AGA 0505 - Análise de Dados em Astronomia

Inferência na Análise de Dados Bayesiana

Laerte Sodré Jr.

1o. semestre, 2025

aula de hoje:

1. o teorema de Bayes e a análise de dados
2. inferência de parâmetros
3. quando estimativas bayesianas e frequentistas divergem?
4. modelos gerativos
5. ABC: *approximate bayesian computation*

As questões mais importantes da vida são, na maior parte, apenas problemas de probabilidades.

Pierre Simon de Laplace

o teorema de Bayes

- o teorema de Bayes:
  - Bayes $\sim$1740, Laplace 1774
  - a e b: variáveis aleatórias
  - probabilidade conjunta de a e b, dado um certo modelo M:

$$P(a,b|M) = P(a|b,M) \times P(b|M) =$$
$$= P(b|a,M) \times P(a|M)$$

e, portanto:

$$P(a|b,M) = \frac{P(b|a,M) \times P(a|M)}{P(b|M)}$$

o teorema de Bayes e a análise de dados

- **a) inferência dos parâmetros de um modelo:**
  - temos um conjunto de dados D que queremos explicar com um modelo M que tem um conjunto de parâmetros w
  - objetivo: determinar w
  - exemplo: $D = \{x_1, y_1, x_2, y_2, ..., x_n, y_n\}$, $M : y = a + bx + \epsilon, \quad w = \{a, b\}$
  - teorema de Bayes ($a \to w$, $b \to D$):

$$P(w|D, M) = \frac{P(D|w, M)P(w|M)}{P(D|M)}$$

$$P(D|M) = \int P(D|w, M)P(w|M)dw$$

- **b) comparação de modelos:**
  - temos dois modelos, M_1 e M_2, para os dados D; qual modelo escolher?
  - exemplo: M_1 : reta, M_2 : parábola
  - teorema de Bayes ($a \to M$, $b \to D$):

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

  - comparação de M_1 e M_2:

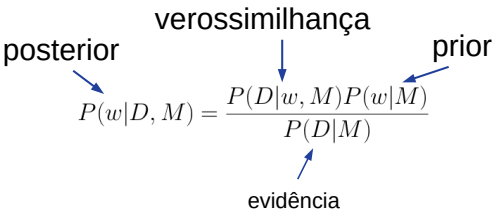
$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \frac{P(M_1)}{P(M_2)}$$

  escolhemos o modelo mais provável

o teorema de Bayes e a análise de dados

- cada termo tem um nome:
  - $P(w|D,M)$: o **posterior**: a distribuição de probabilidades *a posteriori* de w
  - $P(D|w,M)$: a **verossimilhança** (*likelihood*): a probabilidade dos dados com um modelo M e parâmetros w

    aqui P é a probabilidade dos dados, não uma função dos parâmetros!
  - $P(w|M)$: o **prior** de w, a probabilidade *a priori* dos valores dos parâmetros; o que se sabe de w independentemente dos dados
  - $P(D|M)$: a **evidência** dos dados (ou *distribuição marginal dos dados*)



o teorema de Bayes e a análise de dados

- **c) previsão de novos dados**
- podemos usar métodos bayesianos para fazer previsões probabilísticas sobre observações futuras
  - seja $P(w|D)$ o posterior dos parâmetros baseado em dados D (e assumindo um modelo M)
  - seja d um novo dado: a *distribuição preditiva* de d é

$$P(d|D) = \int P(d, D, w)dw = \int P(d|w)P(w|D)dw$$

  - a distribuição preditiva bayesiana considera automaticamente tanto a incerteza nas estimativas dos parâmetros ($P(w|D)$) quanto a aleatoriedade inerente ao processo de geração de dados ($P(d|w)$)

porque usar o teorema de Bayes?

- permite uma abordagem lógica e totalmente probabilística da análise de dados
- oferece maior flexibilidade na construção de modelos
  ex.: o método da máxima verossimilhança pode ser considerado um caso particular
- pode-se incluir informações adicionais de muitas fontes, como outros dados ou opiniões de especialistas
- quantifica a incerteza nos parâmetros e nas predições: *intervalos de credibilidade*

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)}$$

o teorema de Bayes e a inferência de parâmetros

- dados D, modelo M, parâmetros w
- teorema de Bayes:

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)}$$

- estimativa bayesiana dos parâmetros:

  $P(w|D,M) \propto P(D|w,M)P(w|M)$,

  não depende da evidência, já que ela não depende de w

inferência dos parâmetros w:

- estimativa frequentista:
  - os dados D são uma amostra de uma população com parâmetros w fixos
  - máxima verossimilhança:
    estimativa de ponto - valor de w que maximiza $\mathcal{L}(w) = P(D|w,M)$
- inferência bayesiana:
  - os dados D são fixos e os parâmetros têm uma *distribuição de probabilidades*, $P(w|D,M)$

    estimativa de uma distribuição de probabilidades

exemplo 1: inferência da média de uma gaussiana com um prior uniforme

- temos uma única observação, x_1, que queremos modelar como uma gaussiana de média μ desconhecida e desvio padrão σ conhecido:

  queremos determinar (a distribuição de) μ

- dados: $\{x_1\}$
- a verossimilhança é $N(\mu, \sigma)$:

$$P(x_1|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{(x_1-\mu)^2}{2\sigma^2}\Big]$$

- vamos supor um prior uniforme para μ:

$$P(\mu) = \text{cte}$$

- nesse caso, o posterior de μ é proporcional à verossimilhança:

$$P(\mu|x_1,\sigma) \propto P(x_1|\mu,\sigma)P(\mu) \propto P(x_1|\mu,\sigma)$$

ou

$$P(\mu|x_1,\sigma) \propto \exp\Big[-\frac{(x_1-\mu)^2}{2\sigma^2}\Big]$$

- isto é, o posterior também é gaussiano,

$$P(\mu|x_1,\sigma) \sim N(x_1,\sigma),$$

com média x_1 e desvio padrão σ

exemplo 2: inferência da média de uma gaussiana com um prior uniforme

- mesmo modelo, mas agora temos n observações que obedecem a uma gaussiana de média μ desconhecida e desvio padrão σ conhecido
- $D = \{x_1, x_2, ...x_n\}$
- prior de μ uniforme
- verossimilhança de um dado: $P(x_i|\mu, \sigma) \sim N(\mu, \sigma)$
- verossimilhança da amostra (supondo dados independentes):

$$P(D|\mu,\sigma) = \prod_{i=1}^{n} P(x_i|\mu,\sigma) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right] \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

- note que $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$, onde $\bar{x} = \frac{1}{n}\sum_i^n x_i$
- logo, o posterior de μ fica: $P(\mu|D,\sigma) \propto P(D|\mu,\sigma) \propto \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] \sim N(\bar{x}, \sigma/\sqrt{n})$

exemplo 3: inferência da média de uma gaussiana com um prior gaussiano

- temos uma única observação, x_1, que queremos modelar como uma gaussiana de média μ desconhecida e desvio padrão σ conhecido: queremos determinar μ
- a verossimilhança é $N(\mu, \sigma)$
- vamos supor para μ um prior gaussiano de média μ_0 e desvio padrão τ_0, $\quad P(\mu) \sim N(\mu_0, \tau_0)$
- posterior de μ: $\quad P(\mu|x_1, \sigma) \propto P(x_1|\mu, \sigma)P(\mu) \propto \exp\left[-\frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau_0^2}\right] \propto \exp\left[-\frac{(\mu-\mu_1)^2}{2\tau_1^2}\right] \sim N(\mu_1, \tau_1)$
- o posterior também é gaussiano com

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{x_1}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

note que se $\tau_0 >> \sigma$, a influência do prior é desprezível

se $\tau_0 << \sigma$, o prior é muito *informativo*

exemplo 4: inferência da média de uma gaussiana com um prior gaussiano

- agora temos n observações: $D = \{x_1, x_2, ...x_n\}$
- modelo: gaussiana de média μ desconhecida e desvio padrão σ conhecido: $P(x_i, \sigma|\mu) \sim N(\mu, \sigma)$
- prior de μ: $P(\mu) \sim N(\mu_0, \tau_0)$
- verossimilhança dos dados (supostos independentes):

$$P(D|\mu) = \prod_{i=1}^{n} P(x_i, \sigma|\mu) \propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

- posterior de μ: $P(\mu|D, \sigma) \sim N(\mu_n, \tau_n)$, onde

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \quad \text{e} \quad \bar{x} = \frac{1}{n}\sum_i^n x_i$$

- para n grande o efeito do prior fica desprezível!
- *o prior não deve dominar o resultado, a menos que seja relevante!*

exemplo 5: regressão linear bayesiana

- temos um conjunto de dados $D = \{x_i, y_i\},\ i = 1, ..., n$ independentes e com a mesma variância σ^2
- modelo: $y_i = a + bx_i + \epsilon_i$, com erros gaussianos: $\epsilon_i \sim N(0, \sigma^2)$
- parâmetros: $w = \{a, b, \sigma\}$
- objetivo: determinar as distribuições de probabilidades de a, b, σ
- teorema de Bayes: $P(a, b, \sigma|D) \propto P(D|a, b, \sigma)P(a, b, \sigma)$
- verossimilhança de uma medida: $P(D_i|a, b, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - y_c(x_i))^2}{2\sigma^2} \right]$,

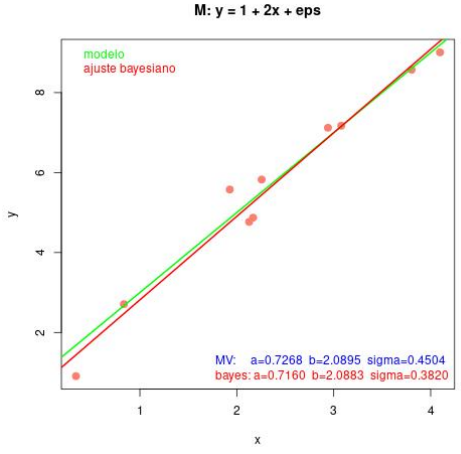
  onde $y_c(x_i) = a + bx_i$

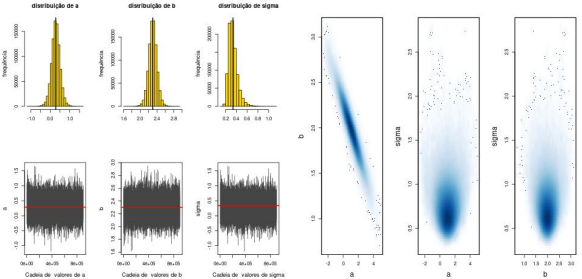
- verossimilhança da amostra: $P(D|a, b, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\sum_i \frac{(y_i - y_c(x_i))^2}{2\sigma^2} \right]$
- vamos supor priores uniformes para a e b e o prior de Jeffreys para σ:

$$P(a, b, \sigma) \propto \frac{1}{\sigma^2}$$

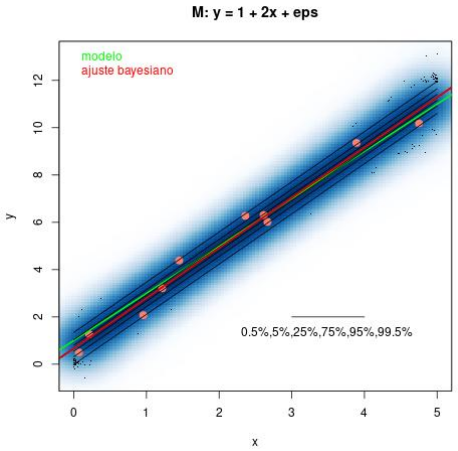
- posterior: $P(a, b, \sigma|D) \propto \sigma^{-(n+2)} \exp\left[-\sum_i \frac{(y_i - y_c(x_i))^2}{2\sigma^2} \right]$
- este posterior pode ser amostrado por MCMC e os parâmetros estimados das cadeias simuladas

exemplo 5: regressão linear bayesiana





exemplo 5: regressão linear bayesiana



o método da máxima verossimilhança

- na abordagem frequentista, a verossimilhança é considerada uma função dos parâmetros:

$$\mathcal{L}(w) = P(D|w)$$

- exemplo: verossimilhança gaussiana para um único dado

$$\mathcal{L}(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_1 - \mu)^2}{2\sigma^2} \right]$$

- note que $\mathcal{L}(w)$ não é uma distribuição de probabilidades: ela não é (em geral) normalizada com respeito aos parâmetros w

- procedimento frequentista para estimativa dos parâmetros de um modelo com base nos dados disponíveis:
a melhor estimativa, $\hat{w}$, é a que maximiza a verossimilhança

- $\hat{w}$ é o ponto no espaço de parâmetros que maximiza $\mathcal{L}(w)$

- incertezas em $\hat{w}$ calculadas a partir da largura de $\mathcal{L}(w)$

quando estimativas bayesianas e frequentistas divergem?

- a estimativa dos parâmetros frequentista é uma *estimativa de ponto*:

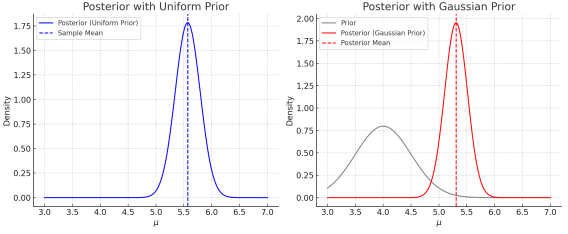
$$\hat{w} = \text{valor de } w \text{ que maximiza } \mathcal{L}(w)$$

- na abordagem bayesiana obtém-se *distribuições de probabilidades* para w:

$$P(w|D) \propto P(D|w)P(w)$$

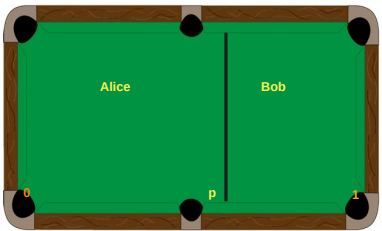
- em muitos casos o máximo do posterior (MAP) é idêntico à estimativa de MV
- em outros casos não; por exemplo:
  - quando se usa priores informativos
  - no trato de hiperparâmetros ou "nuisance parameters"

- com priores uniformes, $\hat{w}$ é a moda do posterior $P(w|D)$
- com priores informativos a moda se desloca



quando estimativas bayesianas e frequentistas divergem?

- em muitos casos o máximo do posterior (MAP) é idêntico à estimativa de MV
- em outros casos não; por exemplo no trato de hiperparâmetros ou “nuisance parameters” *parâmetros que não interessam*



- exemplo: *um jogo de bilhar bayesiano* baseado no trabalho de 1763 de Bayes

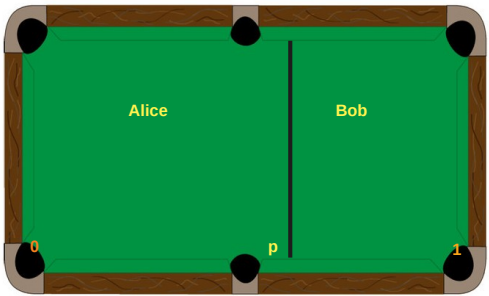
  ver http://jakevdp.github.io/blog/2014/06/06/frequentism-and-bayesianism-2-when-results-differ/

- Carol joga bolas, de costas e sem viés, numa mesa de bilhar que tem uma marca: se elas caem de um lado da marca, Alice ganha um ponto, se caem do outro, Bob ganha um ponto; ganha o jogo quem primeiro fizer 6 pontos
- num certo jogo, após 8 bolas, Alice tem 5 pontos e Bob tem 3
- qual é a probabilidade de Bob ganhar o jogo?

quando estimativas bayesianas e frequentistas divergem?

- após 8 bolas, Alice tem 5 pontos e Bob tem 3
- abordagem frequentista:
  - a probabilidade p da bola cair do lado da Alice é $p = 5/8$
  - a probabilidade p da bola cair do lado de Bob é $1 - p$
  - para Bob ganhar o jogo, ele tem que marcar 3 pontos seguidos
  - probabilidade disso:

$$P_{freq} = (1 - p)^3 = 0.0527$$



quando estimativas bayesianas e frequentistas divergem?

- abordagem Bayesiana:
  - seja B o evento “Bob ganha”
  - dados: $D = \{n_A, n_B\} = \{5, 3\}$
  - p: probabilidade (desconhecida) que a bola caia na área de Alice
  - queremos $P(B|D)$
  - note que o valor de p não interessa! ele é um *nuisance parameter*

- podemos incluir e “sumir” com p via marginalização:

$$P(B|D) = \int P(B, p|D)dp =$$

$$= \int P(B|p, D)P(p, D)dp =$$

$$= \int P(B|p, D)\frac{P(D|p)P(p)}{P(D)}dp =$$

$$= \frac{\int P(B|p, D)P(D|p)P(p)dp}{\int P(D|p)P(p)dp}$$

quando estimativas bayesianas e frequentistas divergem?

abordagem Bayesiana:

- marginalização:

$$P(B|D) = \frac{\int P(B|p,D)P(D|p)P(p)dp}{\int P(D|p)P(p)dp}$$

  - para ganhar a partida, Bob tem que ganhar 3 jogadas seguidas:

$$P(B|p,D) = P(B|p) = (1-p)^3$$

  - vamos supor $P(p)$ uniforme entre 0 e 1
  - verossimilhança: distribuição binomial

$$P(D|p) \propto p^5(1-p)^3$$

- logo,

$$P(B|D) = \frac{\int_0^1 (1-p)^6 p^5 dp}{\int_0^1 (1-p)^3 p^5 dp}$$

- a função beta é:

$$\beta(n,m) = \int_0^1 (1-p)^{n-1} p^{m-1} dp$$

- logo,

$$P(B|D) = \frac{\beta(6+1,5+1)}{\beta(3+1,5+1)} \simeq 0.091$$

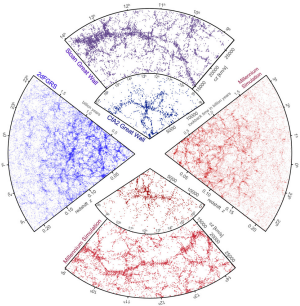
- compare com MV: $P_{freq} = 0.053$!

modelos gerativos

- modelos gerativos:
  inferência baseada em simulações dos dados
- simulamos parâmetros a partir de um prior, $P(w)$,
- usamos esses parâmetros para simular “dados”, e
- usamos os dados simulados para estimar o posterior dos parâmetros
- porquê? há situações onde a verossimilhança é intratável

modelos gerativos

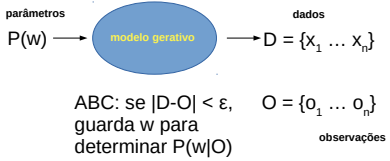
Simulação de dados a partir do prior dos parâmetros e do modelo



- exemplo: estimativa de parâmetros cosmológicos comparando observações e simulações da distribuição de galáxias
- como proceder:
  ao invés de se comparar diretamente observações e modelos , pode-se comparar estatísticas que “resumem” propriedades importantes dos dados tanto nas observações quanto nas simulações
  exemplos: distância média entre galáxias, variância do número de galáxias em esferas de raio 8 Mpc, ...

ABC: Approximate Bayesian Computation

- objetivo: estimar o posterior dos parâmetros w de um modelo, usando um processo gerativo a partir dos priores $P(w)$
- ABC: permitem a inferência bayesiana quando a verossimilhança é intratável, mas os dados podem ser simulados a partir de um modelo
- ABC: apenas dados simulados que concordam com as observações dentro de uma certa "tolerância" são considerados para amostrar o posterior



- algoritmo ABC:
  - 1. amostre w_p do prior $P(w)$
  - 2. simule dados D_p com w_p
  - 3. calcule as estatísticas que sumarizam os dados: x_p =resumo(w_p)
  - 4. aceite w_p se $|x_p - x_{obs}| < \epsilon$ (tolerância)
  - 5. retorne a 1
- atenção: em geral (mas nem sempre),
  - dados contínuos: tolerância $\epsilon > 0$
  - dados discretos: tolerância $\epsilon = 0$

exemplo: estimativa de uma distribuição binomial por ABC

- Dados: $\{1, 0, 0, 0, 1, 0, 1, 0, 1..\}$
  sequência com n valores x_i igual a 0 ou 1, obtidos de uma distribuição binomial com uma certa probabilidade w de sair 1
- objetivo: se a amostra observada tem X valores 1, usar ABC para estimar o posterior de w
- modelo gerativo com prior uniforme:
  - sorteamos um w uniformemente entre 0 e 1
  - simulamos n dados y_i com uma distribuição binomial com parâmetro w
  - determinamos $Y = \sum_i y_i$
    (= número de '1's)

- distância entre os conjuntos de dados x e y:

$$\rho = |X - Y|$$

- exemplo: tolerância zero
  só vamos aceitar parâmetros com distribuições com o mesmo número de 1s, $\rho = 0$