Inferência Estatística com Abordagem Bayesiana

Rosangela Helena Loschi [1]

[1]Departamento de Estatística
Universidade Federal de Minas Gerais

16 de dezembro de 2021

INFERÊNCIA INTERVALAR

- Regiões de credibilidadee
- Regiões com mais alta densidade *a posteriori*

Estimação por Regiões

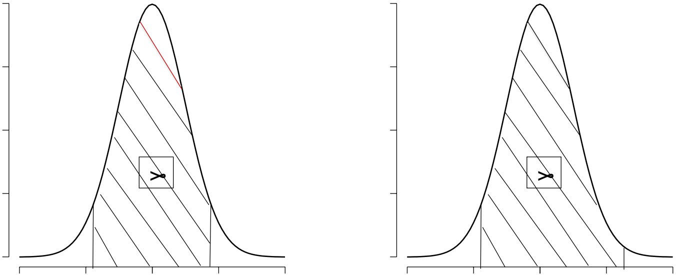
- As estimativas pontuais são resumos muito radicais da distribuição *a posteriori* $\pi(\theta \mid \mathrm{x})$.
- As regiões de credibilidade são mais informativas pois são escolhidas de forma a conter parte substancial da massa probabilística da distribuição *a posteriori*.
- Uma região do espaço paramétrico $R_\theta(\mathrm{x}) \subseteq \Theta$ é uma região com credibilidade γ se

$$P(\theta \in R_\theta(\mathrm{x}) \mid \mathrm{x}) \geq \gamma \tag{1}$$

- Obs: No caso contínuo

$$P(\theta \in R_\theta(\mathrm{x}) \mid \mathrm{x}) = \int_{R_\theta(\mathrm{x})} \pi(\theta \mid \mathrm{x}) d\theta = \gamma$$

Estimação por Regiões: Regiões HPD

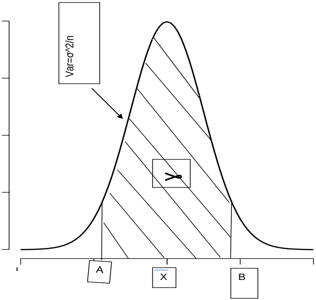


- Uma região de credibilidade tem uma interpretação probabilística direta pois não é aleatória.
- Concluímos que a probabilidade *a posteriori* de θ pertencer à região de credibilidade $R_\theta(\mathrm{x})$ é γ.

Estimação por Regiões

Exemplo: Suponha que $X|\mu \overset{iid}{\sim} Normal(\mu, \sigma^2)$, σ^2 conhecido, e que a incerteza *a priori* sobre μ é bem descrita pela distribuição de Jeffreys. Encontre regiões de $\mathbb{R}$ com credibilidade γ.

- A distribuição *a posteriori* é $\mu \mid \mathrm{x} \sim Normal(\bar{x}, \sigma^2/n)$.
  - simétrica e unimodal



- A região com credibilidade γ é um intervalo $[A, B]$ definido sob a distribuição *a posteriori* de μ tal que $P(A < \mu < B \mid \boldsymbol{x}) = \gamma$

Estimação por Regiões

- Devemos ter

$$Z_{\alpha_1} = \frac{A - \bar{x}}{\sqrt{\sigma^2/n}} \qquad Z_{\alpha_2} = \frac{B - \bar{x}}{\sqrt{\sigma^2/n}}$$

- Um intervalo com credibilidade γ para μ tem a seguinte forma

$$IC_\gamma(\mu) = \left[\bar{x} + \frac{\sigma}{\sqrt{n}} Z_{\alpha_1}; \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{\alpha_2}\right]$$

  onde Z_{α_i} é encontrado na tabela da distribuição $N(0,1)$ e tal que $P(Z > |Z_{\alpha_i}|) = \alpha_i$, $i = 1, 2$ e $\alpha_1 + \alpha_2 = 1 - \gamma$.

- Note que os limites são números determinados a partir da distribuição *a posteriori* para μ.

Estimação por Regiões

- Um intervalo com credibilidade γ e simétrico em torno da média *a posteriori* $\bar{x}$ é

$$IC_\gamma(\mu) = \left[\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{\frac{1-\gamma}{2}}\,;\, \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{\frac{1-\gamma}{2}}\right]$$

onde $Z_{\frac{1-\gamma}{2}}$ é encontrado na tabela da distribuição $N(0,1)$ e tal que $P(Z > Z_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2}$.

Estimação por Regiões

Coloquemos números: Assuma que $\bar{x} = 10$, $\sigma^2 = 4$ e $n = 4$. Construa intervalos com credibilidade $\gamma = 0,95$ e compare-os.

- Se $\alpha_1 = \alpha_2 = 0,025$

  $$ICr(\mu \mid \boldsymbol{x}) = [10 - 1,96; 10 + 1,96] = [8,04; 11,96] \Rightarrow Amp = 3,92.$$

  $\Rightarrow$ A probabilidade da média populacional estar entre $8,04$ e $11,96$ é $0,95$.

- Se $\alpha_1 = 0,03$ e $\alpha_2 = 0,02$

  $$ICr(\mu \mid \boldsymbol{x}) = [10 - 1,88; 10 + 2,05] = [8,12; 12,05] \Rightarrow Amp = 3,93$$

- Se $\alpha_1 = 0,01$ e $\alpha_2 = 0,04$

  $$ICr(\mu \mid \boldsymbol{x}) = [10 - 2,33; 10 + 1,75] = [7,67; 11,75] \Rightarrow Amp = 4,08$$

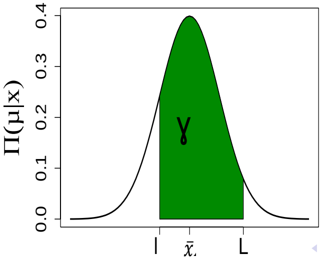
- Se $\alpha_1 = 0,00$ e $\alpha_2 = 0,05$

  $$ICr(\mu \mid \boldsymbol{x}) = [-\infty; 10 + 1,64] = [-\infty; 11,64] \Rightarrow Amp = \infty$$

Todos são intervalos com credibilidade $0,95$ e todos são interpretados probabilisticamente. Como escolher?

Estimação por Regiões

- As regiões de credibilidade como definida na equação (1) são, em certo sentido, melhores resumos *a posteriori* do que as estimativas pontuais.
- No entanto, existem infinitas (no caso contínuo) regiões com credibilidade γ.
- Como está definida em (1), a região de credibilidade pode conter valores de θ que estejam na regiões com menos massa probabilística que alguns outros θ que estejam fora da região construída.



Estimação por Regiões

- Buscamos a melhor! Como definir melhor?
- Buscamos a região que contém todos os valores de θ mais prováveis *a posteriori*.

Estimação por Regiões: regiões HPD

Uma região do espaço paramétrico $R_\theta(\mathrm{x}) \subseteq \Theta$, com credibilidade γ, é uma região com mais alta densidade *a posteriori* (Highest Posterior Density) se

(1)

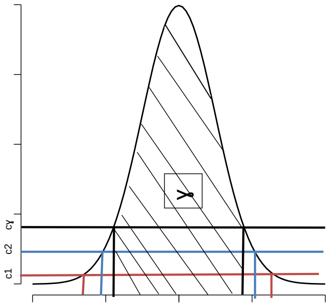
$$R_\theta(\mathrm{x}) = \{\theta \in \Theta : \pi(\theta \mid \mathrm{x}) \geq C_\gamma\}$$

onde $C_\gamma > 0$ é a maior constante tal que

(2)

$$P(\theta \in R_\theta(\mathrm{x}) \mid \mathrm{x}) \geq \gamma.$$

Estimação por Regiões: Regiões HPD



- Se a região HPD é um intervalo $[A, B] \rightarrow \pi(A \mid \mathrm{x}) = \pi(B \mid \mathrm{x})$.
- A região HPD é única.
- O intervalo HPD é o de menor amplitude (A região HPD é a de menor área ou menor volume).

Estimação por Regiões; Regiões HPD

Exemplo: (cont.) Suponha que $X|\mu \overset{iid}{\sim} Normal(\mu, \sigma^2)$, σ^2 conhecido, e que a incerteza *a priori* sobre μ é bem descrita pela distribuição de Jeffreys. Encontre o intervalo HPD com credibilidade γ.

Solução: Como a distribuição *a posteriori* é $\mu \mid \boldsymbol{x} \sim N(\bar{x}, \sigma^2/n)$ é unimodal e simétrica em torno de $\bar{x}$, o intervalo simétrico em torno da média *a posteriori* $\bar{x}$ o Intervalo HPD

$$IC_\gamma(\mu) = \left[\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{\frac{1-\gamma}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{\frac{1-\gamma}{2}}\right]$$

onde $Z_{\frac{1-\gamma}{2}}$ é encontrado na tabela da distribuição $N(0,1)$ e tal que $P(Z > Z_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2}$.

- Não foi coincidência que na análise numérica que fizemos, ele foi o mais curto.

Estimação por Regiões: regiões HPD

OBSERVAÇÕES Importantes Referentes ao Exemplo:

- Pós-experimentalmente, o intervalo clássico com confiança γ tem os mesmos limites que o intervalo simétrico (HPD) em torno da média *a posteriori* $\bar{x}$.
- Apesar disto, as inferências sobre μ não são as mesmas.

Num contexto geral,

- A interpretação clássica é em termos de confiança, não é em termos de probabilidade;
- Em sua construção, o intervalo de confiança leva em conta todas as as amostras (observadas e não) e não apenas a que foi observada.
- O intervalo HPD segue o Princípio da Verossimilhança.

Estimação por Regiões: regiões HPD

Outro Exemplo: (Paulino, Turkeman, Murteira, 2003)

Se $X_1, \ldots, X_n | \theta \overset{iid}{\sim} Bernoulli(\theta)$ e, *a priori*, $\theta \sim Beta(2, 1)$ a distribuição *a posteriori* é $\theta \mid X_1, \ldots, X_n \sim Beta(9, 3)$.

- A distribuição *a posteriori* é unimodal e assimétrica
- A região HPD será um intervalo.

Os intervalos HPD e central são

IC	90%	Amp.	95%	Amp.
HPD	[0.566;0.944]	0.378	[0.516;0.959]	0.443
Central	[0.530;0.921]	0.391	[0.482;0.94]	0.458

Estimação por Regiões: regiões HPD

Observação importante sobre as regiões HPD

- As regiões de credibilidade (HPD ou não), se transformadas, preservam a mesma credibilidade.
- Mas as regiões HPD não são invariantes sob transformações um-a-um.
- Assim, se $R_\theta(\mathrm{x})$ é uma região HPD para θ e $\phi = g(\theta)$, a região transformada $g(R_\theta(\mathrm{x}))$
  - + não é, necessariamente, uma região HPD para ϕ;
  - + será uma região com a mesma credibilidade γ.

Exemplo: Se $\mathrm{X}|\theta \overset{iid}{\sim} \exp(\theta)$ e, *a priori*, $\theta \sim Gamma(a, b)$ então

+ A distribuição *a posteriori* é

$$\theta \mid \mathrm{x} \sim Gamma\left(n + a, b + \sum x_i\right) = Gamma(A, B).$$

$$\pi(\theta \mid \boldsymbol{x}) \propto \theta^{A-1} \exp\{-B\theta\}$$

+ Como é uma distribuição unimodal, a região HPD é um intervalo. Denote-o por $[\bar{\theta}; \hat{\theta}] \Rightarrow \pi(\bar{\theta} \mid \mathrm{x}) = \pi(\hat{\theta} \mid \mathrm{x})$.

Estimação por Regiões: regiões HPD

Considere a transformação $\alpha = 2B\theta$. Consequentemente, segue que

* $\alpha \mid \mathrm{x} \sim Gamma(A, 1/2) \Rightarrow \pi(\alpha \mid \boldsymbol{x}) \propto \alpha^{(A-1)} \exp\{-\alpha/2\}$
* $\gamma = P(\bar{\theta} < \theta < \hat{\theta} \mid \mathrm{x}) = P(2B\bar{\theta} < \alpha < 2B\hat{\theta} \mid \mathrm{x})$
* Mais ainda , segue que avaliando $\pi(\alpha \mid \boldsymbol{x})$ em $2B\bar{\theta}$

$$
\begin{aligned}
\pi(2B\bar{\theta} \mid \mathrm{x}) &\propto (2B\bar{\theta})^{A-1} \exp\{-2B\bar{\theta}/2\} \\
&\propto (2B)^{A-1} \bar{\theta}^{A-1} \exp\{-B\bar{\theta}\} \\
&\propto (2B)^{A-1} \hat{\theta}^{A-1} \exp\{-B\hat{\theta}\} = \pi(2B\hat{\theta} \mid \mathrm{x})
\end{aligned}
$$

* Assim, o intervalo $[2B\bar{\theta}; 2B\hat{\theta}]$ é um intervalo HPD para α.

Estimação por Regiões: regiões HPD

Considere uma outra transformação $\alpha = \frac{1}{\theta}$. Segue que

* $\alpha \mid \mathrm{x} \sim GamaInv(A, B) \Rightarrow \pi(\alpha \mid \boldsymbol{x}) \propto \left(\frac{1}{\alpha}\right)^{A+1} \exp\{-B/\alpha\}$
* $\gamma = P(\bar{\theta} < \theta < \hat{\theta} \mid \mathrm{x}) = P(\frac{1}{\hat{\theta}} < \alpha < \frac{1}{\bar{\theta}} \mid \mathrm{x})$
* Neste caso, avaliando $\pi(\alpha \mid \boldsymbol{x})$ em $1/\bar{\theta}$ temos

$$\begin{aligned} \pi\left(\frac{1}{\bar{\theta}} \mid \mathrm{x}\right) &\propto (\bar{\theta})^2 (\bar{\theta})^{A-1} \exp\{-B\bar{\theta}\} \\ &\propto (\bar{\theta})^2 \hat{\theta}^{A-1} \exp\{-B\hat{\theta}\} \neq \pi\left(\frac{1}{\hat{\theta}} \mid \mathrm{x}\right) \end{aligned}$$

* Assim $[1/\hat{\theta}; 1/\bar{\theta}]$ não é um intervalo HPD para α.

Estimação por Regiões e Teoria de decisão

- A construção de uma região HPD é um problema de decisão.
- Buscamos uma região $R \subseteq \Theta$ com credibilidade γ tal que

$$\mathbb{P}(\theta \in R \mid x) = \int_R \pi(\theta \mid \boldsymbol{x}) d\theta = \gamma.$$

- Fixado γ NÃO existe apenas uma região R satisfazendo esta condição
- Nossa tarefa é buscar pela região que melhor resume a informação *a posteriori*.$\leftarrow$ um problema de decisão.

Estimação por Regiões e Teoria de decisão

Proposição: Seja $\pi(\theta \mid \boldsymbol{x})$ a distribuição *a posteriori* para θ. Seja um número real γ, $0 < \gamma < 1$ e seja o conjunto não-vazio $\mathcal{A}$ de regiões com credibilidade γ, isto é,

$$\mathcal{A} = \{R \subseteq \Theta : \mathbb{P}(\theta \in R \mid \boldsymbol{x}) = \gamma\}$$

e considere a função de perda

$$L(R, \theta) = K\|R\|I(\theta \in R)$$

em que $R \in \mathcal{A}$, $\theta \in \Theta, K > 0$ e $\|R\|$ denota o comprimento, área ou volume de R. Então R é ótima se, e somente se, $\pi(\theta_1 \mid \boldsymbol{x}) \geq \pi(\theta_2 \mid \boldsymbol{x})$ para todo $\theta_1 \in R$ e $\theta_2 \notin R$, salvo para conjuntos de medida nula.

Estimação por Regiões e Teoria de decisão

Prova: Para todo $R \in \mathcal{A}$ temos

$$
\begin{aligned}
E(L(R,\theta) \mid \boldsymbol{x}) &= \int_{\Theta} K\|R\| I(\theta \in R)\pi(\theta \mid \boldsymbol{x}) d\theta \\
&= K\|R\| \int_{\Theta} \pi(\theta \mid \boldsymbol{x}) I(\theta \in R) d\theta \\
&= K\|R\| \int_{R} \pi(\theta \mid \boldsymbol{x}) d\theta \\
&= K\|R\| \mathbb{P}(\theta \in R \mid \boldsymbol{x}) \\
&= K\|R\|\gamma
\end{aligned}
$$

$\Leftarrow\rfloor$ Suponha que $\pi(\theta_1 \mid \boldsymbol{x}) \geq \pi(\theta_2 \mid \boldsymbol{x})$ para todo $\theta_1 \in R$ e $\theta_2 \in R^c$

$\Rightarrow$ R está na região de maior massa probabilística *a posteriori*
$\Rightarrow$ Como para todo $R \in \mathcal{A}$ temos que $\mathbb{P}(\theta \in R \mid \boldsymbol{x})$, então $\|R\|$ é mínima entre todas as regiões $R \in \mathcal{A}$
$\Rightarrow$ $E(L(R,\theta) \mid \boldsymbol{x}) = K\|R\|\gamma$ é mínima $\Rightarrow$ R é a região ótima.

Estimação por Regiões e Teoria de decisão

$\Rightarrow \rfloor$ Suponha que R é uma região ótima $\Rightarrow E(L(R, \theta) \mid \boldsymbol{x})$ é mínima $\Rightarrow K\|R\|\gamma$ é mínima $\Rightarrow \|R\|$ é mínima.

Como o volume sob $\pi(\theta \mid \boldsymbol{x})$ restrito a R é fixo e igual a γ então $\|R\|$ é mínimo se R contém todos os valores $\theta_1 \in \Theta$ que têm mais alta densidade *a posteriori*. Além disto, todos os pontos $\theta_2 \in R^c$ têm menos densidade *a posteriori* que qualquer um dos pontos θ_1 pertencentes à R.

Regiões HPD: Algoritmo

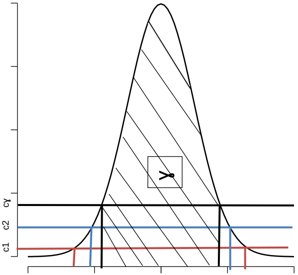
Na prática, não é muito simples obter as regiões HPD de forma exata. A construção do algoritmo passa por

*P*1 : Construir uma rotina para determinar soluções para

$$\pi(\theta \mid \mathrm{x}) = C,$$

para todo C, definindo as regiões

$$R_\theta^C(\mathrm{x}) = \{\theta \in \Theta : \pi(\theta \mid \mathrm{x}) \geq C\}$$



Regiões HPD: Algoritmo

$P2:$ Construir uma rotina para calcular

$$P(\theta \in R_{\theta}^{C}(\mathrm{x}) \mid \mathrm{x}) = \int_{R_{\theta}^{C}(\mathrm{x})} \pi(\theta \mid \mathrm{x}) d\theta.$$

$P3:$ A região HPD é determinada quando encontrarmos o maior $C = C_{\gamma}$ tal que

$$P(\theta \in R_{\theta}^{C_{\gamma}}(\mathrm{x}) \mid \mathrm{x}) = \int_{R_{\theta}^{C_{\gamma}}(\mathrm{x})} \pi(\theta \mid \mathrm{x}) d\theta \geq \gamma$$

Regiões HPD: Algoritmo

Se a região HPD é um intervalo e se temos uma amostra $\theta_1, \ldots,$ θ_n da distribuição *a posteriori* $\pi(\theta \mid \mathrm{x})$, então se pode aproximar os intervalos HPD como segue:

Algoritmo (Chen and Shao (J. Comp Graph. Stat., 1999)):

$P1$: Ordene a amostra $\theta_{(1)}, \ldots, \theta_{(n)}$

$P2$: Determine os intervalos com credibilidade γ fazendo

$$R_i(x) = [\theta_{(i)}; \theta_{i+\lfloor n\gamma \rfloor - 1}], \forall i = 1, \ldots n - \lfloor n\gamma \rfloor + 1.$$

$P3$: O intervalo HPD é o intervalo $R_{i_0}(x)$ tal que

$$\theta_{i_0+\lfloor n\gamma \rfloor - 1} - \theta_{i_0} = \min_i \left\{ \theta_{i+\lfloor n\gamma \rfloor - 1} - \theta_i, \; i = 1, \ldots, n - \lfloor n\gamma \rfloor + 1 \right\}.$$

onde $\lfloor a \rfloor$ é o maior inteiro menor ou igual que a.

Regiões HPD: Algoritmo

Exemplo: Suponha que queiramos obter um intervalo com credibilidade $\gamma = 0,8$

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
1	3	8	8	9	12	14	18	26	30
$\theta_{(1)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(4)}$	$\theta_{(5)}$	$\theta_{(6)}$	$\theta_{(7)}$	$\theta_{(8)}$	$\theta_{(9)}$	$\theta_{(10)}$

Temos que $n - \lfloor n\gamma \rfloor + 1 = 10 - \lfloor 10 \times 0.8 \rfloor + 1 = 3$

Candidatos

$$R_1 = [\theta_{(1)}, \theta_{(1+8-1)}] = [\theta_{(1)}, \theta_{(8)}] = [1, 18] \rightarrow Amp = 17, p = 0,8$$

$$R_2 = [\theta_{(2)}, \theta_{(2+8-1)}] = [\theta_{(2)}, \theta_{(9)}] = [3, 26] \rightarrow Amp = 23, p = 0,8$$

$$R_3 = [\theta_{(3)}, \theta_{(3+8-1)}] = [\theta_{(3)}, \theta_{(10)}] = [8, 30] \rightarrow Amp = 22, p = 0,8$$

O Intervalo HPD com credibilidade $0,8$ é

$$R_1 = [1, 18]$$