

**Estatística para Ciências de Dados**

# **Aula 8: Regressão padronizada e Inferência bayesiana**

Mariana Cúri  
ICMC/USP  
[mcuri@icmc.usp.br](mailto:mcuri@icmc.usp.br)



# Conteúdo

## 1. Regressão linear múltipla: tópico adicional

- Problemas detectados na análise de resíduos
- Controle de erros computacionais: regressão padronizada

## 2. Inferência bayesiana

- Provocação
- Filosofia
- Distribuições a priori e a posteriori
- Prioris conjugadas

## 3. Estimação bayesiana

- Estimador de Bayes
- Intervalo de credibilidade
- Distribuição preditiva

# Regressão linear múltipla: análise de resíduos

Problemas detectados em um modelos de regressão linear múltipla:

- Relação não linear entre variável resposta e preditores
  - transformação das variáveis (resposta e/ou preditoras)
  - incluir novos preditores no modelo
  - método de transformação Box-Cox para a variável resposta
- Heterocedasticidade (variância não é constante)
  - transformação da variável resposta
  - se a variação está relacionada a um preditor: Mínimos Quadrados Ponderados
- Resposta não gaussiana
  - transformação da variável resposta
  - Modelos Lineares Generalizados
- Outliers e pontos influentes
  - modelos robustos
- Erros computacionais: cálculo da inversa de  $X^tX$  é a principal fonte

# Regressão linear múltipla: análise de resíduos

- A inversão da matriz pode gerar graves erros de arredondamento principalmente quando:
  - as variáveis preditoras são correlacionadas em um alto grau:  $\det(\mathbf{X}^t\mathbf{X})$  perto de 0
  - a ordem de grandeza dos preditores é muito diferente
  - decomposição de Cholesky e decomposição por autovalores, reduzem os problemas de cálculo da inversa
  - padronização dos preditores (transformação da correlação) também ajuda: elementos de  $(\mathbf{X}^*)^t\mathbf{X}^*$  entre -1 e 1 (base da Regressão Padronizada)
  - Regressão Lasso ou Ridge

# Regressão linear múltipla: análise de resíduos

- Multicolinearidade

- não gera falta de ajuste do modelo e tende a não afetar inferências sobre a resposta média
- infla o erro padrão dos estimadores dos parâmetros de regressão
- pode distorcer o sentido da relação do preditor e da resposta
- $VIF = (1 - R_k^2)^{-1}$ , em que  $R_k^2$  é o coeficiente de determinação do modelo de regressão de  $X_k$  em função dos demais preditores do modelo; note que se  $R_k^2 = 0$ , então  $VIF = 1$
- $\max(VIF)$  dos preditores do modelo acima de 10: multicolinearidade importante
- cortes para  $(1/VIF)$  sugeridos: 0,01 ou 0,001 ou 0,0001 (preditor deve ser excluído)
- limitações do VIF: incapaz de distinguir várias multicolinearidades simultâneas

# Regressão padronizada

- Transformação das variáveis resposta e explicativas (quantitativas):

$$X^* = \frac{X_i - \bar{X}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

incluir o denominador (n-1)  
faz com que os elementos  
de  $(X^*)^t X^*$  não estejam  
entre -1 e 1

$$Y^* = \frac{Y_i - \bar{Y}}{\sqrt{\sum (Y_i - \bar{Y})^2}}$$

- Como as médias das variáveis transformadas são iguais a zero, a regressão padronizada passa pela origem  $(0, \mathbf{0}^t)$ , o modelo não tem intercepto


$$Y_i^* = \beta_1^* X_1^* + \beta_2^* X_2^* + \cdots + \beta_{p-1}^* X_{p-1}^* + \epsilon_i^*$$

- $\beta_i = \beta_i^* \left( \frac{s_Y}{s_{X_i}} \right)$  o quanto o aumento de 1 desvio padrão em  $X_i$  impacta em  $Y$  (em unidades de desvios padrão)



# Inferência bayesiana: provocação

## Sally Clark, advogada, 1964-2007

- 1º filho morreu com 11 semanas, em 1996, e o 2º com 8 semanas, em 1998
- foi condenada pelo assassinato dos dois filhos em 1999-2000
- culpada até conseguir provar inocência
- testemunho do pediatra Roy Meadow: *'uma morte súbita na infância é uma tragédia para a família, duas são suspeitas e três são assassinio a menos que existam provas em contrário'*
- evidência estatística falha: 'a probabilidade de dois lactentes de uma família abastada vir a óbito por morte súbita (SMSL) é 1 em 73 milhões  $\approx (1/8500)^2$ '  
 **independentes?**  
**sem comparar com uma referência (assassinados pela mãe)?**
- Libertada após 3 anos: evidências de infecção no líquido de Harry (2º filho)

# Inferência bayesiana: provocação



## Beyond reasonable doubt

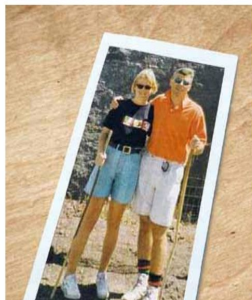
By Helen Joyce

Submitted by plusadmin on September 1, 2002

Five years ago, a young couple from Cheshire suffered one of the most devastating losses imaginable - their baby Christopher died in his sleep, aged 11 weeks. Doctors, neighbours, all were sympathetic, and the death was certified as natural causes - there was evidence of a respiratory infection, and no sign of any failure of care.

But just a year later, in what must have felt like a horribly familiar nightmare, the Clarks' second child Harry died, aged 8 weeks. This time, there was no sympathy from the professionals. Four weeks after Harry's death the couple were arrested, and eventually Sally Clark was charged with murdering both children. She was tried and convicted in 1999 and is now almost three years into a life sentence.

The forensic evidence was slim to nonexistent - certainly neither case would have stood up alone. Even the prosecution team disagreed among themselves as to how the two children had died. They claimed



## Inferência Bayesiana:

†† : os 2 morreram inesperadamente e sem causa aparente

MS: 2 mortes por SMSL (síndrome de morte súbita em lactentes)

MS<sup>c</sup>: não morrer por SMSL

baseado em: Confidential Enquiry for Stillbirths and Deaths in Infancy (CESDI), 1993-1996

Estabelecendo que ambos morreram da mesma causa:





# Inferência bayesiana: provocação

$$P(MS) = \frac{1}{1300} \frac{1}{100} = 0,0000077$$

$$P(MS^c) = 0,99999223$$

Falácia do promotor: = P(inocência de Sally)

distribuição de probabilidade a priori

$$P(\dagger\dagger | MS) = 1$$

$$P(\dagger\dagger | MS^c) = \frac{30}{650000} \frac{1}{10} = 0,0000046$$

distribuição de probabilidade condicionada ao fato de que as crianças morreram inesperadamente

$$P(MS^c | \dagger\dagger) \approx 0,37$$

$$P(MS | \dagger\dagger) = \frac{P(\dagger\dagger|MS)P(MS)}{P(\dagger\dagger)} = \frac{P(\dagger\dagger|MS)P(MS)}{P(\dagger\dagger|MS)P(MS)+P(\dagger\dagger|MS^c)P(MS^c)} \approx 0,63$$

# Inferência bayesiana: Filosofia

Mesmo contexto que inferência frequentista:

- modelo probabilístico  $f(\mathbf{x}|\theta)$
- deseja-se fazer inferência sobre  $\theta$
- diferença:  $\theta$  é tratado como uma quantidade aleatória
- inferência é baseada em  $f(\theta|\mathbf{x})$  ao invés de  $f(\mathbf{x}|\theta)$
- **distribuição a priori**  $f(\theta)$ : informação sobre a distribuição de  $\theta$ , antes de coletadas as observações
- é a **principal vantagem** ou **maior armadilha**
- prioris diferentes levam a inferências diferentes sobre  $\theta$

# Inferência bayesiana: *priori* e *posteriori*

Passos da abordagem bayesiana:

1. Escolha do modelo probabilístico:  $L(\theta|\mathbf{x}) \equiv f(\mathbf{x}|\theta)$
2. Escolha da distribuição a *priori*  $f(\theta)$
3. Obtenção da distribuição a *posteriori*,  $f(\theta|\mathbf{x})$ , pelo Teorema de Bayes
4. Inferências sobre  $\theta$  a partir da distribuição a *posteriori*

## Teorema de Bayes (variáveis aleatórias)

posteriori

$$f(\theta | x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

verossimilhança

marginal ou  
evidência ou  
constante  
normalizadora

priori

se  $\theta$  é discreto,  
substitui-se  
 $\int$  por  $\sum$  no  
denominador

# Inferência bayesiana: *priori* e *posteriori*

- $f(x)$ , no denominador, pode ser de difícil obtenção, devido à  $\int$
- obtenção de  $f(\theta|x)$  pode ser computacionalmente difícil
- para certas combinações verossimilhança-priori, o cálculo de  $f(x)$  pode ser evitado: **prioris conjugadas**

$$f(\theta \mid x) \propto f(x \mid \theta) f(\theta)$$



**C**

constante normalizadora

# Inferência bayesiana: priori conjugada

Voltemos ao exemplo do Sildenafil:

X: pelo menos 60% das tentativas bem sucedidas

$X \sim \text{Bernoulli}(p)$ ,  $i = 1, \dots, n=379$ , independentes

p: proporção de pessoas (na população) com pelo menos 60% das tentativas bem sucedidas

**interesse: fazer inferências sobre p**

$$0 < p < 1$$

Encontrar a posteriori: verossimilhança  
priori  
marginal (?)

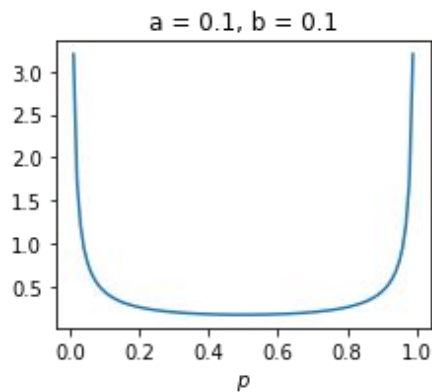


# Inferência bayesiana: priori conjugada

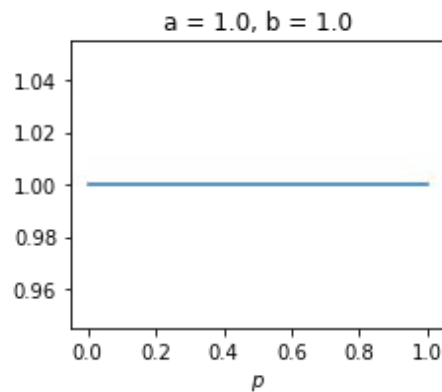
Verossimilhança:  $f(\mathbf{x} \mid p) = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}$

priori: o que já se sabe (ou se supõe) sobre o Sildenafil?

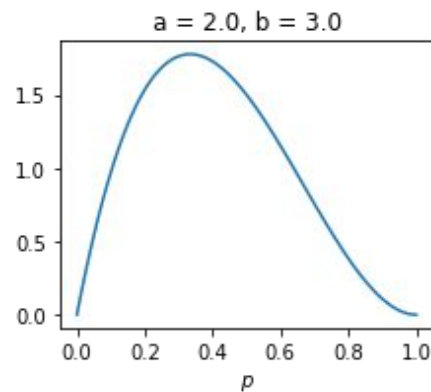
*métodos para encontrar  
priors não informativas,  
método de Jeffreys, por ex.*



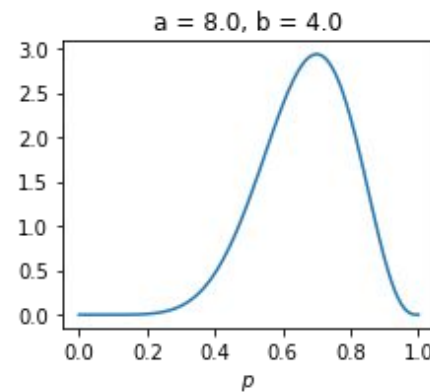
funciona muito bem ou  
quase não tem efeito



não se sabe nada



tem efeito em  
alguns casos



parece ser eficaz  
em vários casos

$$p \sim \text{Beta}(a, b) \quad f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}$$

$$0 < p < 1$$

# Inferência bayesiana: priori conjugada

obtenção da  
posteriori não  
exigiu  
integração,  
nem cálculos  
que não têm  
solução  
analítica;

$$f(p | \mathbf{x}) \propto p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i} p^{a-1} (1 - p)^{b-1}$$

$$\propto p^{\sum_{i=1}^n x_i + a - 1} (1 - p)^{n - \sum_{i=1}^n x_i + b - 1}$$

núcleo de uma  $Beta(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b)$

a maioria das  
situações exige  
obtenção por  
simulação:  
MCMC (Monte  
Carlo via  
Cadeia de  
Markov)

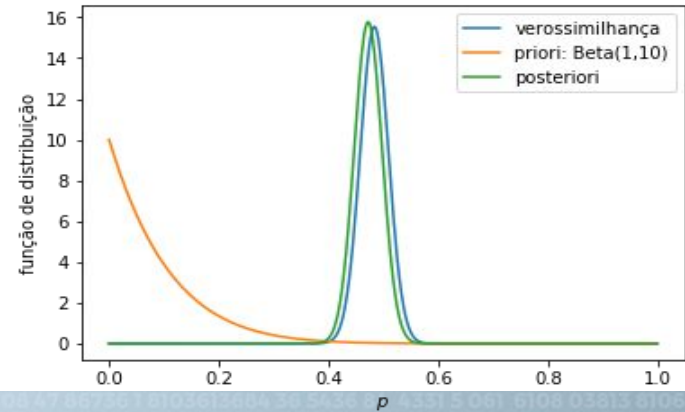
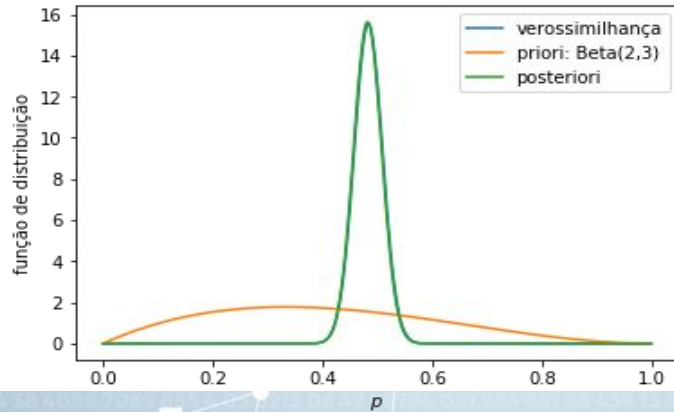
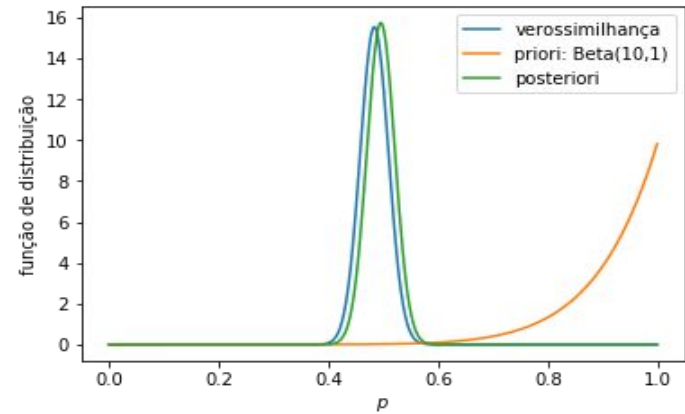
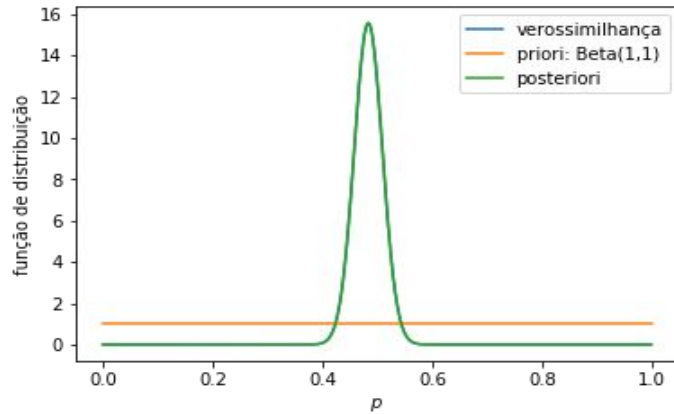
distribuição a posteriori

$$f(p | \mathbf{x}) = \frac{\Gamma(n+a+b)}{\Gamma(\sum_{i=1}^n x_i + a) \Gamma(n - \sum_{i=1}^n x_i + b)} p^{\sum_{i=1}^n x_i + a - 1} (1 - p)^{n - \sum_{i=1}^n x_i + b - 1} \quad 0 < p < 1$$

**PRIORIS CONJUGADAS:** a distribuição *a priori* e a *a posteriori* são da mesma família

*A Beta é priori conjugada para amostras da Binomial*

# Inferência bayesiana: priori Beta, ex Sildenafil



# Inferência bayesiana: prioris conjugadas

Verossimilhança	Priori	Posteriori
$Y \sim B(n, \theta)$	$\theta \sim \text{Beta}(p, q)$	$\theta y \sim \text{Beta}(p + y, q + n - y)$
$Y \sim P(\theta)$	$\theta \sim \text{Ga}(p, q)$	$\theta y \sim \text{Ga}(p + \sum_{i=1}^n y_i, q + n)$
$Y \sim N(\theta, \tau^{-1}), (\tau \text{ conhecido})$	$\theta \sim N(b, c^{-1})$	$\theta y \sim N(\frac{cb + n\tau\bar{y}}{c + n\tau}, \frac{1}{c + n\tau})$
$Y \sim \text{Ga}(k, \theta), (k \text{ conhecido})$	$\theta \sim \text{Ga}(p, q)$	$\theta y \sim \text{Ga}(p + nk, q + \sum_{i=1}^n y_i)$
$Y \sim \text{Geo}(\theta)$	$\theta \sim \text{Beta}(p, q)$	$\theta y \sim \text{Beta}(p + n, q + \sum_{i=1}^n y_i - n)$
$Y \sim \text{BN}(r, \theta)$	$\theta \sim \text{Beta}(p, q)$	$\theta y \sim \text{Beta}(p + r, q + y - r)$

Fonte: <http://www.leg.ufpr.br/~paulojus/CE227/InferenciaBayesiana.pdf>

# Aplicação real: voo AF 447





## Bayesian Search for Missing Aircraft

20 April 2017  
Lawrence D. Stone

**Bayesian search theory provides a principled and  
successful method for planning searches for lost  
aircraft and other objects**

Fonte:

[https://www.nps.edu/documents/103424533/106018074/Bayes+Search+for+Missing+Aircraft+NPS+20+Apr+2017.pdf/051a76bc-18cc-](https://www.nps.edu/documents/103424533/106018074/Bayes+Search+for+Missing+Aircraft+NPS+20+Apr+2017.pdf/051a76bc-18cc-47a7-b8b8-52d92d618dfe)

[47a7-b8b8-52d92d618dfe](https://www.nps.edu/documents/103424533/106018074/Bayes+Search+for+Missing+Aircraft+NPS+20+Apr+2017.pdf/051a76bc-18cc-47a7-b8b8-52d92d618dfe)

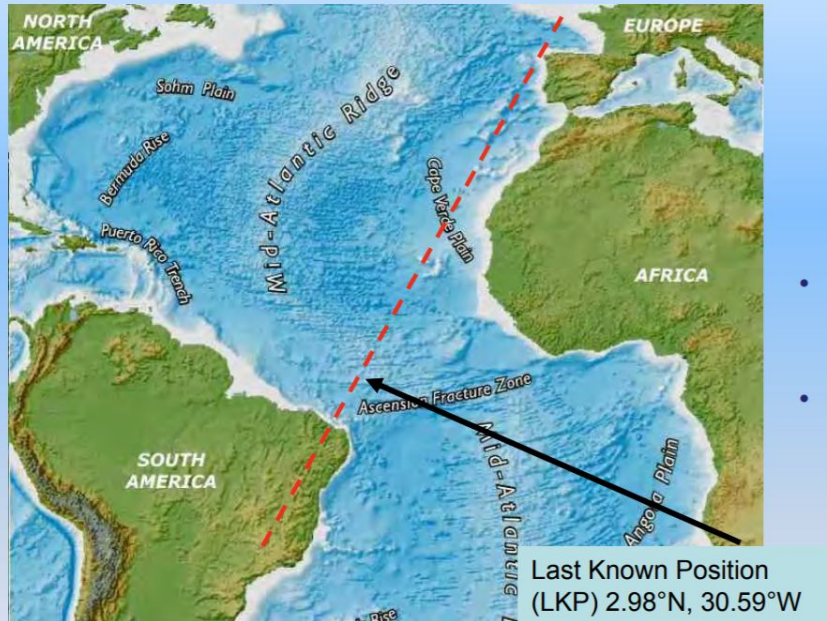
Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.





# Aplicação real: voo AF 447

## Air France Flight 447 Disappears



- In the early morning hours of June 1, 2009, Air France Flight AF 447 from Rio to Paris with 228 passengers and crew aboard, disappeared during stormy weather over the South Atlantic
- The French Bureau of Enquiries and Analyses (BEA) took charge of the search.
- On April 3, 2011, an autonomous underwater vehicle (AUV) found the wreckage on the ocean bottom at a depth of 13,060 ft

Fonte:

<https://www.nps.edu/documents/103424533/106018074/Bayes+Search+for+Missing+Aircraft+NPS+20+Apr+2017.pdf/051a76bc-18cc-47a7-b8b8-52d92d618dfe>

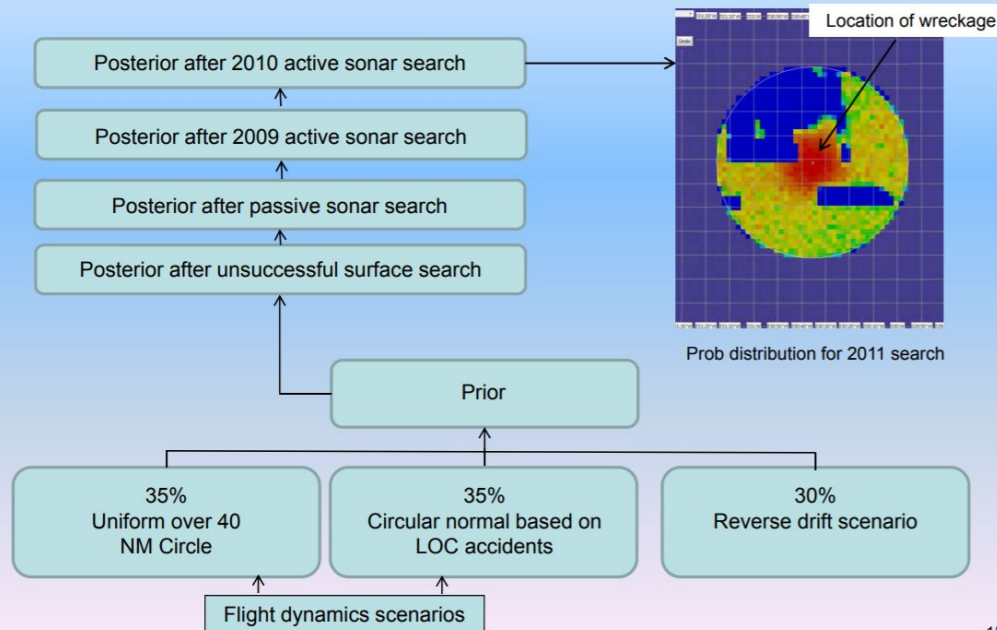
47a7-b8b8-52d92d618dfe

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.



# Aplicação real: voo AF 447

## Analysis Process



Fonte:

<https://www.nps.edu/documents/103424533/106018074/Bayes+Search+for+Missing+Aircraft+NPS+20+Apr+2017.pdf/051a76bc-18cc-47a7-b8b8-52d92d618dfe>

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

# Aplicação real: voo AF 447

## Table of Contents

<b>GLOSSARY.....</b>	<b>1</b>
<b>1 INTRODUCTION .....</b>	<b>2</b>
<b>2 APPROACH .....</b>	<b>3</b>
<b>3 PRIOR PROBABILITY DISTRIBUTION FOR IMPACT LOCATION.....</b>	<b>5</b>
3.1 FLIGHT DYNAMICS PRIOR.....	5
3.2 REVERSE DRIFT PRIOR.....	6
3.3 PRIOR PROBABILITY DISTRIBUTION BEFORE SURFACE SEARCH.....	9
<b>4 POSTERIOR DISTRIBUTION GIVEN UNSUCCESSFUL SEARCH.....</b>	<b>11</b>
4.1 ACCOUNTING FOR UNSUCCESSFUL SEARCH.....	11
4.2 AIRCRAFT, SHIP, AND SATELLITE SURFACE SEARCHES.....	12
4.3 PHASE I SEARCHES.....	18
4.4 PHASE II SEARCHES .....	24
4.5 PHASE III SEARCHES .....	27
4.6 POSTERIOR ASSUMING THE FINGERS FAILED .....	34
<b>5 CONCLUSIONS AND RECOMMENDATIONS .....</b>	<b>36</b>
<b>6 ACKNOWLEDGEMENTS .....</b>	<b>36</b>
<b>7 APPENDIX A: CRASH DISTANCES .....</b>	<b>37</b>
<b>8 APPENDIX B: ULB DATA.....</b>	<b>39</b>
<b>9 REFERENCES .....</b>	<b>40</b>

Fonte: [https://www.bea.aero/uploads/tx\\_elyextendttnews/metron.search.analysis\\_01.pdf](https://www.bea.aero/uploads/tx_elyextendttnews/metron.search.analysis_01.pdf)

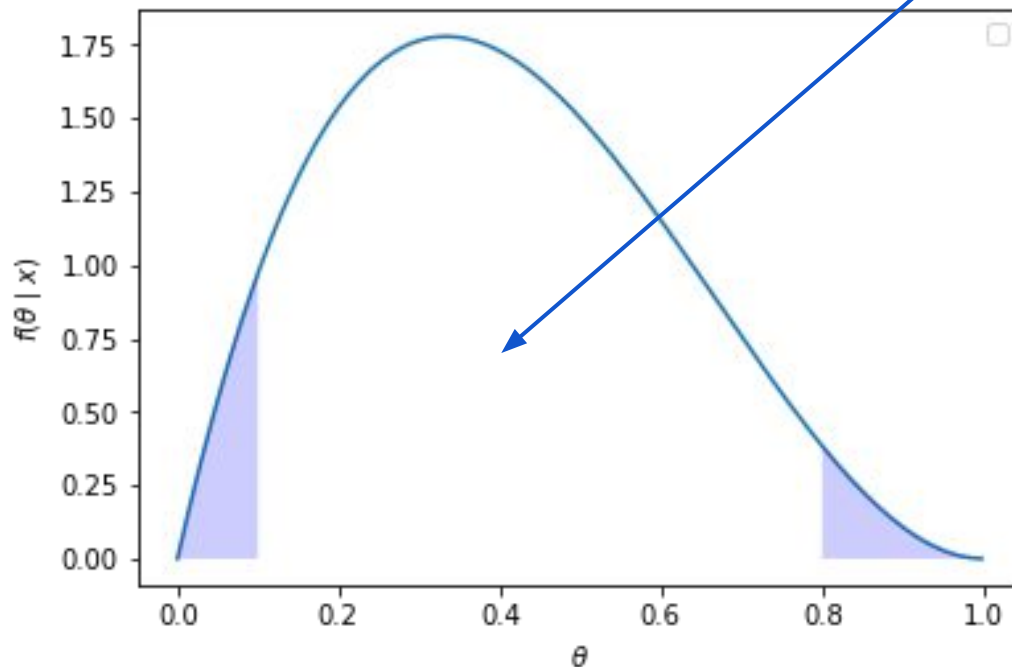
# Estimação: estimador pontual

- Resumir a informação da posteriori
- Função de perda:  $L(\theta, a)$ : “perda” na estimativa de  $\theta$  por  $a$ 
  - $L(\theta, a) > 0$  e, se  $\theta = a$ ,  $L(\theta, a) = 0$
  - Perda quadrática:  $L(\theta, a) = (\theta - a)^2$
  - Perda erro absoluto:  $L(\theta, a) = |\theta - a|$
  - Perda 0-1:  $L(\theta, a) = 0$ , se  $|a - \theta| \leq \varepsilon$ , e  $L(\theta, a) = 1$  se  $|a - \theta| > \varepsilon$
- $a$  que minimiza a perda esperada à posteriori, é o estimador de Bayes:  
minimizar  $\int L(\theta, a) f(\theta | \mathbf{x}) d\theta$
- No caso da perda quadrática,  $a = E(\theta | \mathbf{x})$ : estimador EAP (esperança da posteriori)
- Outra opção usual: MAP (máximo da posteriori)



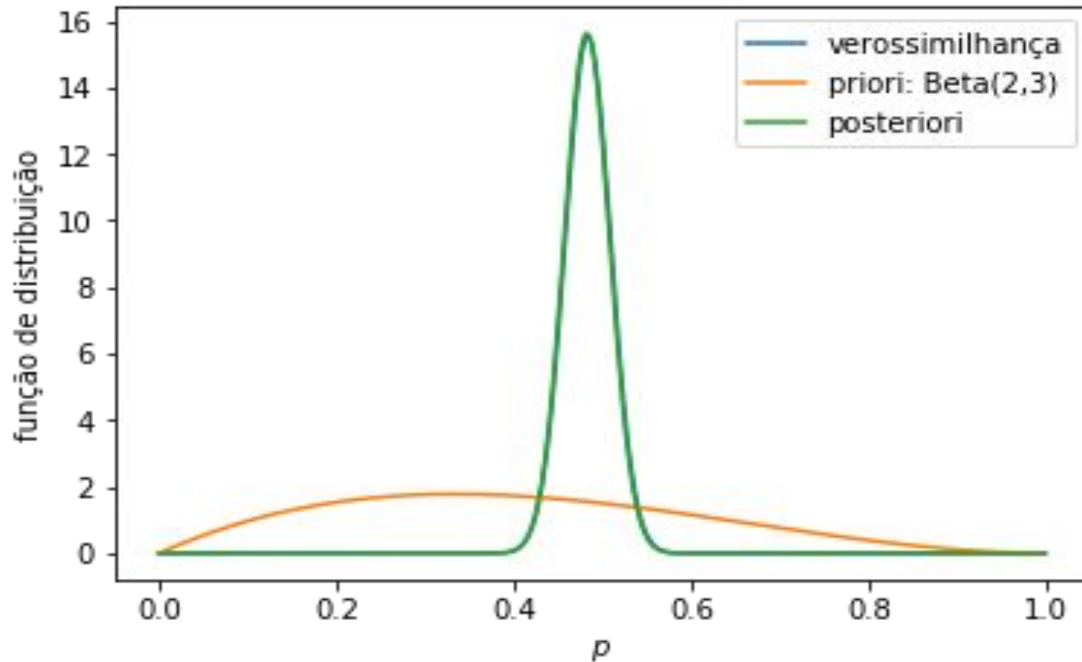
# Estimação: intervalo de credibilidade

- Intervalo de credibilidade 100( $\gamma$ )% para  $\theta$ , (U,L):  $\int_L^U f(\theta | \mathbf{x}) d\theta = \gamma$





# Estimação: exemplo Sildenafil



da teoria:

$$X \sim \text{Beta}(a, b)$$

$$E(X) = \frac{a}{a+b}$$

$$V(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

$$p \mid \mathbf{x} \sim \text{Beta}(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b)$$

$$p \mid \mathbf{x} \sim \text{Beta}(185, 199)$$

$$E(p \mid \mathbf{x}) = \frac{185}{185+199} = 0,482$$

# Distribuição preditiva

Previsões com base no modelo estimado:

- incerteza sobre os valores dos parâmetros estimados com base nos dados
- incerteza pela previsão em si

$$\begin{aligned}f(x_p \mid \mathbf{x}) &= \int f(x_p, \theta \mid \mathbf{x}) d\theta \\&= \int f(x_p \mid \theta, \mathbf{x}) f(\theta \mid \mathbf{x}) d\theta \\&\stackrel{x_p \text{ e } \mathbf{x} \text{ cond indep.}}{=} \int f(x_p \mid \theta) f(\theta \mid \mathbf{x}) d\theta\end{aligned}$$

- verossimilhança da variável a ser predita, ponderada pela posteriori
- cálculo analítico ou por simulação