

MAT02034 - Métodos bayesianos para análise de dados

Apresentações

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Porto Alegre, 2022

O professor

Olá!



Olá!

- ▶ Desde outubro de 2017 eu sou Professor do Departamento de Estatística e faço parte do Corpo Docente do Programa de Pós Graduação em Epidemiologia da Universidade Federal do Rio Grande do sul (UFRGS). Além disso, eu atuo como pesquisador no Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil).
- ▶ Eu me formei Bacharel em Estatística pelo Departamento de Estatística da UFRGS em 2007, e Mestre (2010) e Doutor (2014) em Estatística pelo Programa de Pós Graduação em Estatística da Universidade Federal de Minas Gerais.
- ▶ A minha dissertação de mestrado, intitulada *Técnicas estatísticas para avaliação de novos marcadores de risco: aplicações envolvendo o Modelo de Cox*, foi orientada pelos Professores Enrico A. Colosimo e Maria do Carmo P. Nunes.

Olá!

- ▶ A minha tese de doutorado, intitulada *Análise hierárquica de múltiplos sistemas reparáveis*, foi orientada pelos Professores Enrico A. Colosimo e Gustavo L. Gilardoní.
- ▶ Os meus interesses de pesquisa são Inferência causal em epidemiologia, Análise de mediação, Modelos de predição de risco e Análise de sobrevivência.
- ▶ Em estatística aplicada eu tenho interesse na epidemiologia do Diabetes Mellitus.

A disciplina

Objetivos

- ▶ Apresentar e discutir os conceitos fundamentais do método **Monte Carlo via Cadeias de Markov (MCMC)** no contexto da **inferência bayesiana**.
- ▶ Capacitar o aluno para a utilização e interpretação de modelos estatísticos sob enfoque bayesiano.
- ▶ Estimular o aluno a desenvolver espírito crítico e maturidade de julgar aplicações de inferência bayesiana.

Organização

- ▶ **Disciplina:** Métodos bayesianos para análise de dados
- ▶ **Turma:** U
- ▶ **Modalidade:** Ensino remoto emergencial (**Moodle**)
- ▶ **Professor:** Rodrigo Citton Padilha dos Reis
 - ▶ e-mail: citton.padilha@ufrgs.br ou rodrigocpdosreis@gmail.com
 - ▶ Sala: B215 do Instituto de Matemática e Estatística

Aulas e material didático

- ▶ **Aulas** (teóricas e práticas)
 - ▶ Exposição e **discussão** dos conteúdos
 - ▶ Faremos leituras semanais de artigos e capítulos de livros
 - ▶ Exemplos
- ▶ **Notas de aula**
 - ▶ Slides
 - ▶ Arquivos de rotinas em R (e outras linguagens)
- ▶ **Exercícios**
 - ▶ Listas de exercícios
 - ▶ Para casa
 - ▶ Questionários do Moodle
- ▶ **Canais de comunicação:**
 - ▶ Durante as aulas
 - ▶ Moodle: aulas, materiais, listas de exercícios
 - ▶ Sala de aula virtual: notas das avaliações
 - ▶ e-mail do professor

Aulas e material didático



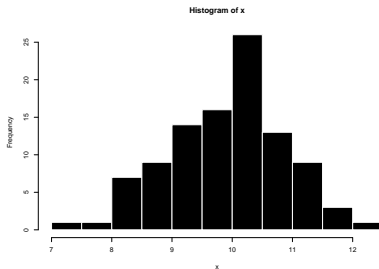
- ▶ **Aulas:** Quartas, das 10hs 30min às 11hs 30min, no MConf do Moodle da disciplina
 - ▶ As aulas serão realizadas de maneira **síncrona** com **gravação** e disponibilizadas para posterior consulta

Aulas e material didático

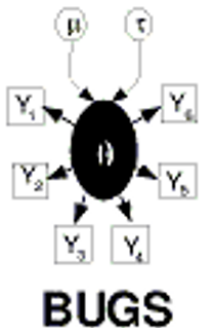


- ▶ Exemplos e exercícios com o apoio do computador:
 - ▶ R e RStudio

```
hist(x, col = 'black', border = 'white')
```



Aulas e material didático



Conteúdo programático

▶ Área 1

- ▶ Algoritmos de Metropolis-Hastings e Gibbs-Sampling
- ▶ Aplicativos para MCMC
- ▶ Métodos para avaliação de convergência

▶ Área 2

- ▶ Modelos hierárquicos: definições Básicas
- ▶ Modelo linear geral (regressão e análise de variância com 1 fator)
- ▶ Modelo de regressão logística
- ▶ Outros modelos
- ▶ Métodos para avaliação de ajuste de modelos
- ▶ Métodos para seleção de covariáveis

Avaliação

- ▶ Serão realizadas pelo menos uma (e no máximo três) avaliação(ões) pontuais de cada área por meio de questionários e tarefas do Moodle
- ▶ Será realizada uma avaliação parcial de cada área por meio de questionários e tarefas do Moodle
- ▶ Cada atividade de avaliação vale 10 pontos
- ▶ Será realizado um teste no Moodle (individual) como atividade de recuperação (*TR*)
 - ▶ Para os alunos que não atingirem o conceito mínimo
 - ▶ **Este teste abrange todo o conteúdo da disciplina**

Avaliação

$$MF = [(Avaliação\ parcial\ Área_1 \times 3) + (Avaliação\ parcial\ Área_2 \times 3) + (nota\ média\ das\ avaliações\ pontuais \times 4)]/10.$$

- ▶ **A:** $9 \leq MF \leq 10$
- ▶ **B:** $7,5 \leq MF < 9$
- ▶ **C:** $6 \leq MF < 7,5$
- ▶ Se $MF < 6$ o aluno poderá realizar o teste de recuperação e neste caso

$$MF' = MF \times 0,4 + TR \times 0,6$$

- ▶ **C:** $MF' \geq 6$
- ▶ **D:** $MF' < 6$

Referências bibliográficas



Principais

Albert, J. **Bayesian Computation with R**. New York: Springer-Verlag, 2009.

Cowles, M. K. **Applied Bayesian Statistics with R and OpenBugs examples**. New York: Springer, 2013.

Complementares

Marin, J. M, Robert, C. **Bayesian essentials with R**. New York: Springer 2014.

Estatística bayesiana

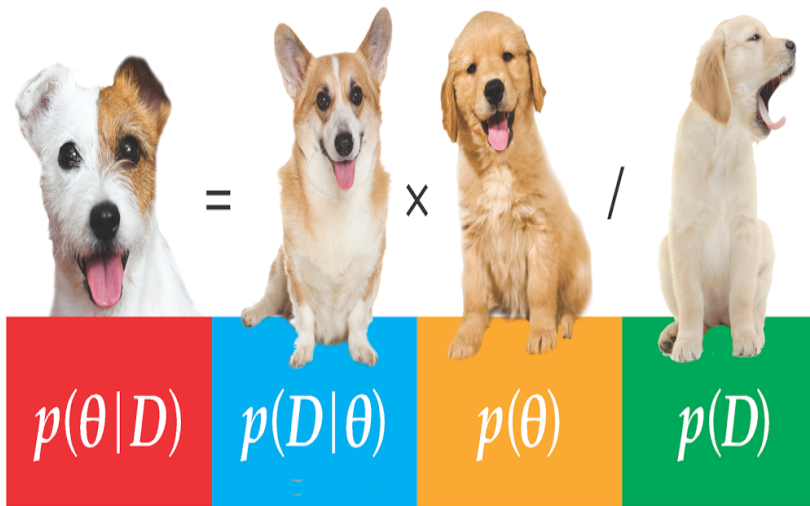
Estatística bayesiana

- ▶ A **estatística bayesiana** expandiu enormemente sua cobertura nas últimas três décadas.
- ▶ Os **métodos bayesianos** são agora aplicados a uma ampla variedade de **empreendimentos científicos, sociais e empresariais**, incluindo áreas como:
 - ▶ astronomia,
 - ▶ biologia,
 - ▶ economia,
 - ▶ educação,
 - ▶ engenharia,
 - ▶ genética,
 - ▶ marketing,
 - ▶ medicina,
 - ▶ psicologia,
 - ▶ saúde pública,
 - ▶ esportes, entre muitos outros.

Estatística bayesiana

- ▶ Existem certas situações em que a estatística bayesiana aparece como o único paradigma que oferece soluções viáveis e isso se tornou possível devido ao tremendo **desenvolvimento da teoria, metodologia, computação e aplicações bayesianas**.
- ▶ O tópico tornou-se a vanguarda da estatística prática com o advento de **computadores de alta velocidade** e **técnicas computacionais sofisticadas**, especialmente na forma de **métodos Monte Carlo via Cadeias de Markov** e **abordagens baseadas em amostras**.
 - ▶ De fato, a modelagem bayesiana em problemas complexos combina livremente componentes de diferentes tipos de abordagens de modelagem com informações *a priori* estruturais, sem restrições se tais combinações de modelos já foram estudadas ou analisadas antes.

Inferência bayesiana


$$p(\theta|D) = p(D|\theta) \times p(\theta) / p(D)$$

Inferência bayesiana

- ▶ A distribuição *a posteriori*, $p(\theta|D)$, é a descrição completa do conhecimento corrente sobre θ obtido da quantificação da informação *a priori* em $p(\theta)$ e da informação amostral em $p(D|\theta)$, materializando-se na expressão matemática

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{\int_{\Theta} p(\theta)p(D|\theta)d\theta} \propto p(\theta)p(D|\theta), \theta \in \Theta.$$

Inferência bayesiana

Estimação pontual (caso de θ escalar)

- ▶ Moda *a posteriori*

$$\hat{\theta} = \arg \max_{\theta \in \Theta} p(\theta|D) = \arg \max_{\theta \in \Theta} p(\theta)p(D|\theta).$$

- ▶ Média *a posteriori*

$$\hat{\theta} = E[\theta|D] = \int_{\Theta} \theta p(\theta|D) d\theta.$$

- ▶ Mediana *a posteriori*: $\hat{\theta}$ o menor valor tal que

$$\Pr(\theta \geq \hat{\theta}|D) \geq 1/2 \quad \text{e} \quad \Pr(\theta \leq \hat{\theta}|D) \geq 1/2.$$

Inferência bayesiana

Estimação por intervalo (caso de θ escalar)

$$\Pr(\theta_1 \leq \theta \leq \theta_2 | D) = \int_{\theta_1}^{\theta_2} p(\theta | D) d\theta.$$

Algumas questões

- ▶ Como especificar a distribuição *a priori*?
 - ▶ θ é discreto ou contínuo?
 - ▶ θ é escalar ou vetor?
- ▶ A distribuição *a posteriori* é obtida analiticamente ou precisa ser aproximada?
- ▶ No caso em que a distribuição *a posteriori* não pode ser expressa de forma fechada, quais aproximações podemos utilizar?
 - ▶ Aproximações analíticas e numéricas (aproximação normal, método de Laplace, INLA)?
 - ▶ Aproximações estocásticas (Monte Carlo, MCMC)?
 - ▶ Métodos híbridos?
- ▶ Como implementar estas aproximações?
- ▶ Como avaliar a qualidade das aproximações?

Algumas questões

- ▶ Estas questões constituem o nosso objeto de estudo:
os **Métodos bayesianos para análise de dados.**

Próxima aula

- ▶ Introdução ao raciocínio bayesiano.

Por hoje é só!

Sejam tod@s bem-vind@s!

