Laboratório Nacional de Computação Científica
UNIDADE DE PESQUISA DO MCTI

Mais sobre Inferência Bayesiana e Exemplos

Modelos Compartimentais em Epidemiologia e Inferência Bayesiana

Gustavo Libotte e Regina Almeida

12 de janeiro de 2022

Distribuição de probabilidade marginal
Motivação

- Antes de vermos mais sobre inferência Bayesiana, vamos ver duas definições importantes.
- **Exemplo**: duas linhas de produção fabricam um certo tipo de peça.
- Suponha que a capacidade (em qualquer dia) seja 5 peças na linha I e 3 peças na linha II.
- Admita que o número de peças realmente produzidas em qualquer linha seja uma variável aleatória, e que (X, Y) represente a variável aleatória bidimensional que fornece o número de peças produzidas pela linha I e pela linha II, respectivamente.
- A tabela a seguir fornece a distribuição de probabilidade conjunta de (X, Y).
- Cada casa representa
$$p\left(x_i, y_j\right) = P\left(X = x_i, Y = y_j\right).$$
- Assim, $p(2, 3) = P(X = 2, Y = 3) = 0{,}04$ etc.

Distribuição de probabilidade marginal
Motivação

- Vamos calcular os totais "marginais", isto é, a soma das 6 colunas e 4 linhas da tabela.
- As probabilidades que aparecem nas margens, linha e coluna, representam a distribuição de probabilidade de Y e de X, respectivamente.
- Por exemplo, $P(Y = 1) = 0,26, P(X = 3) = 0,21$ etc.

Y \ X	0	1	2	3	4	5	Soma
0	0	0,01	0,03	0,05	0,07	0,09	0,25
1	0,01	0,02	0,04	0,05	0,06	0,08	0,26
2	0,01	0,03	0,05	0,05	0,05	0,06	0,25
3	0,01	0,02	0,04	0,06	0,06	0,05	0,24
Soma	0,03	0,08	0,16	0,21	0,24	0,28	1,00

Distribuição de probabilidade marginal
Definição

- No caso **discreto**, procederemos assim: desde que $X = x_i$ deve ocorrer junto com $Y = y_j$ para algum j e pode ocorrer com $Y = y_j$ somente para um j, teremos

$$p\left(x_i\right) = P\left(X = x_i\right) = P\left(X = x_i, Y = y_1 \text{ ou } X = x_i, Y = y_2 \text{ ou } \cdots\right)$$
$$= \sum_{j=1}^{\infty} p\left(x_i, y_j\right)$$

- A função p definida para $x_1, x_2, \ldots,$ representa a distribuição de probabilidade marginal de X.
- Analogamente definimos $q\left(y_j\right) = P\left(Y = Y_j\right) = \sum_{i=1}^{\infty} p\left(x_i, y_j\right)$ como a distribuição de probabilidade marginal de Y.

Distribuição de probabilidade marginal
Definição

- No caso **continuo**, procederemos do seguinte modo: seja f a função densidade de probabilidade conjunta da variável aleatória bidimensional contínua (X, Y).
- Definiremos g e h, respectivamente as funções densidade de probabilidade marginal de X e de Y, assim:

$$g(x) = \int_{-\infty}^{+\infty} f(x,y)dy; \quad h(y) = \int_{-\infty}^{+\infty} f(x,y)dx.$$

- Essas funções densidade de probabilidade correspondem às funções densidade de probabilidade básicas das variáveis aleatórias unidimensionais X e Y, respectivamente.
- Por exemplo

$$\begin{aligned} P(c \leq X \leq d) &= P[c \leq X \leq d, -\infty < Y < \infty] \\ &= \int_c^d \int_{-\infty}^{+\infty} f(x,y)dydx \\ &= \int_c^d g(x)dx. \end{aligned}$$

Variáveis aleatórias independentes

- Suponha que $Y_1, \ldots, Y_n$ sejam variáveis aleatórias e que θ seja um parâmetro que descreve as condições sob as quais as variáveis aleatórias são geradas.
- Dizemos que $Y_1, \ldots, Y_n$ são **condicionalmente independentes** dado θ se para cada coleção de n conjuntos $\{A_1, \ldots, A_n\}$, temos que
$$\mathrm{P}\left(Y_1 \in A_1, \ldots, Y_n \in A_n \mid \theta\right) = \mathrm{P}\left(Y_1 \in A_1 \mid \theta\right) \times \cdots \times \mathrm{P}\left(Y_n \in A_n \mid \theta\right)$$
onde cada $\left\{Y_j \in A_j\right\}$ é um evento.
- A independência condicional pode ser interpretada como que Y_j não fornece **nenhuma informação adicional** sobre Y_i além de saber θ. Além disso, sob independência, a densidade da junta é dada por
$$p\left(y_1, \ldots, y_n \mid \theta\right) = p\left(y_1 \mid \theta\right) \times \cdots \times p\left(y_n \mid \theta\right) = \prod_{i=1}^{n} p\left(y_i \mid \theta\right)$$
o **produto das densidades marginais**.
- Neste caso, dizemos que $Y_1, \ldots, Y_n$ são **condicionalmente independentes e identicamente distribuídas** (i.i.d.).

Inferência Bayesiana

- Considere o (já conhecido) Teorema de Bayes:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

- Se substituirmos θ por *hipótese* e y por *dados*, o teorema de Bayes nos diz como calcular a probabilidade de uma hipótese θ, conhecendo os dados y.
- Mas, como transformamos uma hipótese em algo que podemos "colocar dentro" do teorema de Bayes? Fazemos isso usando **distribuições de probabilidade**.
- Portanto, em geral, nossa *hipótese* é uma hipótese em um sentido muito, muito, muito restrito; seremos mais precisos se falarmos em **encontrar um valor adequado para os parâmetros** de nossos modelos, ou seja, parâmetros de distribuições de probabilidade.
- Vamos falar de cada termo do teorema de Bayes:
  - $p(\theta)$: Distribuição *a priori*
  - $p(y \mid \theta)$: Função de verossimilhança
  - $p(y)$: Distribuição marginal
  - $p(\theta \mid y)$: Distribuição *a posteriori*

Inferência Bayesiana
Um exemplo analítico

LNCC

- Os participantes de uma pesquisa social foram questionados se eles eram ou não felizes. Seja Y_i a variável aleatória associada a esta questão, de modo que

$$Y_i = \begin{cases} 1, & \text{se o participante } i \text{ se declara feliz} \\ 0, & \text{caso contrário} \end{cases}$$

- Suponha que θ seja a taxa de felicidade entre os $N = 1.272$ respondentes da pesquisa.
- Deste conjunto, selecionamos $n = 129$ respostas ao acaso.
- Podemos afirmar que $Y_1, \ldots, Y_{129}$ são condicionalmente independentes e identicamente distribuídas (i.i.d.). Portanto

$$P\left(Y_i = y_i \mid \theta, Y_j = y_j, j \neq i\right) = \theta^{y_i}(1-\theta)^{1-y_i}$$

$$\begin{aligned} P\left(Y_1 = y_1, \ldots, Y_{129} = y_{129} \mid \theta\right) &= \prod_{i=1}^{129} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_i y_i}(1-\theta)^{129-\sum_i y_i} \end{aligned}$$

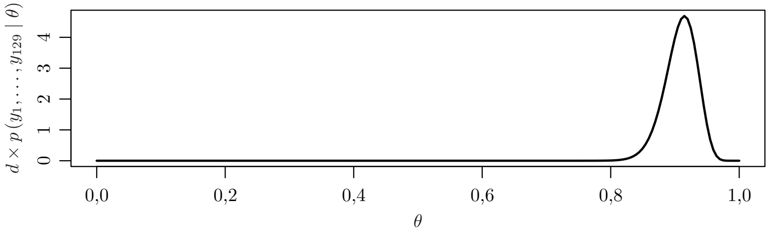
Inferência Bayesiana
Um exemplo analítico

- **Qual seria a nossa distribuição *a priori*?**
- O parâmetro θ é algum número desconhecido entre 0 e 1.
- Suponha que nossa informação *a priori* seja tal que todos os subintervalos de $[0,1]$ com o mesmo comprimento também tenham a mesma probabilidade. Isso significa que
$$P(a \leq \theta \leq b) = P(a+c \leq \theta \leq b+c) \quad \text{para} \quad 0 \leq a < b < b+c \leq 1.$$
- Esta condição implica que θ deve ser uniformemente distribuído:
$$p(\theta) = 1 \quad \text{para todo} \quad \theta \in [0,1].$$
- Para esta distribuição *a priori* e o modelo de amostragem definido, o teorema de Bayes resulta em
$$\begin{aligned} p\left(\theta \mid y_1, \ldots, y_{129}\right) &= \frac{p\left(y_1, \ldots, y_{129} \mid \theta\right) p(\theta)}{p\left(y_1, \ldots, y_{129}\right)} \\ &= p\left(y_1, \ldots, y_{129} \mid \theta\right) \times \frac{1}{p\left(y_1, \ldots, y_{129}\right)} \\ &\propto p\left(y_1, \ldots, y_{129} \mid \theta\right). \end{aligned}$$

Inferência Bayesiana
Um exemplo analítico

- Agora, vamos **conhecer os dados**:
  - 129 indivíduos entrevistados;
  - 118 indivíduos responderam serem felizes, em geral (91%)
  - 11 indivíduos responderam não serem felizes, em geral (9%)
- Da nossa função de verossimilhança, temos que

$$\begin{aligned} p\left(y_1, \ldots, y_{129} \mid \theta\right) &= \theta^{\sum_i y_i}(1-\theta)^{129-\sum_i y_i} \\ &= \theta^{118}(1-\theta)^{11} \end{aligned}$$



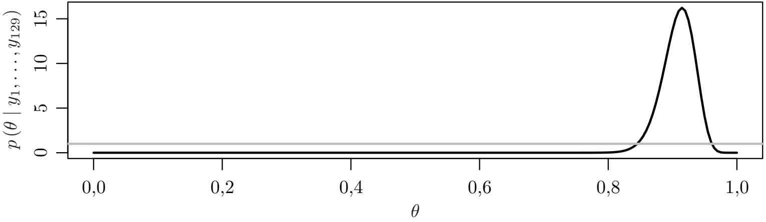
Por que o gráfico é de $d \times p\left(y_1, \ldots, y_{129} \mid \theta\right)$ e não apenas de $p\left(y_1, \ldots, y_{129} \mid \theta\right)$?

Inferência Bayesiana

Um exemplo analítico

- Neste resultado, vemos que a distribuição *a posteriori* $p\left(\theta \mid y_1, \ldots, y_{129}\right)$ terá **a mesma forma** que esta função.
- Portanto, sabemos que o valor verdadeiro de θ é muito provável próximo a 0,91.
- No entanto, muitas vezes queremos ser mais precisos do que isso e precisaremos saber **a escala** de $p\left(\theta \mid y_1, \ldots, y_{129}\right)$, bem como a forma.
- Pela regra de Bayes, temos

$$\begin{aligned} p\left(\theta \mid y_1, \ldots, y_{129}\right) &= \theta^{118}(1-\theta)^{11} \times p(\theta)/p\left(y_1, \ldots, y_{129}\right) \\ &= \theta^{118}(1-\theta)^{11} \times 1/p\left(y_1, \ldots, y_{129}\right) \end{aligned}$$



Inferência Bayesiana
Um exemplo analítico

- Podemos calcular a escala ou "constante de normalização" $1/p\left(y_1,\ldots,y_{129}\right)$ usando o seguinte resultado do cálculo [1, p. 18–19]:

$$p\left(y_1,\ldots,y_{129}\right) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta$$
$$= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

- Como o resultado do cálculo nos ajuda a calcular $p\left(\theta \mid y_1,\ldots,y_{129}\right)$?

  1. $\int_0^1 p\left(\theta \mid y_1,\ldots,y_{129}\right) d\theta = 1$, uma vez que todas as distribuições de probabilidade somam para 1;
  2. $p\left(\theta \mid y_1,\ldots,y_{129}\right) = \theta^{118}(1-\theta)^{11} / p\left(y_1,\ldots,y_{129}\right)$, do teorema de Bayes.

[1] E. Artin.
The Gamma Function.
Athena Series, Selected Topics in Mathematics. Holt, Rinehart and Winston, New York, 1 edition, 1964.

Inferência Bayesiana
Um exemplo analítico

- Usando estes resultados, temos:

$$
\begin{aligned}
1 &= \int_0^1 p\left(\theta \mid y_1, \ldots, y_{129}\right) d\theta, \quad \text{usando (1)} \\
1 &= \int_0^1 \theta^{118}(1-\theta)^{11} / p\left(y_1, \ldots, y_{129}\right) d\theta, \quad \text{usando (2)} \\
1 &= \frac{1}{p\left(y_1, \ldots, y_{129}\right)} \int_0^1 \theta^{118}(1-\theta)^{11} d\theta \\
1 &= \frac{1}{p\left(y_1, \ldots, y_{129}\right)} \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}
\end{aligned}
$$

- Assim, concluímos que:

$$
p\left(y_1, \ldots, y_{129}\right) = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}
$$

- Juntando todos os resultados, chegamos a

$$
p\left(\theta \mid y_1, \ldots, y_{129}\right) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)} \theta^{118}(1-\theta)^{11}
$$

Inferência Bayesiana
Um exemplo numérico

- **Problema**: jogamos uma moeda várias vezes e registramos quantas caras e coroas obtemos.
- Com base nesses dados, tentamos responder a perguntas como: *a moeda é viciada?* Ou, de forma mais geral, *quão enviesada é a moeda?*
- Vamos supor que já jogamos uma moeda várias vezes e temos um registro do número de **caras** observadas, então a parte de coleta de dados já está feita.
- Diremos que uma moeda com viés de 1 sempre dará **cara**, uma com viés 0 sempre dará **coroa** e outra com viés de 0,5 dará cara e coroa na mesma proporção.
- Para representar a **tendência**, usaremos o parâmetro θ e, para representar o **número total de caras** para um número N de **lançamentos**, usaremos a variável y.

Inferência Bayesiana
Um exemplo numérico

- Vamos supor que apenas **dois resultados** sejam possíveis—cara ou coroa—e também supor que um lançamento de moeda não afeta os outros lançamentos, ou seja, estamos assumindo que os lançamentos de moeda são **independentes** um do outro.
- Iremos ainda assumir que todos os lançamentos de moeda vêm da mesma distribuição. Assim, o lançamento da moeda é uma variável aleatória **i.i.d**.
- Dadas essas premissas, um bom candidato para a probabilidade é a **distribuição binomial**:

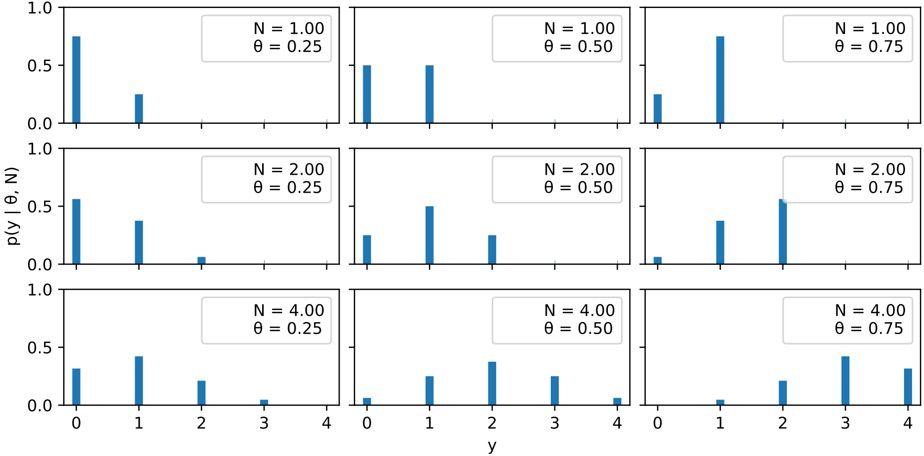
$$p(y \mid \theta, N) = \frac{N!}{y!(N-y)!}\theta^y(1-\theta)^{N-y}$$

- Relembrando: esta é uma distribuição **discreta** que retorna a probabilidade de se obter y caras (ou, em geral, sucessos) de N jogadas (ou, em geral, tentativas ou experimentos) dado um valor θ fixo.

Inferência Bayesiana
Um exemplo numérico

```python
n_params = [1, 2, 4] # Numero de tentativas
p_params = [0.25, 0.5, 0.75] # Probabilidade de sucesso
x = np.arange(0, max(n_params)+1)
f,ax = plt.subplots(len(n_params), len(p_params), sharex=True, sharey=True, figsize=(8, 4), constrained_layout=True)
for i in range(len(n_params)):
    for j in range(len(p_params)):
        n = n_params[i]
        p = p_params[j]
        y = stats.binom(n=n, p=p).pmf(x)
        ax[i,j].vlines(x, 0, y, colors='C0', lw=5)
        ax[i,j].set_ylim(0, 1)
        ax[i,j].plot(0, 0, label="N = {:3.2f}\nθ = {:3.2f}".format(n,p), alpha=0)
        ax[i,j].legend()
        ax[2,1].set_xlabel('y')
        ax[1,0].set_ylabel('p(y | θ, N)')
        ax[0,0].set_xticks(x)
```

Inferência Bayesiana
Um exemplo numérico



Inferência Bayesiana
Um exemplo numérico

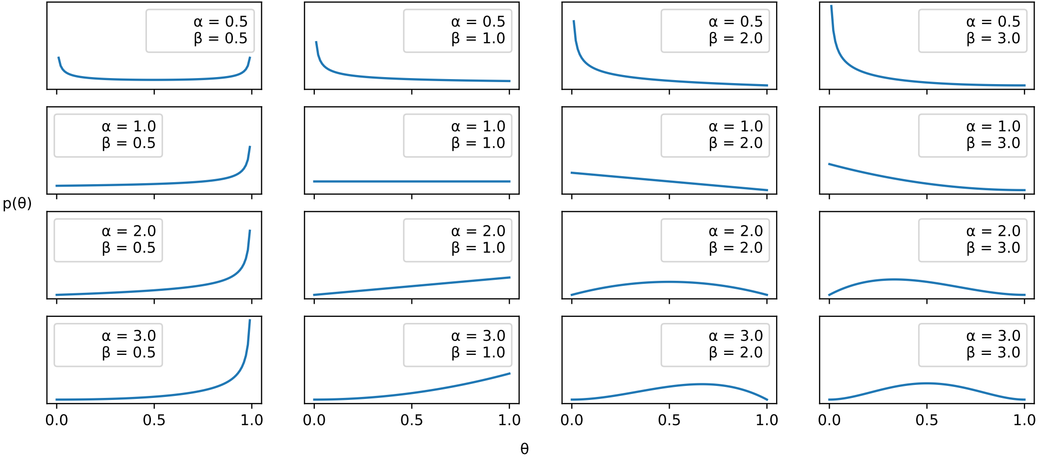
- Se somarmos a altura de todas as barras, obteremos 1, ou seja, para distribuições discretas, a altura das barras representa as probabilidades reais.
- A **distribuição binomial** é uma escolha razoável para a **função de verossimilhança**: podemos ver que θ indica a probabilidade de se obter uma cara ao jogar uma moeda—basta comparar o valor de θ com a altura das barras para $y = 1$ (caras).
- Se soubermos o valor de θ, a distribuição binomial nos dirá a **distribuição esperada de caras**. O único problema é que **não conhecemos** θ!
- Para a distribuição *a priori*, usaremos uma **distribuição beta**:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Inferência Bayesiana
Um exemplo numérico

```
params = [0.5, 1, 2, 3]
x = np.linspace(0, 1, 100)
f, ax = plt.subplots(len(params), len(params), sharex=True, sharey=True, figsize=(12, 5))
for i in range(4):
    for j in range(4):
        a = params[i]
        b = params[j]
        y = stats.beta(a, b).pdf(x)
        ax[i,j].plot(x, y)
        ax[i,j].plot(0, 0, label="α = {:2.1f}\nβ = {:2.1f}".format(a,b), alpha=0)
        ax[i,j].legend()
ax[1,0].set_yticks([])
ax[1,0].set_xticks([0, 0.5, 1])
f.text(0.5, 0.03, 'θ', ha='center')
f.text(0.09, 0.5, 'p(θ)', va='center', rotation=0)
```

Inferência Bayesiana
Um exemplo numérico



Inferência Bayesiana
Um exemplo numérico

LNCC

- Existem algumas razões para usar uma distribuição beta para este problema:
  1. A distribuição beta está **restrita** entre 0 e 1, da mesma forma que nosso parâmetro θ.
  2. Outro motivo é sua **versatilidade**. Como podemos ver na figura anterior, a distribuição adota várias formas, incluindo uma distribuição uniforme, distribuições tipo gaussiana e distribuições tipo U.
  3. A distribuição beta é uma **distribuição *a priori* conjugada**[1] à distribuição binomial (que estamos usando como função de verossimilhança). Cada vez que usarmos uma distribuição beta como a distribuição *a priori* e uma distribuição binomial como a função de verossimilhança, teremos uma distribuição beta como a distribuição *a posteriori*.
- A distribuição *a posteriori* é **proporcional** à função de verossimilhança vezes a distribuição *a priori*. Então, para o nosso problema, temos que multiplicar as distribuições binomial e beta:

$$p(\theta \mid y) \propto \frac{N!}{y!(N-y)!}\theta^{y}(1-\theta)^{N-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

[1] Uma **distribuição *a priori* conjugada** é tal que, quando combinada com uma dada função de verossimilhança, leva a uma distribuição *a posteriori* com a mesma forma funcional da distribuição *a priori* (da mesma família). Um distribuição *a priori* conjugada é uma conveniência algébrica, pois resulta em uma expressão de forma fechada para a distribuição *a posteriori*.

Inferência Bayesiana
Um exemplo numérico

- Podemos simplificar essa expressão. Para nossas questões práticas, podemos descartar todos os termos que não dependem de θ e nossos resultados ainda serão válidos. Assim, podemos escrever:

$$p(\theta \mid y) \propto \theta^{y}(1-\theta)^{N-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Reordenando, obtemos:

$$p(\theta \mid y) \propto \theta^{y+\alpha-1}(1-\theta)^{N-y+\beta-1}$$

- Se prestarmos atenção, veremos que esta expressão tem a mesma forma funcional de uma **distribuição beta** (exceto para o termo de normalização) com $\alpha_{\text{posterior}} = \alpha_{\text{prior}} + y$ e $\beta_{\text{posterior}} = \beta_{\text{prior}} + N - y$. Na verdade, a distribuição *a posteriori* do nosso problema é a distribuição beta:

$$p(\theta \mid y) \propto \text{Beta}\left(\alpha_{\text{prior}} + y, \beta_{\text{prior}} + N - y\right)$$

Inferência Bayesiana
Um exemplo numérico

```python
plt.figure(figsize=(12, 5))
n_trials = [0, 2, 3, 4, 8, 16, 32, 50, 150]
data = [0, 1, 1, 1, 4, 6, 9, 13, 48]
theta_real = 0.35
beta_params = [(1, 1), (20, 20), (1, 4)]
dist = stats.beta
x = np.linspace(0, 1, 200)
for idx, N in enumerate(n_trials):
    plt.subplot(3, 3, idx+1)
    plt.xlabel('θ')
    y = data[idx]
    for (a_prior, b_prior) in beta_params:
        p_theta_given_y = dist.pdf(x, a_prior + y, b_prior + N - y)
        plt.fill_between(x, 0, p_theta_given_y, alpha=0.7)
    plt.axvline(theta_real, ymax=0.3, color='k')
    plt.plot(0, 0, label=f'{N:4d} jogadas\n{y:4d} caras', alpha=0)
    plt.xlim(0, 1); plt.ylim(0, 12); plt.legend(); plt.yticks([])
plt.tight_layout()
```

Inferência Bayesiana
Um exemplo numérico