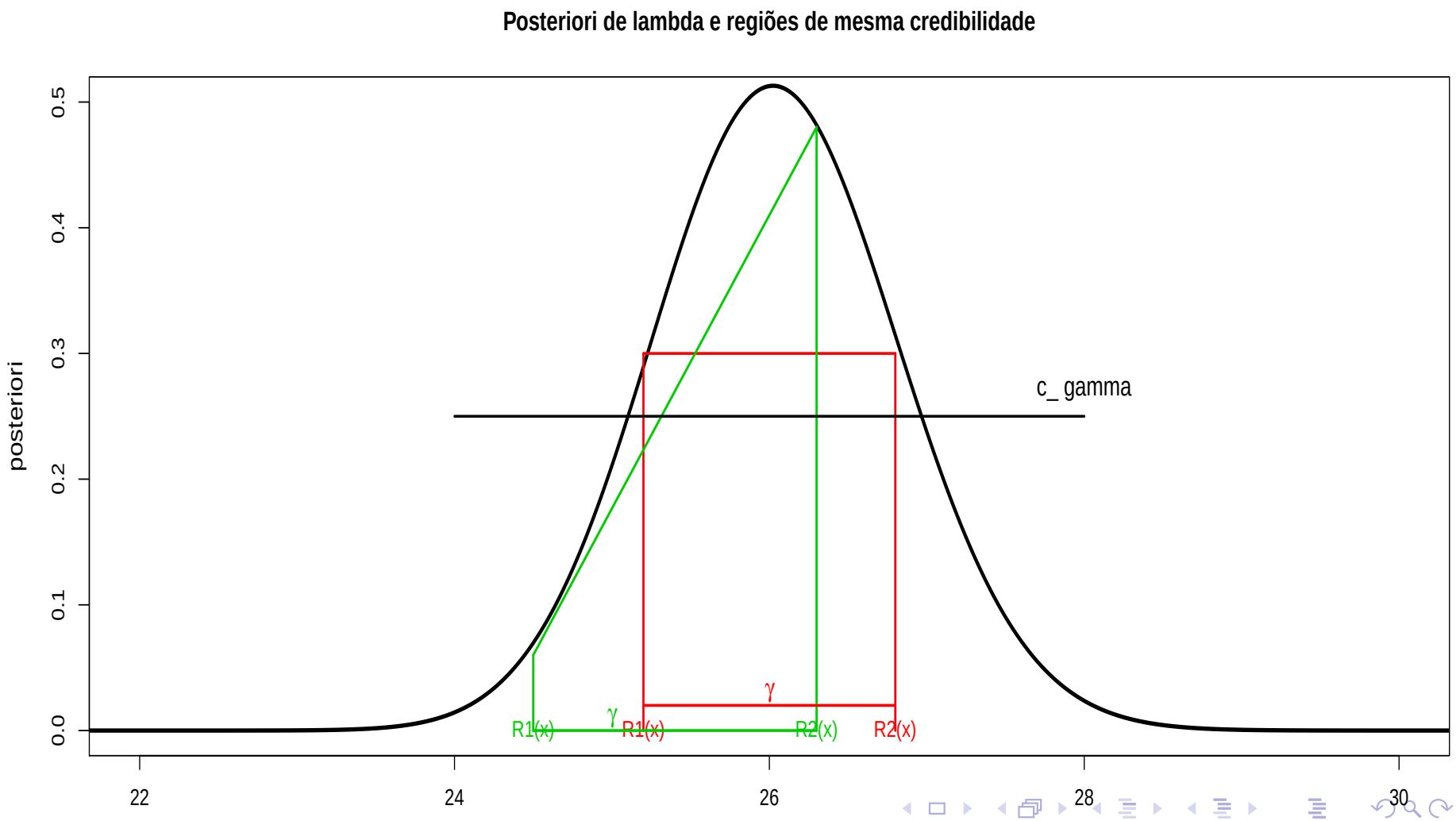
- A definição para um intervalo (região) HPD é a seguinte:

$$R(\mathbf{x}) = \{\theta : p(\theta|\mathbf{x}) \geq c_\gamma\}$$

em que c_γ é a maior constante tal que:

$$P(\theta \in R(\mathbf{x})|\mathbf{x}) = \int_{R(\mathbf{x})} p(\theta|\mathbf{x}) d\theta \geq \gamma$$

Intervalos de credibilidade para o exemplo dos acidentes



Intervalos (regiões) de credibilidade (cont.)

- Se a posteriori for simétrica e unimodal, o intervalo HPD corresponde ao intervalo simétrico, ou seja

$$P(\theta \leq R_1(\mathbf{x})|\mathbf{x}) = \frac{1-\gamma}{2} \text{ e } P(\theta \geq R_2(\mathbf{x})|\mathbf{x}) = \frac{1-\gamma}{2}$$

- Se a posteriori for assimétrica e unimodal, o intervalo HPD corresponde àquele em que a posteriori apresenta o mesmo valor nos limites, ou seja

$$p(R_1(\mathbf{x})|\mathbf{x}) = p(R_2(\mathbf{x})|\mathbf{x})$$

- Em geral, no segundo caso, o HPD não é obtenível analiticamente e, assim, deve-se empregar métodos numéricos para se construí-los. Há pacotes no R que obtêm, numericamente, intervalos HPD.

Intervalos (regiões) de credibilidade (cont.)

- Voltando ao exemplo do número de acidentes, temos que

$$\lambda|\mathbf{x} \sim gama(n\overline{x} + a, (n+b)^{-1})$$

- Neste caso, a posteriori é assimétrica e unimodal. Para obtermos um $IC_B(\lambda, \gamma)$, podemos utilizar os quantis da distribuição gama, da seguinte forma

  $P(\lambda \leq R_1(\mathbf{x})|\mathbf{x}) = \frac{1-\gamma}{2}$ e $P(\lambda \geq R_2(\mathbf{x})|\mathbf{x}) = \frac{1-\gamma}{2}$

- Utilizando o programa R, para $\gamma = 0,95$, obtemos $R_1(\mathbf{x}) = 24,54$ e $R_2(\mathbf{x}) = 27,59$.

Intervalos (regiões) de credibilidade (cont.)

- Uma outra opção é, primeiramente, encontrar um IC_B para uma transformação de λ que apresente uma "distribuição-tabelada" (N(0,1), t de Student, qui-quadrado, F de Snedcor).
- No nosso caso, $\theta = 2(n+b)\lambda|\mathbf{x} \sim \chi^2_{[2(n\bar{x}+a)]}$
- Nesse caso, o intervalo simétrico para θ é $IC^*_B(\lambda, 95\%) = R^*(\mathbf{x}) = [2112,04; 2374,47]$.

Intervalos (regiões) de credibilidade (cont.)

- Isto implica que o intervalo (simétrico) para λ é $IC_B(\lambda, 95\%) = R(\mathbf{x}) = R^*(\mathbf{x})/(2(n+b)) = [24,54; 27,59]$.
- O IC_B simétrico para a transformação gera o IC_B simétrico, através da transformação inversa, para o parâmetro original. Contudo, isto, em geral, não é verdade para o intervalo HPD (a menos que a transformação seja linear).

Intervalos (regiões) de credibilidade (cont.)

- Note que, além da restrição natural:

  $P_{\lambda|\mathbf{x}}(R_1(\mathbf{x}) \leq \lambda \leq R_2(\mathbf{x})|\mathbf{x}) = \gamma$, o intervalo HPD de credibilidade γ, deve satisfazer à

$$p_{\lambda|\mathbf{x}}(R_1(\mathbf{x})|\mathbf{x}) = p_{\lambda|\mathbf{x}}(R_2(\mathbf{x})|\mathbf{x})$$

- Assim, para a obtenção do intervalo HPD, precisamos resolver o seguinte sistema de equações:

$$\begin{cases} p_{\lambda|\mathbf{x}}(R_2(\mathbf{x})|\mathbf{x}) - p_{\lambda|\mathbf{x}}(R_1(\mathbf{x})|\mathbf{x}) & = 0 \\ P_{\lambda|\mathbf{x}}(\lambda \leq R_2(\mathbf{x})|\mathbf{x}) - P_{\lambda|\mathbf{x}}(\lambda \leq R_1(\mathbf{x})|\mathbf{x}) - \gamma & = 0 \end{cases}$$

Intervalos (regiões) de credibilidade (cont.)

- A função *hpd* do pacote *TeachingDemos* resolve tal sistema de equações.
- Em nosso caso, devemos utilizar o comando

  hpd(qgamma,shape=shapeFC,rate=rateFC)

  em que $shapeFC = n\bar{x} + a$ e $rateFC = n + b$
- Assim $IC_{HPD}(\lambda, \gamma) = [24,53; 27,58]$, que neste caso (como era de se esperar) praticamente coincide com o IC_B simétrico.