AGA 0505 - Análise de Dados em Astronomia

7. Inferência Bayesiana

Laerte Sodré Jr.

1o. semestre, 2023

aula de hoje:

1. o teorema de Bayes
2. o teorema de Bayes e a análise de dados
3. inferência de parâmetros
4. exemplo: quantos peixes tem no lago?
5. métodos frequentistas x bayesianos
6. priores
7. exemplo: inferência da média de uma gaussiana
8. exemplo: análise sequencial ou conjunta?
9. exemplo: uma modelagem bayesiana- o problema do farol
10. quando estimativas bayesianas e frequentistas divergem?
11. modelos gerativos
12. ABC: *automated 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 ∼1740, Laplace 1774
  - x e y: variáveis aleatórias
  - probabilidade conjunta de x e y, dado um certo modelo M:

$$P(x, y|M) = P(x|y, M) \times P(y|M) =$$
$$= P(y|x, M) \times P(x|M)$$

e, portanto:

$$P(x|y, M) = \frac{P(y|x, M) \times P(x|M)}{P(y|M)}$$

o teorema de Bayes e a análise de dados

- 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
  - teorema de Bayes ($x \to w, y \to D$):

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

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

- comparação de modelos ($x \to M, y \to D$):

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

  - temos dois modelos, M_1 e M_2, para os dados D; qual escolher?

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

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
  - $P(w|M)$: 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)

verossimilhança

posterior

prior

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

evidência

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

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.: máxima verossimilhança é 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

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

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:

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

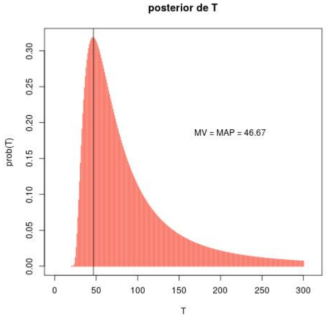
não depende da evidência, já que ela não depende de w

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 $P(D|w,M)$
- inferência bayesiana:
  - os dados D são fixos e os parâmetros têm uma *distribuição de probabilidades*, $P(w|D,M)$

análise de dados bayesiana

- para ser útil, a modelagem dos dados precisa fornecer estimativas dos:
  - parâmetros
  - incertezas nos parâmetros
  - qualidade do ajuste/modelo



- teorema de Bayes:

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

  - parâmetros: distribuição de probabilidades de $P(w|D,M)$
  - incertezas nos parâmetros: estatísticas, 'larguras',... de $P(w|D,M)$
    intervalo de credibilidade de w
  - qualidade do ajuste/modelo: comparação de modelos

exemplo: quantos peixes tem no lago?

- exemplo proposto por Rasmus Bååth
U. de Lund (www.sumsar.net)
- Vamos a um lago e pescamos 7 peixes. Marcamos cada um e os devolvemos ao lago. Alguns dias depois voltamos ao lago, pescamos 20 peixes (dia de sorte!) e verificamos que 3 deles estão marcados. Quantos peixes tem no lago?
- Seja
  - T: número total de peixes no lago: ?
  - M: número de peixes marcados: 7
  - K: número de peixes pescados: 20
  - X: número de peixes pescados que estão marcados: 3

- Solução de MV: supomos que a fração de peixes marcados (X/K) resultante de nossa medida (pescaria) é representativa
- regra de 3: se com K=20 peixes temos X=3 marcados, com T teremos M=7, logo:

$$\hat{T} = MK/X = 7 \times 20/3 \simeq 46.67$$

Lagoa Feia, Coração de Jesus, MG

exemplo: quantos peixes tem no lago?

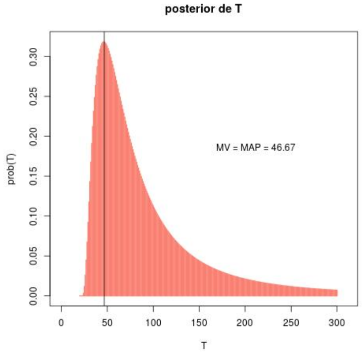
- Dados:
  - T: número total de peixes no lago: ?
  - M: número de peixes marcados: 7
  - K: número de peixes pescados: 20
  - X: número de pescados que estão marcados: 3
- Solução de MV: $\hat{T} \simeq 46.67$
- Pode-se mostrar que a verossimilhança é dada pela distribuição multinomial:

$$P(X|T,M,K) = \frac{\binom{M}{X}\binom{T-M}{K-X}}{\binom{T}{K}}$$

- Para determinar o posterior de T, vamos supor um prior uniforme entre 20 (o número mínimo de peixes no lago) e 300

$$P(T|X,M,K) \propto P(X|T,M,K)P(T)$$

- Note como a cauda da distribuição é extensa: a solução bayesiana é *informativa*!



- MAP: *maximum a posteriori*
- A solução bayesiana é uma distribuição, não um ponto

métodos frequentistas x bayesianos

- frequentistas:
  - probabilidades: frequência de eventos
  - parâmetros: constantes fixas
  - procedimentos estatísticos devem ter propriedades assintóticas bem definidas: um intervalo de confiança de 95% deve conter o “valor verdadeiro” do parâmetro 95% das vezes
- bayesianos:
  - probabilidades medem plausibilidades, não apenas frequências
  - pode-se definir probabilidades para dados, parâmetros, modelos
  - inferência estatística produz distribuições de probabilidades: estatísticas podem ser obtidas dessas distribuições: *intervalo de credibilidade*
- comparações:
  - métodos bayesianos são computacionalmente intensivos
  - em muitos casos os dois enfoques dão o mesmo resultado (MAP)
  - os métodos frequentistas são úteis, mas os futuro é bayesiano

que prior escolher?

- prior: distribuição de probabilidades cujo papel é levar em conta a informação *a priori* sobre os parâmetros do modelo

- Bernoulli (1713):
  princípio da razão insuficiente
  - se conseguimos enumerar um conjunto de possibilidades mutuamente exclusivas e não temos nenhuma razão para achar que uma é mais provável que outra, então devemos atribuir a mesma probabilidade a todas

  - exemplo- probabilidade a priori de cada face de um dado: 1/6

- *priores conjugados*: quando o posterior tem a mesma forma funcional que o prior

  - uma gaussiana é o prior conjugado de uma verossimilhança gaussiana
  - a distribuição beta é o prior conjugado da verossimilhança binomial
  - a distribução gama é o prior conjugado da verossimilhança poissoniana

- em geral *o prior não deve dominar o resultado*
  se isso acontece é porque os dados são pouco informativos!

que prior escolher?

- em muitos casos usam-se *priores não-informativos*
- priores não-informativos para parâmetros de localização (ex.: média de uma gaussiana):
  - ex.: *prior uniforme* dentro de um intervalo de interesse:

$$P(w) = \begin{cases} \frac{1}{b-a} & \text{se } a < w < b \\ 0 & \text{se } w < a \text{ ou } w > b \end{cases}$$

  - nesse caso, o valor mais provável do posterior (MAP, *maximum a posteriori*) com prior uniforme é igual à solução de máxima verossimilhança

- priores não-informativos para parâmetros de escala (ex.: dispersão de uma gaussiana):

$$P(w) = \begin{cases} \frac{1}{w} & \text{se } a < w < b \\ 0 & \text{se } w < a \text{ ou } w > b \end{cases}$$

prior de Jeffreys

- os parâmetros que aparecem nos priores (como a e b acima) são chamados *hiperparâmetros*, para distinguir dos parâmetros do modelo
- priores podem ser próprios (normalizados) ou impróprios (não normalizados)

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

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

- vamos supor para μ um prior <u>uniforme</u>:

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

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

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

ou

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

- 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 1: 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
- $D = \{x_i, \sigma\}$
- prior de μ uniforme
- verossimilhança de um dado: $P(x_i, \sigma|\mu) \sim N(\mu, \sigma)$
- verossimilhança da amostra (supondo dados independentes):

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

- 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 $P(\mu|D) \propto P(D|\mu) \propto \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] \sim N(\bar{x}, \sigma/\sqrt{n})$

exemplo 2: 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: $\quad P(\mu|x_1, \sigma) \propto P(x_1|\mu, \sigma)P(\mu) \propto \exp\left[-\frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau_0^2}\right] \propto \exp\left[-\frac{(\mu-\mu_1)^2}{2\tau_1^2}\right] = N(\mu_1, \tau_1)$
- o posterior também é gaussiano com

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

note que se $\tau_0 >> \sigma$, a influência do prior é desprezível
se $\tau_0 << \sigma$, o prior é muito *informativo*

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

- agora temos n observações: $D = \{x_i, \sigma\}$
- 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 $\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$, $\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$ e $\bar{x} = \frac{1}{n}\sum_i^n x_i$
- para n grande o efeito do prior fica desprezível!
- *o prior não deve dominar o resultado, a menos que seja relevante!*

exemplo 3: análise sequencial ou conjunta?

- podemos estimar w via uma análise conjunta de todos os dados $D = \{x_1, x_2, ...\}$:

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

- ou sequencialmente, usando um dado de cada vez:
  - chega um dado x_1:

  $$P(w|x_1) \propto P(x_1|w)P(w)$$

  - chega um dado x_2:

  $$P(w|x_1, x_2) \propto P(x_1, x_2|w)P(w)$$

  se os dados são independentes: $P(x_1, x_2|w) = P(x_1|w)P(x_2|w)$

- logo,

$$P(w|x_1, x_2) \propto P(x_1|w)P(x_2|w)P(w)$$

$$\propto P(x_2|w)P(w|x_1)$$

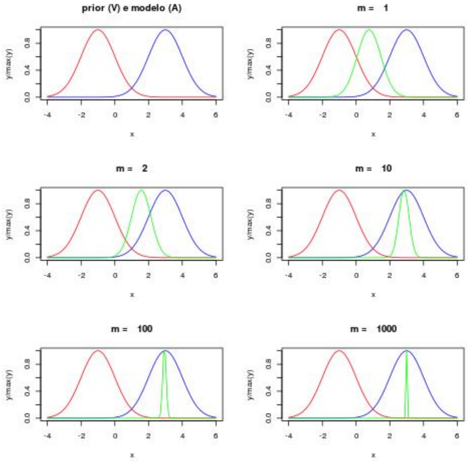
- se chega um dado x_3, analogamente, teremos:

$$P(w|x_1, x_2, x_3) \propto P(x_3|w)P(w|x_1, x_2)$$

- note que, no caso sequencial, *o prior dos parâmetros é o posterior da estimativa de w com os dados anteriores*

exemplo 3: análise sequencial ou conjunta?

- simulação:
  - dados: $N(x; \mu = 3, \sigma = 1)$
  - prior: $N(x; \mu = -1, \sigma = 1)$
  - figuras: como o posterior depende do número de pontos (m)
- vejam como a moda do posterior migra rapidamente de próximo da moda do prior para a dos dados simulados com o aumento do número de dados!
- *conforme analisamos as evidências, mudamos nossas mentes; conforme conseguimos mais informações, tomamos decisões mais racionais frente às evidências*



exemplo 4: o problema do farol

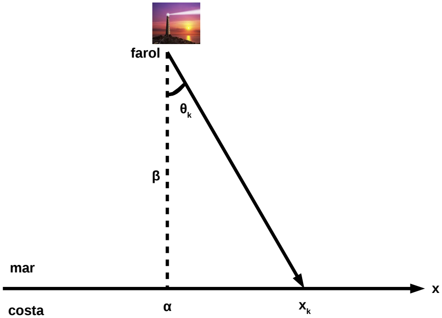
exemplo de uma modelagem bayesiana
Sivia & Skilling, sec. 2.4 (Gull, 1988)

- um farol está em uma posição α ao longo de uma costa reta e a uma distância β no mar;
- girando, ele emite uma série de pulsos curtos altamente colimados em intervalos de tempo aleatórios (e portanto em azimutes θ também aleatórios);
- N destes pulsos são detectados por sensores na costa, mas só as posições $D = \{x_1, x_2..., x_N\}$, não as direções:

onde está o farol?

- objetivo: determinar α e β

$$P(\alpha, \beta|D) \propto P(D|\alpha, \beta)P(\alpha, \beta)$$



exemplo 4: o problema do farol

- da geometria do problema: $\text{tg}\theta = (x - \alpha)/\beta$
- vamos atribuir um prior uniforme para θ:
  - para o pulso ser detectado, ele deve ter sido emitido com θ em $-\frac{\pi}{2} \leq \theta_k \leq \frac{\pi}{2}$: $P(\theta_k) = \frac{1}{\pi}$
- mudança de variáveis: $|P(x)dx| = |P(\theta)d\theta|$
- derivando em relação a x:

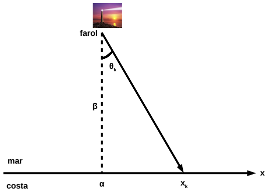
$$\frac{d\,\text{tg}\theta}{d\theta} = \sec^2\theta = \frac{1}{\beta}\frac{dx}{d\theta},$$

logo,

$$\frac{d\theta}{dx} = \frac{1}{\beta\sec^2\theta} = \frac{1}{\beta(1+\text{ tg}^2\theta)} = \frac{\beta}{\beta^2 + (x-\alpha)^2}$$

- como $P(x|\alpha,\beta) = P(\theta|\alpha,\beta)\left|\frac{d\theta}{dx}\right|$ vem que $P(x|\alpha,\beta) = \frac{\beta}{\pi[\beta^2+(x-\alpha)^2]}$,

que é uma distribuição de Cauchy!

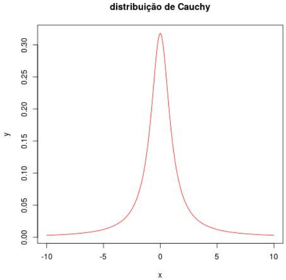


exemplo 4: o problema do farol

- distribuição de Cauchy

$$P(x|\alpha,\beta) = \frac{\beta}{\pi[\beta^2 + (x-\alpha)^2]}$$

distribuição simétrica em relação a α e com $FWHM = 2\beta$ (FWHM: largura a meia altura)



- verossimilhança:

$$P(D|\alpha,\beta) = \prod_{k=1}^{N} P(x_k|\alpha,\beta)$$

e, portanto,

$$\ln P(D|\alpha,\beta) =$$

$$= \mathrm{cte} \; + N \ln\beta - \sum_{k=1}^{N} \ln[\beta^2 + (x_k - \alpha)^2]$$

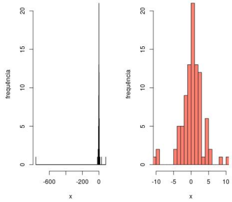
- vamos adotar priores não informativos para α e β: nesse caso o posterior é proporcional à verossimilhança

exemplo 4: o problema do farol

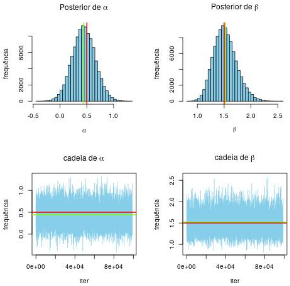
- log do posterior:

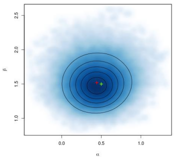
$$\ln P(\alpha,\beta|D) = \text{cte} + N\ln\beta - \sum_{k=1}^{N} \ln[\beta^2 + (x_k - \alpha)^2]$$

- simulação com $N = 100$, $\alpha = 0.5$, $\beta = 1.5$



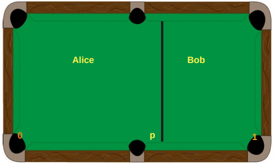
- inferência com MCMC:
$\alpha = 0.44 \pm 0.21$, $\beta = 1.52 \pm 0.22$





quando estimativas bayesianas e frequentistas divergem?

- métodos bayesianos lidam com distribuições de probabilidades, métodos frequentistas lidam com estimativas de ponto
- 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"



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

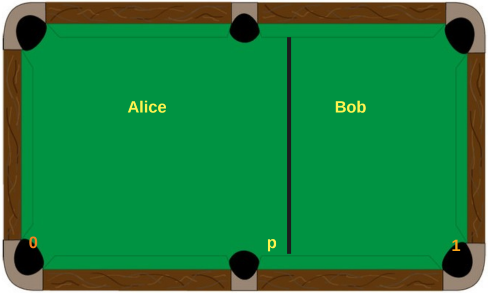
  ver http://jakevdp.github.io/blog/2014/06/06/frequentism-and-bayesianism-2-when-results-differ/

- 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*

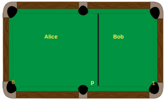
- “sumimos” com p via marginalização:

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

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

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

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



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,

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

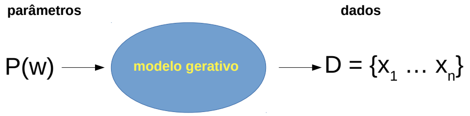
- a função beta é:
  $\beta(n,m) = \int_0^1 (1-p)^{n-1} p^{m-1} dp$
- logo, $P(B|D) = \frac{\beta(6+1,5+1)}{\beta(3+1,5+1)} \simeq 0.091$
- compare com MV: $P_{freq} = 0.053$!

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

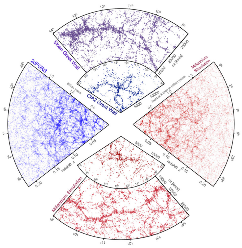
modelos gerativos

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



ABC: Approximate Bayesian Computation

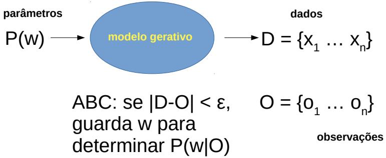
- objetivo: estimar o posterior dos parâmetros de um modelo
  - a) sem calcular (necessariamente) a verossimilhança
  - b) mas usando um processo gerativo de gerar dados a partir de parâmetros



- porquê?
  -há situações onde a verossimilhança é intratável
  -exemplo: estimativa de parâmetros cosmológicos comparando observações e simulações da distribuição de galáxias
- como proceder?
  ao invés de se comparar os dados (e.g. posições de galáxias) diretamente, 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

- vamos estimar o posterior dos parâmetros w de um modelo, $P(w|D)$, usando um processo gerativo a partir dos priores $P(w)$
- ABC: apenas dados simulados que concordam com as observações dentro de uma certa "tolerância" são considerados para amostrar o posterior



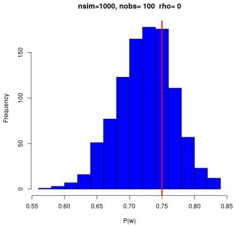
- algoritmo ABC:
  - 1. amostre w_p do prior $P(w)$
  - 2. simule dados D_p com w_p
  - 3. calcule as estatísticas que sumarizam os dados: $x_p =$resumo(w_p)
  - 4. aceite w_p se $|x_p - x_{obs}| < \epsilon$ (tolerância)
  - 5. retorne a 1
- atenção: em geral (mas nem sempre),
  - dados contínuos: tolerância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

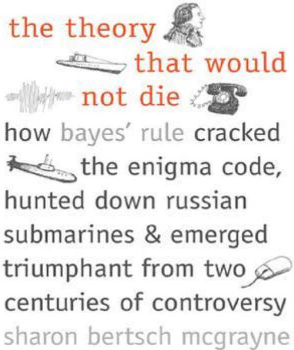
$$\rho = |X - Y|$$

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



sugestões

- *The theory that would not die*,
por Sharon Bertsch Mcgrayne
um livro muito interessante, com a história e grandes contribuições (desconhecidas) dos métodos bayesianos



- visite www.sumsar.net
recomendo fortemente os 3 vídeos de Rasmus Bååth sobre introdução à análise de dados bayesiana