Notas de aula de Estatística Bayesiana

Lia Hanna Martins Morita

2020

Contents

Materiais de apoio

- (abrir em uma nova janela do navegador)
- contato por email: profaliaufmt@gmail.com ou através do AVA UFMT
- Guia de Estudos
- Apostila da disciplina em pdf
- Aplicação de RJAGS na disciplina
- Material didático Estatística Bayesiana prof. Ricardo Ehlers ;
- Apêndice B do Mood - distribuições ;
- Material didático Inferência Estatística prof. Ricardo Ehlers ;

Novidades

- Atividades síncronas no google meet (dia & horários no Guia de Estudos)
- Tarefas para a próxima aula
- Exercícios para entregar
- Exemplo de resolução de exercício com fórmulas no R Markdown:
  - Exemplo de figura (baixar no computador e colocar na mesma pasta de sua rotina em R, para poder "rodar")
  - Rotina em markdown (para abrir no R studio)
  - Saída em PDF
- Exemplo de relatório - para iniciar os trabalhos em R markdown:
  - **Passo 1:** abrir o R studio;
  - **Passo 2:** criar um novo projeto em um diretório existente (escolher no seu computador);
  - **Passo 3:** criar um novo arquivo formato R markdown;
  - **Passo 4:** sugiro pegar estes exemplos abaixo:
  - Arquivo 1: para fazer um relatorio
  - Arquivo 2: Exemplo de banco de dados
  - Arquivo 3: Exemplo de figura ;
  - se não conseguir baixar um arquivo, basta copiar e colar:
    - se for arquivo .txt, abrir um novo bloco de notas, copiar e colar;
    - se for arquivo .Rmd, abrir um novo arquivo Rmd (no R studio), copiar e colar.
- **Diário de bordo dos encontros síncronos** (contém a descrição detalhada de cada aula síncrona): verificar no fórum do AVA (somente alunos da disciplina podem acessar)
- **Grupo de Whats App** link de convite no AVA

Recursos Computacionais - Sofware R

- download R studio
- download R *base* via *cran mirror* da UFPR
- Aprendendo com R Markdown cookbook

Pacotes

```r
require(captioner)  #para colocar numeração nas figuras e tabelas ao longo do texto (opcional)
```

```
## Loading required package: captioner
```

```r
require(DT)         #para construir tabelas
```

```
## Loading required package: DT
```

Referências na literatura

- BERNARDO, J. M., SMITH, A. F. M.. Bayesian theory. New York: John Wiley & Sons. 1994;
- BERRY, D.A. Statistics: A Bayesian Perspective. Duxbury Press, Belmont, 1996
- BOX, G.E.P.; TIAO, G.C. Bayesian inference in statistical analysis. New York: J. Wiley, 1973. 360p.
- CASELLA, G.; BERGER, R. L. Inferência estatísitca. São Paulo: Cengage Learning, c2011. xi, 588 p.
- DEGROOT, M. H. & SCHERVISH, M. J. Probability and Statistics. New York: Addison Wesley, 2002.
- GELMAN, A., CARLIN, J.B., STERN, H.S., RUBIN, D.B. Bayesian data analysis. 2. ed. London: Chapman and Hall, 2004.
- KINAS, P.G., ANDRADE, H.A. Introdução à análise Bayesiana (com R). Porto Alegre: MaisQnada 2010
- LEE, P.M. Bayesian Statistics: an Introduction. 2. ed. New York: Edward Arnold, 1996
- PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B.. Estatistica bayesiana. Fundacao Calouste Gulbenkian 2003 ed. 446 p.

UNIDADE I – Medição de Incertezas

Os métodos Bayesianos têm aplicação em muitas áreas como epidemiologia, bioestatística, engenharia, ciência da computação, entre outros.

- Thomas Bayes (1764) introduziu a inferência Bayesiana para o modelo binomial com uma *priori* constante;
- Laplace (1862) estudou o resultado de Bayes para qualquer distribuição;
- A teoria das probabilidades foi originalmente introduzida entre 1764 e 1838;
- O conceito de probabilidade inversa foi usado entre 1838 e 1945;
- Fisher introduziu a estatística clássica entre 1938 e 1955;
- 1955 surgiram os testes Bayesianos;
- De Finetti (1974) introduziu a existência da *priori* como principal fundamento da inferência Bayesiana;
- 1990 surgiram os Métodos MCMC (em inglês: Markov Chain Monte Carlo, ou em português: Monte Carlo com cadeias de Markov).

Teoria das probabilidades e axiomas

Definição 1.1: Partição de um Espaço Amostral Dizemos que os eventos $A_1, A_2, \ldots, A_n$ formam uma partição do espaço amostral Ω se as seguintes propriedades são satisfeitas:

- $A_i \neq \emptyset, i = 1, \ldots, n$: significa que nenhum evento pode ser igual ao conjunto vazio;
- $A_i \cap A_j = \emptyset$, para $i \neq j$: significa que os eventos são disjuntos;
- $\cup_{i=1}^{n} A_i = \Omega$: significa que a união (ou reunião) de todos os eventos totaliza o espaço amostral.

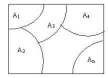


Figure 1: Figura 1: Representação Gráfica de partição de um espaço amostral

Definição 1.2: Classe de eventos do espaço amostral A classe de eventos do espaço amostral Ω, também chamada de *classe de subconjuntos do espaço amostral* Ω, ou *Conjunto das partes de* Ω, é o conjunto que contém todos os subconjuntos de Ω e é representado por $\mathcal{P}(\Omega)$.

Conceito de probabilidade A probabilidade é definida numa classe de eventos do espaço amostral, com certas propriedades.

Definição 1.3: Probabilidade é uma função $P(.)$ que associa a cada evento de $\mathcal{P}(\Omega)$ (ou subconjunto de Ω) um número real pertencente ao intervalo $[0, 1]$, satisfazendo aos Axiomas de Kolmogorov

- Axioma 1: $P(A) \geq 0$ para todo evento A, $A \subset \Omega$: significa que a probabilidade é sempre um número real não negativo;
- Axioma 2: $P(\Omega) = 1$: significa que Ω é um evento certo pois reúne todas as possibilidades;
- Axioma 3: $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, se $A_1, A_2, \ldots$ forem, dois a dois, mutuamente exclusivos:

significa que a probabilidade da união de dois ou mais eventos é igual à soma de suas respectivas probabilidades, se os eventos forem mutuamente exclusivos aos pares.

Teorema 1.4: Se os eventos $A_1, A_2, \ldots, A_n$ formam uma partição do espaço amostral, então:

$$\sum_{i=1}^{n} P(A_i) = 1.$$

Demonstração: A demonstração vem da definição de partição e dos Axiomas 2 e 3 de Kolmogorov.

Definição 1.5: Probabilidade condicional A probabilidade condicional de B dado A é dada pela fórmula:

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

sendo que $P(A)$ deve ser maior do que zero.

Teorema 1.6: Teorema do produto

$$P(A \cap B) = P(A)P(B|A) \text{ e também } P(A \cap B) = P(B)P(A|B).$$

Demonstração: A demonstração vem da definição de probabilidade condicional.

Proposição 1.7: Generalização do Teorema do Produto Sejam A_1, A_2,..., A_{n-1}, A_n eventos do espaço amostral Ω, onde está definida a probabilidade P, temos:

$$P(A_1 \cap A_2 \ldots \cap A_{n-1} \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_n|A1 \cap A2 \ldots A_{n-1}).$$

Demonstração: A demonstração é através do Princípio da Indução Finita.

Teorema 1.8: Teorema da Probabilidade Total (ou Fórmula da Probabilidade Total): Sejam $A_1, A_2, \ldots, A_n$ eventos que formam uma partição do espaço amostral. Seja B um evento deste espaço.

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i),$$

onde $A_1, A_2, \ldots A_n$ formam uma partição no espaço amostral.

A fórmula da Probabilidade Total permite calcular a probabilidade de um evento B a partir das probabilidades de um conjunto de eventos disjuntos cuja reunião é o espaço amostral; e as probabilidades condicionais de B dado cada um destes eventos são fornecidas.

Demonstração - Passo 1: Como $A_1, \ldots, A_n$ formam uma partição, então podemos escrever B da seguinte forma:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup \ldots \cup (B \cap A_n).$$

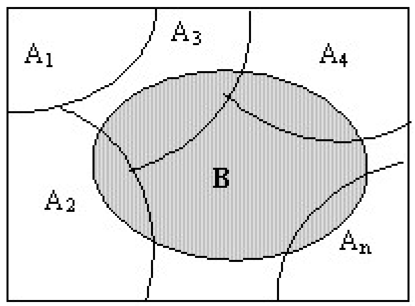


Figure 2: Figura 2: Representação gráfica do teorema da probabilidade total

- Passo 2: Como $A_1, A_2, \ldots, A_n$ são disjuntos, então pelo **axioma 3 de Kolmogorov**, temos:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) \ldots + P(B \cap A_n).$$

- Passo 3: Utilizando o Teorema do Produto, podemos escrever:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + \ldots + P(A_n)P(B|A_n).$$

Teorema 1.9: Teorema de Bayes (ou fórmula de Bayes):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Demonstração: A demonstração vem da Definição de probabilidade condicional, Teorema do Produto e Teorema da Probabilidade Total.

Teorema de Bayes - Caso geral

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)},$$

onde $A_1, A_2, \ldots A_n$ formam uma partição no espaço amostral

Exemplo 1: Exames de diagnóstico não são infalíveis, mas deseja-se que tenha probabilidade pequena de erro. Um exame detecta uma certa doença, caso ela exista, com probabilidade 0,9. Se a doença não existir, o exame corretamente aponta isso com probabilidade 0,8. Considere que estamos aplicando esses exames em uma população com 10% de prevalência dessa doença. Para um indivíduo escolhido ao acaso, pergunta-se:

- a) A probabilidade de ser realmente doente se o exame indicou que era.
- b) A probabilidade de acerto do exame.

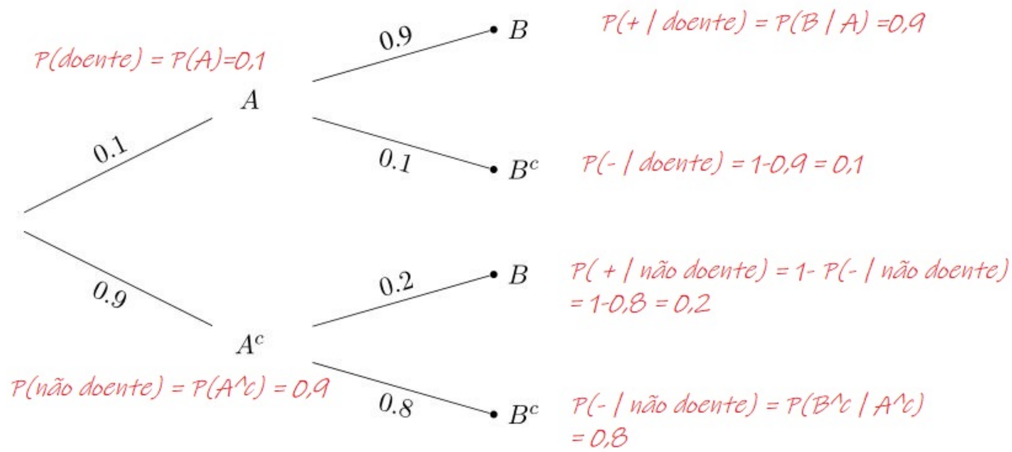


Figure 3: Figura 3: Diagrama da árvore

Resolução:

- Sejam os eventos A: o indivíduo ter a doença e B: o teste dar positivo.
- Podemos construir o diagrama da árvore

- item a) esta probabilidade é denotada por $P(doente|+) = P(A|B)$ não é mesma coisa que $P(B|A)$.

Passo I: Calcular a probabilidade Total - que vai no denominador da probabilidade que queremos:

$$\begin{aligned} P(B) &= P(B|A) \times P(A) + P(B|A^c) \times P(A^c) \\ &= 0,9 \times 0,1 + 0,2 \times 0,9 \\ &= 0,27 \end{aligned}$$

Passo II: Calcular a probabilidade condicional

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(B|A) \times P(A)}{P(B)} \text{ , o numerador vem da fórmula do produto} \\ &= \frac{0,9 \times 0,1}{0,27} = \approx 0,33 = 33\% \end{aligned}$$

- item b) esta probabilidade é dada pela soma: P(+ e o indivíduo ser doente) + P(- e o indivíduo não ser doente):

$$\begin{aligned} P(A \cap B) + P(A^c \cap B^c) &= P(B|A) \times P(A) + P(B^c|A^c) \times P(A^c) \text{ pela fórmula do produto} \\ &= 0,9 \times 0,1 + 0,8 \times 0,9 = 0,81 = 81\% \end{aligned}$$

Exercícios

- 1) Um novo teste para detectar o vírus HIV apresenta 95% de sensitividade e 98% de especificidade. Numa população com uma prevalência de $0,1\%$ para a doença
  - a) qual é a probabilidade de um indivíduo com teste positivo ter o vírus HIV?
  - b) qual é a probabilidade de um indívíduo com teste negativo não ter o vírus HIV?
  - c) Utilize o resultado dos itens *a*) e *b*) para responder à seguinte pergunta: Por que quando o teste dá resultado positivo o laboratório repete o teste, mas do contrário não é necessário repetir o teste?

Ajuda: sensibilidade do teste: é a probabilidade do teste dar resultado positivo para um indivíduo que tem a doença, especificidade do teste: é a probabilidade do teste dar resultado negativo para um indivíduo que não tem doença, prevalência: é a proporção de pessoas com a doença em certa população de interesse. Em testes diagnósticos, temos interesse em encontrar o teste que possui os maiores valores de sensibilidade e especificidade.

- 2) Em um determinado posto de gasolina, 40% dos clientes usam gasolina comum, 35% usam gasolina aditivada e 25% usam gasolina Premium. Dos clientes que usam gasolina comum apenas 30% enchem o tanque; dentre os que usam gasolina aditivada 60% enchem o tanque; e dentre os que usam Premium 50% enchem o tanque.
  - a) Qual é a probabilidade de um cliente encher o tanque, sabendo-se que ele pediu gasolina comum?
  - b) Qual é a probabilidade de um cliente pedir gasolina aditivada e encher o tanque?
  - c) Qual é a probabilidade de um cliente encher o tanque?
  - d) Dado que o cliente encheu o tanque, qual é a probabilidade dele ter pedido gasolina comum? E gasolina aditivada? E gasolina Premium?
- 3) Uma máquina produz 5% de itens defeituosos. Cada item produzido passa por um teste de qualidade que o classifica como bom, defeituoso ou suspeito. Este teste classifica 20% dos itens defeituosos como bons e 30% como suspeitos. Ele também classifica 15% dos itens bons como defeituosos e 25% como suspeitos. Utilize o Teorema de Bayes para responder às perguntas abaixo:
  - a) Que proporção dos itens serão classificados como suspeitos?
  - b) Qual a probabilidade de um item classificado como suspeito ser defeituoso?

Componentes da inferência Bayesiana

- Distribuição *a priori* : utiliza a probabilidade como um meio de quantificar a incerteza sobre quantidades desconhecidas (variáveis), então temos $f(\theta)$: distribuição *a priori* para o parâmetro θ;
- Verossimilhança : relaciona todas as variáveis num modelo de probabilidade completo, então temos $L(\theta|\boldsymbol{y})$: função de verossimilhança de θ dado o conjunto de dados, vem diretametne de $f(\boldsymbol{y}|\theta)$:
- Distribuição *a posteriori* : quando observamos algumas variáveis (os dados), podemos usar a fórmula de Bayes para encontrar as distribuições de probabilidade condicionais para as quantidades de interesse não observadas, então temos $f(\theta|\boldsymbol{y})$: distribuição *a posteriori* para o parâmetro θ.

Nosso principal objetivo é utilizar a distribuição *a posteriori* para a nossa tomada de decisões. Pelo **Teorema de Bayes**, temos:

- a) **Caso discreto:** neste caso, assumimos que θ é uma variável aleatória discreta.

$$P(\theta = \theta_j|\boldsymbol{y}) = \frac{P(\theta = \theta_j)f(\boldsymbol{y}|\theta)}{\sum\limits_j P(\theta = \theta_j)f(\boldsymbol{y}|\theta)},$$

  onde $\theta_j, j = 1, 2, \ldots$ são os valores que θ pode assumir, ou seja, **o espaço paramétrico de θ é discreto**,

- b) **Caso contínuo:** neste caso, assumimos que θ é uma variável aleatória contínua.

$$f(\theta|\boldsymbol{y}) = \frac{f(\theta)f(\boldsymbol{y}|\theta)}{\int\limits_\Theta f(\theta)f(\boldsymbol{y}|\theta)d\theta},$$

  onde Θ é o espaço paramétrico de θ, **o espaço paramétrico de θ é contínuo**

Observações:

- O caso contínuo é mais comumente utilizado na estatística Bayesiana,

- A distribuição *a priori* também pode ser denotada por $p(\theta)$ ou $\pi(\theta)$, assim como a distribuição *a posteriori* denotada por $p(\theta|\boldsymbol{y})$ ou $\pi(\theta|\boldsymbol{y})$,
- Em geral, não é necessário efetuar o cálculo do denominador $\int\limits_{\Theta} f(\theta)L(\theta|\boldsymbol{y})d\theta$ pois se trata de uma constante que não depende de θ, então temos

$$f(\theta|\boldsymbol{y}) \propto \frac{f(\theta)L(\theta|\boldsymbol{y})}{\int\limits_{\Theta} f(\theta)L(\theta|\boldsymbol{y})d\theta},$$

  donde o símbolo $\propto$ significa "é proporcional a",
- As distribuições *a priori* podem ser de vários tipos e características:
- Quanto à propriedade de integrabilidade: existem *prioris* **próprias** ou **impróprias**,
- Quanto ao nível de informação: existem *prioris* **não informativas** ou **informativas**,
- Quanto a depender ou não da amostra (dos dados): existem *prioris* **subjetivas** ou **objetivas**

Exercícios - continuação

- 4) Seja Y uma variável aleatória com distribuição Binomial $Y \sim$ Binomial (n, p).
  - a) Qual é a estimativa de máxima verossimilhança de p?
  - b) Se p tem distribuição *a priori* Beta com parâmetros conhecidos a e b então qual é a distribuição *a posteriori* para p?
  - c) Segundo o item *b*), qual é a média *a posteriori* para p?

Dicas para o item *b*)

$$P(Y = y) =?$$

Qual é a distribuição de Y (os dados)? $f(p) =?$ Qual é a distribuição *a priori* para o parâmetro p? $f(p|\boldsymbol{y}) = ?$ Qual é a distribuição de p condicionada aos dados?

- 5) Foram gerados 10 valores da distribuição de Poisson com taxa $\lambda = 2$, através do seguinte código no software R:

```
set.seed(15052017)
lambda=2  #este é o valor verdadeiro de lambda
n=10
x=rpois(n,lambda)
x
```

```
##  [1] 3 2 0 1 4 4 3 0 3 3
```

- a) Obtenha a estimativa de máxima verossimilhança para λ;
- b) Considerando a distribuição *a priori* Gamma com média 1 e variância 5, obtenha a distribuição *a posteriori* para λ;
- c) Segundo o item *b*), qual é a média *a posteriori* para λ?
- d) Segundo o item *b*), qual é a variância *a posteriori* para λ? Veremos mais tarde que os exercícios 4 e 5 envolvem distribuições *a priori* conjugadas.

Resolução do Exercício 4

Você vai precisar de

- Apêndice B do Mood - distribuições ;
- a) Considerando Y com distribuição de Binomial:

  $P(Y=y) = \begin{pmatrix} n \\ y \end{pmatrix} \times p^y (1-p)^{n-y}$ então
  $\hat{p} = \frac{y}{n}$, ou seja, a e.m.v. de p é igual a contagem do número de sucessos sobre o número de tentativas

- Passo a passo
- b) a *posteriori* para p vem do caso 3 de *prioris* conjugadas (veja seção 2.2):

$$p|\boldsymbol{y} \sim \text{Beta}(a+y, b+n-y), \text{ ou seja} f(p|\boldsymbol{y}) = \frac{1}{B(a+y, b+n-y)} \times p^{a+y-1} \times (1-p)^{b+n-y-1} \times I_{(0,1)}(p)$$

- c) média *a posteriori* - Olhar no apêndice do Mood a fórmula da média:

$$\text{E}(p|\boldsymbol{y}) = \frac{a+y}{(a+y)+(b+n-y)} = \frac{a+y}{a+b+n}$$

- **Aplicação com números:** Em 16 ensaios de Bernoulli, foram obtidos 7 sucessos. Considere uma **priori** Beta de parâmetros $a = 0,5$ e $b = 0,5$.

- Calcule a e.m.v. de p e mostre graficamente.

```
n=16
y=7
logL=function(p){
L=dbinom(y,size=n,prob=p)
log(L)
}
optimize(logL, interval=c(0.01,0.99), maximum=TRUE) #encontra o ponto de máximo
```

```
## $maximum
## [1] 0.4374995
##
## $objective
## [1] -1.620156
```

```
emv=y/n
print(paste0("emv= ",emv))
```

```
## [1] "emv= 0.4375"
```

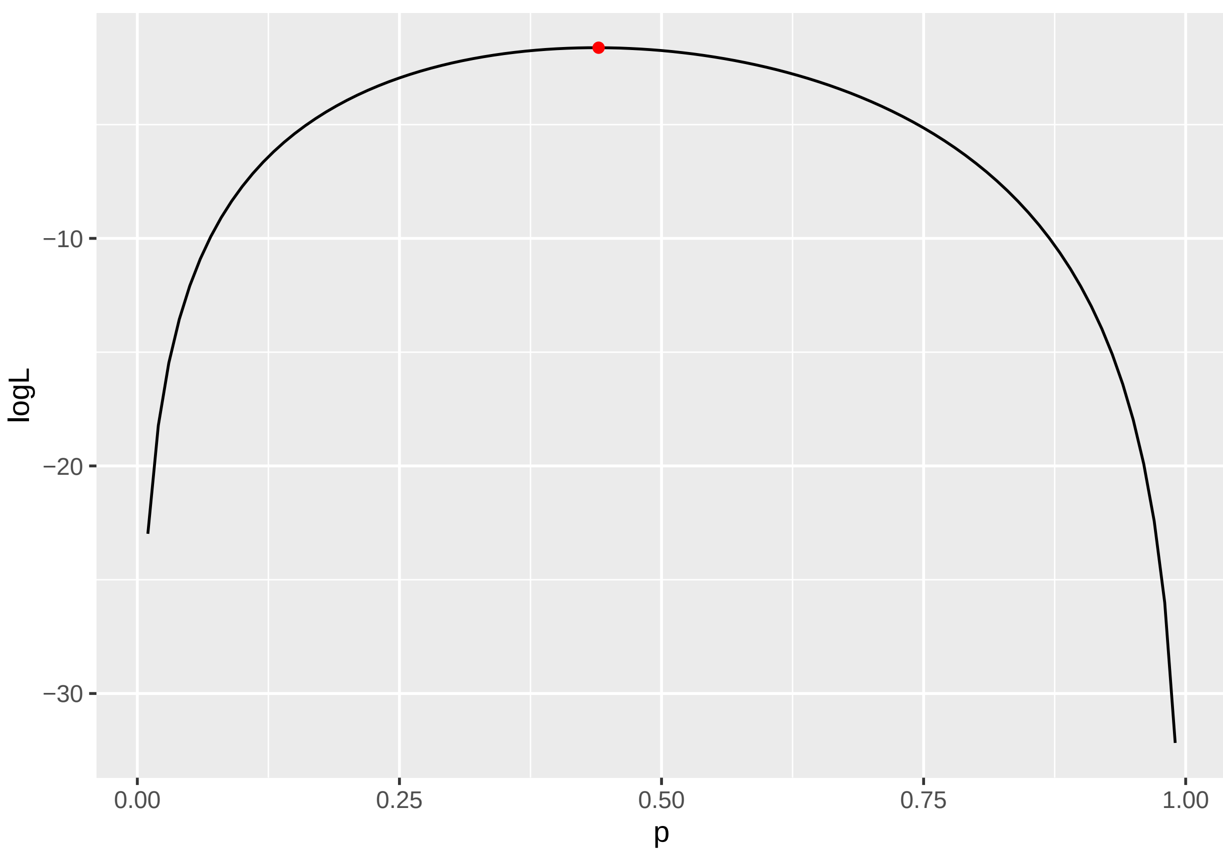
```
p=seq(0.01,0.99,0.01)
temp <- data.frame(p = p, logL = logL(p))
datatable(temp)
```

Show 10 entries Search:

	p	logL
1	0.01	-22.9817730596644
2	0.02	-18.2211141389209
3	0.03	-15.4751668836685
4	0.04	-13.5546574598259
5	0.05	-12.0868942994321
6	0.06	-10.9058823858493
7	0.07	-9.92308522910918
8	0.08	-9.08566372567546
9	0.09	-8.35954411087049
10	0.1	-7.72146902694497

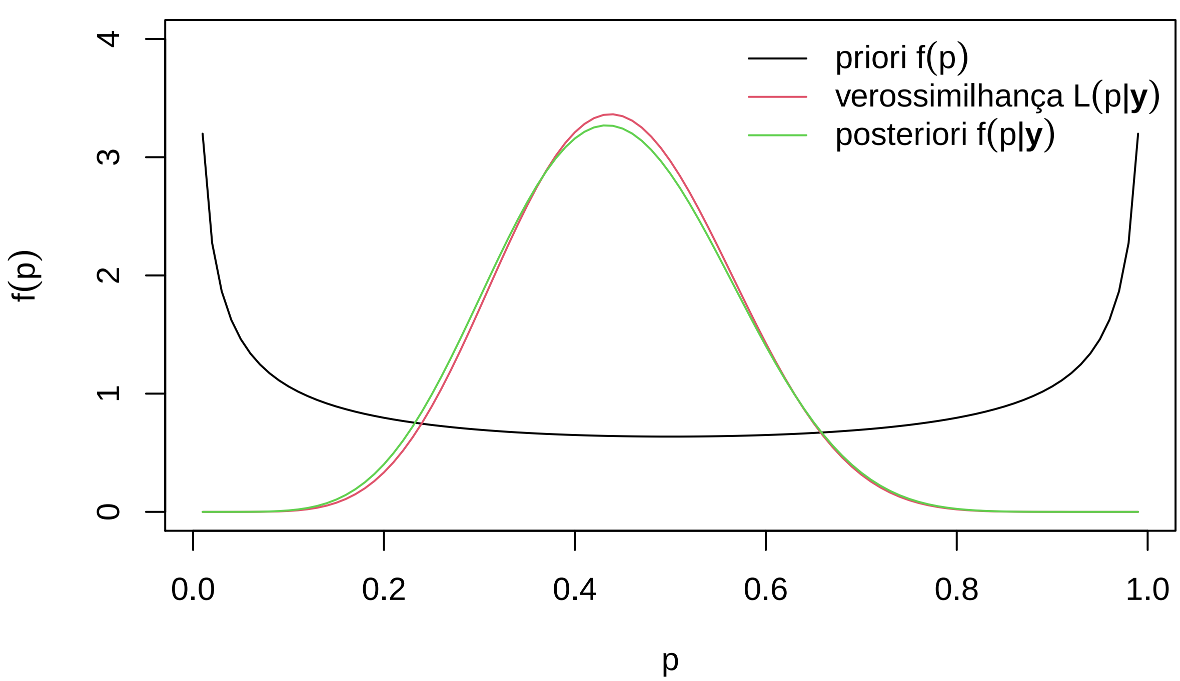
Showing 1 to 10 of 99 entries Previous 1 2 3 4 5 … 10 Next

```
ggplot(data = temp, aes(x = p, y = logL)) + geom_line() + stat_peaks(col = "red")
```



- Implemente os gráficos da *priori*, verossimilhança e *posteriori*:
- Para colocar as três funções no mesmo gráfico, é necessário aproximar a curva da verossimilhança para o formato de uma densidade
  - com soma igual a 1, apenas para fins ilustrativos de construção de gráfico;
  - Sendo assim, para cada valor calculado da função de verossimilhança, divide-se por uma constante que é dada pela soma de todos os valores da função - denotado por "vero_normalizado" na rotina abaixo:

```
a=0.5
b=0.5
p=seq(0.01,0.99,0.01)
pri=dbeta(p,shape1=a,shape2=b)

#Aproxima a verossimilhança para o formato de uma densidade (com soma igual a 1, apenas para fins ilustrat
intervalos=0.01
vero=dbinom(y,size=n,prob=p)
vero_normalizado=vero/sum(vero*intervalos)
pos=dbeta(p,shape1=a+y,shape2=b+n-y)
plot(p,pri,type='l',xlab=expression(p),ylab=expression(f(p)),ylim=c(0,4))
lines(p,vero_normalizado,type='l',col=2)
lines(p,pos,type='l',col=3)
legend("topright",c(expression("priori "*f(p)),expression("verossimilhança "*L(p*"|"*bold(y))),expression(
```



Resolução do Exercício 5

Você vai precisar de

- Apêndice B do Mood - distribuições ;

- a) Considerando uma amostra de tamanho n com distribuição de Poisson:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ então}$$
$$\hat{\lambda} = \bar{x}, \text{ou seja, a e.m.v. de } \lambda \text{ é igual à média amostral}$$

- Passo a passo
- Aplicando nos dados: $\hat{\lambda} = 2.3$, pois

```
x=c(3,2,0,1,4,4,3,0,3,3)
lambda_hat=mean(x) #estimativa de maxima verossimilhança
lambda_hat
```

```
## [1] 2.3
```

- Visualizando a log-verossimilhança numericamente graficamente:

```
n=length(x)
logL=function(lambda){
L=exp(-n*lambda)*lambda^(sum(x))/prod(factorial(x))
log(L)
}
optimize(logL, interval=c(0.01,10), maximum=TRUE) #encontra o ponto de máximo
```

```
## $maximum
## [1] 2.300012
##
## $objective
## [1] -18.05938
```

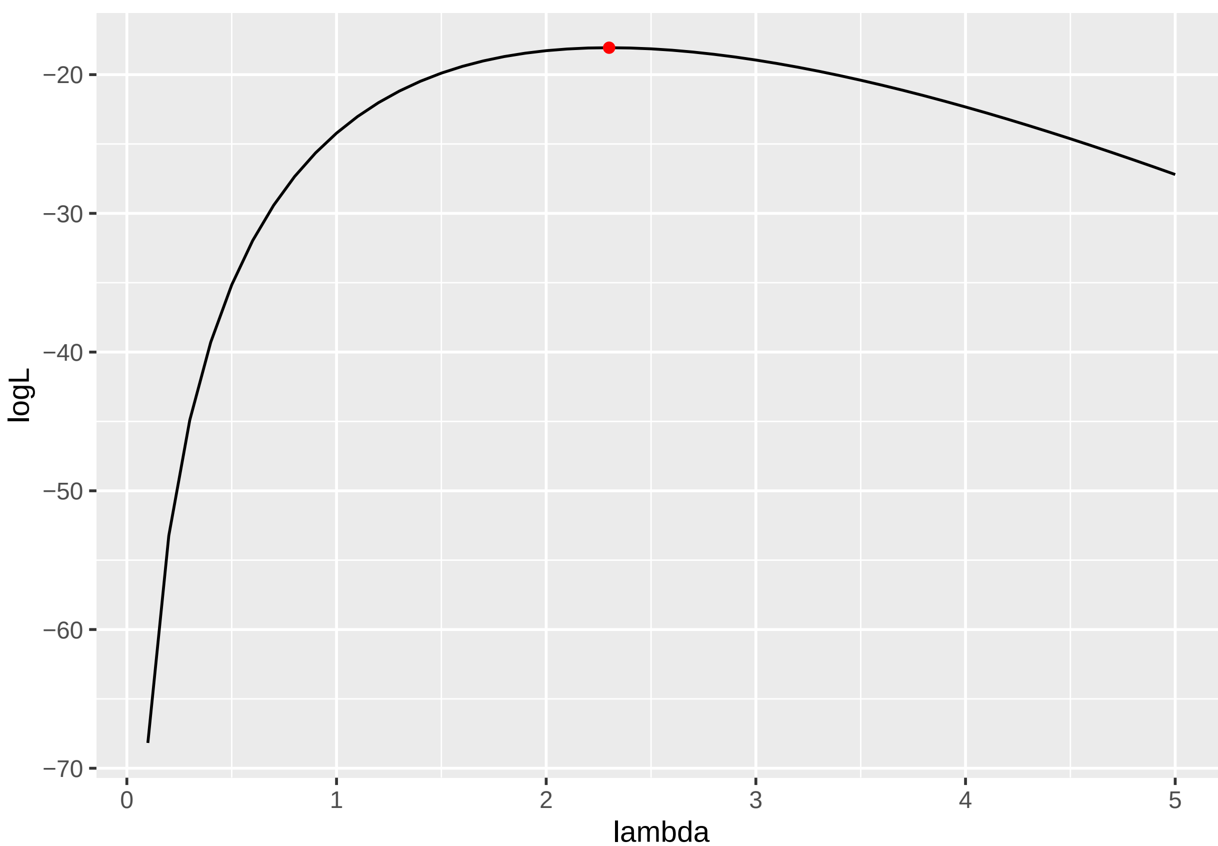
```
lambda=seq(0.1,5,0.1) #só assume valores positivos
temp <- data.frame(lambda=lambda, logL = logL(lambda))
datatable(temp)
```

Show 10 entries Search:

	lambda	logL
1	0.1	-68.1757498570311
2	0.2	-53.2333647041524
3	0.3	-44.9076672176646
4	0.4	-39.2909795512736
5	0.5	-35.1586778710468
6	0.6	-31.9652820647858
7	0.7	-29.4198164287589
8	0.8	-27.3485943983949
9	0.9	-25.6395845782981
10	1	-24.2162927181681

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

```r
ggplot(data = temp, aes(x = lambda, y = logL)) + geom_line() + stat_peaks(col = "red")
```



- b) Seja $\lambda \sim \text{Gamma}(a, r)$, então podemos encontrar os valores de a e r baseados na média e variância *a priori*:

$$f(\lambda) = \frac{a^r}{\Gamma[r]}\lambda^{r-1}e^{-a\lambda}I_{(0,\infty)}(\lambda) \text{ com } \begin{cases} \text{E}(\lambda) = \frac{r}{a} = 1 \\ \text{VAR}(\lambda) = \frac{r}{a^2} = 5 \end{cases} \Rightarrow a = r = \frac{1}{5}$$

Sendo assim, a *posteriori* para λ vem do caso 2 de *prioris* conjugadas (veja seção 2.2):

```
paste0("somatório de x= ",sum(x))
```

```
## [1] "somatório de x= 23"
```

$$\lambda|\boldsymbol{y} \sim \text{Gamma}(a+n, r+\sum_{i=1}^{n} x_i), \text{ ou seja, Como } a = r = \frac{1}{5}, n = 10, \text{ e } \sum_{i=1}^{n} x_i = 23\lambda|\boldsymbol{y} \sim \text{Gamma}\left(\frac{51}{5}, \frac{116}{5}\right)$$

- c) e d) média e variância *a posteriori* - Olhar no apêndice do Mood as fórmulas da média & variância:

$$\begin{aligned} \text{E}(\lambda|\boldsymbol{x}) &= \frac{\frac{116}{5}}{\frac{51}{5}} \approx 2.2745 \\ \text{VAR}(\lambda|\boldsymbol{x}) &= \frac{\frac{116}{5}}{(\frac{51}{5})^2} \approx 0.2230 \end{aligned}$$

- Graficamente: Fazendo os graficos da priori, verossimilhança e posteriori

```
a=1/5
r=1/5

pri_lambda=dgamma(lambda,shape=r, scale=1/a) #na parametrização do R, o parâmetro de  escala (a) é inverti

#Aproxima a verossimilhança para o formato de uma densidade (com soma igual a 1, apenas para fins ilustrat
intervalos=0.1
L_lambda=exp(-n*lambda)*lambda^(sum(x))/prod(factorial(x))
L_lambda_normalizado=L_lambda/sum(L_lambda*intervalos)
pos_lambda=dgamma(lambda,shape=r+sum(x), scale= 1/(a+n)) #na parametrização do R, o parâmetro de escala be
plot(lambda,pri_lambda,type='l',xlab=expression(lambda),ylab=expression(f(lambda)),ylim=c(0,1.5))
lines(lambda,L_lambda_normalizado,type='l',col=2)
lines(lambda,pos_lambda,type='l',col=3)
legend("topright",c(expression("priori "*f(lambda)),expression("verossimilhança "*L(lambda*"|"*bold(x))),e
```

Exemplo 1: Ensaios de Bernoulli com distribuição *a priori* discreta.

- Uma determinada droga tem taxa de resposta θ podendo assumir os seguintes valores *a priori*: $0,2$; $0,4$, $0,6$ ou $0,8$, sendo cada um dos valores com mesma probabilidade de ocorrência. Do resultado de uma amostra unitária, obtivemos sucesso. Como nossa crença pode ser revisada? Podemos representar o problema através de uma tabela.
- Atribui-se *priori* uniforme discreta para a proporção θ;
- E a distribuição *a posteriori* para esta proporção será uniforme discreta.

Resolução: Você vai precisar de

- Apêndice B do Mood - distribuições ;
- Fórmula de Bayes (da seção 1.3) para o caso discreto (pois θ=0,2 ou 0,4 ou 0,6 ou 0,8);
- sem dispensar o termo do denominador - pois é caso discreto.

$$P(\theta = \theta_j|y = 1) = \frac{P(\theta = \theta_j)P(y = 1|\theta = \theta_j)}{\sum_{i=1}^{4} P(\theta = \theta_j)P(y = 1|\theta = \theta_j)} \text{ mesmo raciocínio do Teorema de Bayes!}$$

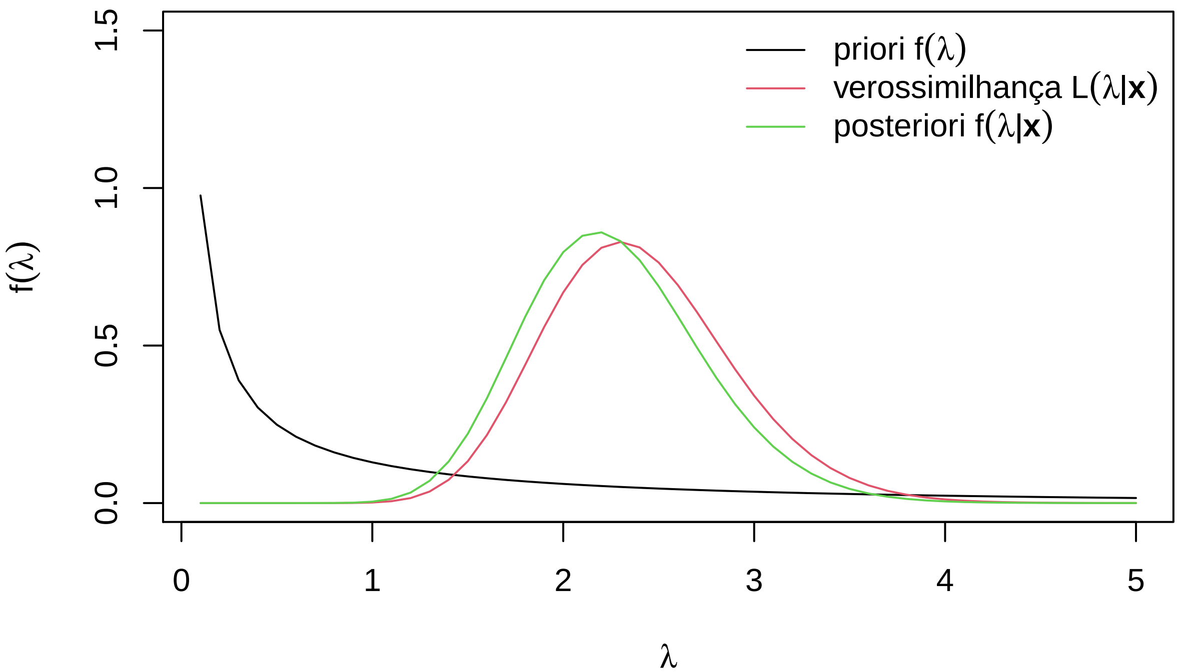


Figure 4: Figura 4: Priori, Verossimilhança e Posteriori

- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Bernoulli), ou seja, temos uma amostra de tamanho 1 de uma distribuição de Bernoulli com proporção igual a θ, e pelo enunciado $y = 1$.

Passo I: atribuir *priori* para θ, sendo θ=0,2 ou 0,4 ou 0,6 ou 0,8:

```
require(kableExtra)
```

```
## Loading required package: kableExtra
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
theta = c(0.2,0.4,0.6,0.8)
priori_theta = c(0.25,0.25,0.25,0.25)

tabela <- data.frame(
          theta,
          priori_theta)

names(tabela) = c("$\\theta_j$","$P(\\theta=\\theta_j)$")

tabela=rbind(tabela,c("Total",1))

 tabela %>%
  kbl(escape = FALSE) %>%
  kable_classic(full_width = F, html_font = "Cambria") %>%
     kable_styling(bootstrap_options = c("striped","hold_position"),
                full_width = F)
```

θ_j	$P(\theta = \theta_j)$
0.2	0.25
0.4	0.25
0.6	0.25
0.8	0.25
Total	1

Passo II: atribuir distribuição Bernoulli para os dados(amostra de tamanho unitário):

$$Y \sim \text{Bernoulli } (\theta) \text{ então } P(Y = 1) = \theta$$

```
theta = c(0.2,0.4,0.6,0.8)
L_theta = theta

tabela <- data.frame(
      theta,
      L_theta)
names(tabela) = c("$\\theta_j$","$P(Y=1 \\vert \\theta = \\theta_j)$")
```

```
 tabela %>%
  kbl(escape = FALSE) %>%
  kable_classic(full_width = F, html_font = "Cambria") %>%
     kable_styling(bootstrap_options = c("striped","hold_position"),
                full_width = F)
```

θ_j	$P(Y=1\|\theta=\theta_j)$
0.2	0.2
0.4	0.4
0.6	0.6
0.8	0.8

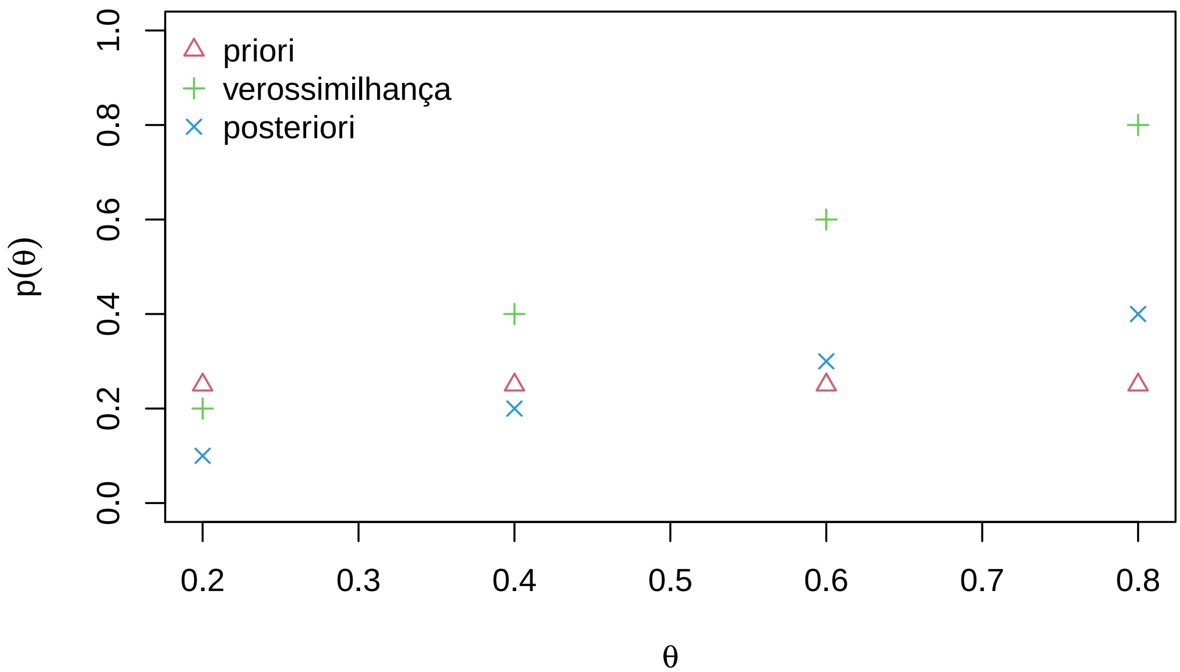
Passo III: Aplicar a fórmula

```
theta = c(0.2,0.4,0.6,0.8)
priori_theta = c(0.25,0.25,0.25,0.25)
L_theta = theta
post_theta=priori_theta*L_theta/sum(priori_theta*L_theta)

tabela <- data.frame(
          theta,
          priori_theta,
          L_theta,
          post_theta)

names(tabela) = c("$\\theta_j$","$P(\\theta=\\theta_j)$","$P(Y=1 \\vert \\theta = \\theta_j)$","$P(\\theta

tabela=rbind(tabela,c("Total",1,"",1))

 tabela %>%
  kbl(escape = FALSE) %>%
  kable_classic(full_width = F, html_font = "Cambria") %>%
     kable_styling(bootstrap_options = c("striped","hold_position"),
                full_width = F)
```

θ_j	$P(\theta=\theta_j)$	$P(Y=1\|\theta=\theta_j)$	$P(\theta=\theta_j\|y=1)$
0.2	0.25	0.2	0.1
0.4	0.25	0.4	0.2
0.6	0.25	0.6	0.3
0.8	0.25	0.8	0.4
Total	1		1

Ilustração:

```
plot(theta,priori_theta,xlab=expression(theta),ylab=expression(p(theta)),type='p',col=2,ylim=c(0,1),pch=2)
lines(theta,L_theta,type='p',col=3,pch=3)
lines(theta,post_theta,type='p',col=4,pch=4)
legend("topleft",c("priori","verossimilhança","posteriori"),col=c(2,3,4),pch=c(2,3,4),bty = "n")
```



Exemplos de aplicação. Material de Inferência Estatística Ricardo Ehlers pag. 7:

1) Considere $X_1, X_2, \ldots, X_n$ uma amostra aleatória de tamanho n da distribuição exponencial com parâmetro λ.

- a) Encontre a expressão para a estimativa de máxima verossimilhança.

Resposta: Considerando a parametrização para a exponencial:

$$f(x) = \lambda e^{-\lambda x} \text{ então}$$
$$\hat{\lambda} = \frac{1}{\bar{x}}, \text{ou seja, a e.m.v. de } \lambda \text{ é dada pelo inverso da média amostral}$$

- Passo a passo

- b) Faça uma aplicação mostrando o gráfico da função de verossimilhança de λ baseado em uma amostra:

```
set.seed(05102020)
lambda=5  #valor utilizado para gerar os dados
n=30
x=rexp(n,lambda)
x_bar=mean(x)
paste0("n= ",n)
```

```
## [1] "n= 30"
```

```r
emv=1/x_bar
paste0("média amostral=",x_bar)
```

```
## [1] "média amostral=0.148695362229909"
```

```r
paste0("emv=",emv)
```

```
## [1] "emv=6.72515931232494"
```

```r
require(tidyverse)
require(ggpmisc)
require(DT)
```

```r
logL=function(lambda){
L=lambda^n *exp(-lambda*sum(x))
log=log(L)
log
}
optimize(logL, interval=c(0.1, 10), maximum=TRUE) #encontra o ponto de máximo
```

```
## $maximum
## [1] 6.725157
##
## $objective
## [1] 27.17567
```

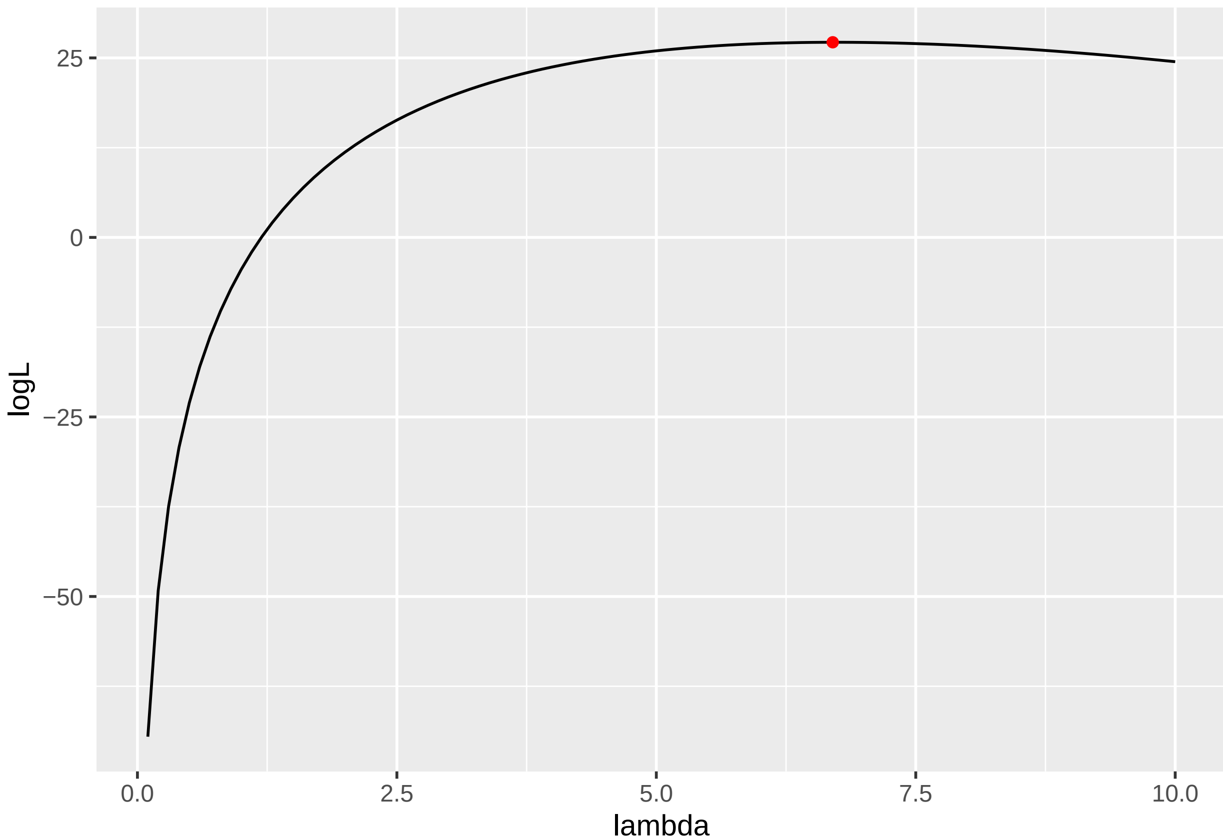
```r
lambda=seq(0.1,10,0.1)
temp <- data.frame(lambda = lambda, logL = logL(lambda))
datatable(temp)
```

Show 10 entries Search:

	lambda	logL
1	0.1	-69.5236388765111
2	0.2	-49.1753095464025
3	0.3	-37.4574423898473
4	0.4	-29.2730663029836
5	0.5	-23.024845850247
6	0.6	-18.0012852331181
7	0.7	-13.82285092499
8	0.8	-10.2629952329441
9	0.9	-7.17559024994232
10	1	-4.46086086689726

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

```r
ggplot(data = temp, aes(x = lambda, y = logL)) + geom_line() + stat_peaks(col = "red")
```



UNIDADE II – Análise Bayesiana de Dados

Prioris conjugadas

Uma família de distribuições *a priori* é conjugada se as distribuições *a posteriori* pertencem a esta mesma família de distribuições.

Casos principais de *prioris* conjugadas

Caso 1: Quando os dados têm distribuição normal com média desconhecida e variância conhecida

- Atribui-se *priori* normal para a média μ;
- E a distribuição *a posteriori* para esta média será normal.

Resolução: Você vai precisar de

- Apêndice B do Mood - distribuições ;
- Fórmula de Bayes (da seção 1.3) para o caso contínuo (pois $-\infty < \mu < \infty$);
- dispensando o termo do denominador (pois é uma constante com relação ao parâmetro μ;

$$f(\mu|\boldsymbol{y}) \propto f(\mu)L(\mu|\boldsymbol{y})$$

- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Normal), ou seja, temos uma amostra de tamanho n i.i.d. (independentes e identicamente distribuídos) de uma distribuição de Normal com média igual a μ e variância igual a σ^2, então $\boldsymbol{y}$ é um vetor de tamanho n.

Passo I: atribuir *priori* para μ, sendo $-\infty < \mu\infty$,

$\mu \sim \text{Normal}(m_0, s_0^2)$:
$f(\mu) = \frac{1}{\sqrt{2\pi}s_0}\exp\left[-\frac{1}{2s_0^2}(\mu - m_0)^2\right]$ considere m_0 e s_0^2 os parâmetros da média e variância da distribuição a priori, respectiva

- com m_0 e s_0^2 conhecidos; - A distribuição *a priori* nos traz o conhecimento *a priori* sobre a média μ; - Se temos pouca informação a respeito de μ, podemos fixar a média m_0 e atribuir uma variância s_0^2 grande; - Se temos muita informação a respeito de μ, podemos fixar a média m_0 e atribuir uma variância s_0^2 pequena.

Passo II: atribuir distribuição Normal para os dados

$$Y_1, Y_2, Y_3, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2) \text{ então } \begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right] \\ &= L(\mu|\boldsymbol{y}) \end{aligned}$$

distribuição dos dados = função de verossimilhança de μ dado os dados

- A função de verossimilhança nos traz toda a informação disponível na amostra (nos dados);
- com μ desconhecido e σ^2 conhecido;

Passo III: Aplicar a fórmula - pode-se dispensar os termos constantes

$$\begin{aligned} f(\mu|\boldsymbol{y}) &\propto f(\mu)L(\mu|\boldsymbol{y}) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{1}{s_0^2}(\mu-m_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right)\right], \\ \text{Então} &\quad \mu|\boldsymbol{y} \sim N\left(\frac{\frac{m_0}{s_0^2}+\frac{n\bar{y}}{\sigma^2}}{\frac{1}{s_0^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{s_0^2}+\frac{n}{\sigma^2}}\right) \end{aligned}$$

- o símbolo $\propto$ significa "é proporcional a", ou seja, todos os termos multiplicativos que não dependem de μ podem ser desconsiderados na fórmula. - A demonstração pode ser encontrada em Box & Tiao (1973)

Prioris Conjugadas - continuação

Exemplo 1 Box & Tiao (1973) Os físicos A e B desejam determinar uma quantidade física μ. O físico A tem mais experiência nesta área e especifica sua priori como $\mu \sim N(900, 20^2)$. O físico B tem pouca experiência e especifica uma priori muito mais incerta em relação à posição de μ: $\mu \sim N(800, 80^2)$. Faz-se então uma medição X de μ em laboratório com um aparelho calibrado com distribuição amostral $X|\mu \sim N(\mu, 40^2)$ e observa-se $X = 850$. Este exemplo corresponde ao "Caso 1)" de *prioris* conjugadas.

Explorando o exemplo e obtendo a *posteriori*:

- para o físico A:
  - distribuição *a priori*: $\mu \sim N(900, 40^2)$;
  - $P(860 < \mu < 940) \approx 0,95$: o intervalo que abrange 95% dos valores é mais estreito.
  - distribuição *a posteriori*: $\mu|\boldsymbol{x} \sim N(890, 320)$.
  - a variância do nosso parâmetro diminuiu $\rightarrow$ significa que ganhamos informação com os dados observados;
- para o físico B:
  - distribuição *a priori*: $\mu \sim N(800, 80^2)$;

- $P(640 < \mu < 960) \approx 0,95$: o intervalo que abrange 95% dos valores é mais largo.
- distribuição *a posteriori*: $\mu|\boldsymbol{x} \sim N(840, 1280)$.
- a variância de μ era igual a 6400, e passou a ser igual a 1280 $\rightarrow$ agregamos informação da amostra;

- A distribuição *a posteriori* representa um compromisso entre a distribuição *a priori* e a verossimilhança. Além disso, como as incertezas iniciais são bem diferentes, o mesmo experimento fornece muito pouca informação adicional para o físico A enquanto que a incerteza do físico B foi bastante reduzida.

- **Cálculos** para este exemplo:

```
require(DT)
```

```
#Informações a priori
m_0=c(900,800)
s_0=c(20,80)

quantis_A=qnorm(c(0.025,0.975),mean=m_0[1],sd=s_0[1])
quantis_B=qnorm(c(0.025,0.975),mean=m_0[2],sd=s_0[2])
quantis=data.frame(prob=c(0.025,0.975),quantis_A,quantis_B)
datatable(quantis,caption="Distribuição a priori")
```

Show 10 entries Search:

Distribuição a priori

	prob	quantis_A	quantis_B
1	0.025	860.800720309199	643.202881236796
2	0.975	939.199279690801	956.797118763204

Showing 1 to 2 of 2 entries Previous 1 Next

```
#Amostra
x=850
n=1
sigma=40
sigma2=sigma^2
x_bar=mean(x)

#Informações a posteriori
media=(m_0/s_0^2+n*x_bar/sigma2)/(1/s_0^2+n/sigma2)
variancia=1/(1/s_0^2+n/sigma2)
```

```
posteriori=data.frame(rbind(media,variancia))
names(posteriori)=c("Fisico A","Fisico B")
datatable(posteriori,caption="Distribuição a posteriori")
```

Show 10 entries Search:

Distribuição a posteriori

	Fisico A	Fisico B
media	890	840
variancia	320	1280

Showing 1 to 2 of 2 entries Previous 1 Next

- Representação Gráfica:

```
mu=seq(500,1000,1)
priori_A=dnorm(mu,mean=m_0[1],sd=s_0[1])
priori_B=dnorm(mu,mean=m_0[2],sd=s_0[2])
L_mu=dnorm(x,mean=mu,sd=40)

#Aproxima a verossimilhança para o formato de uma densidade (com soma igual a 1, apenas para fins ilustrat
intervalos=1
L_mu_normalizado=L_mu/sum(L_mu*intervalos)
posteriori_A=dnorm(mu,mean=media[1],sd=sqrt(variancia[1]))
posteriori_B=dnorm(mu,mean=media[2],sd=sqrt(variancia[2]))
```

```
par(mfrow=c(1,2))
plot(mu,L_mu_normalizado,type='l',xlab=expression(mu),ylab=expression(f(mu)),main="Fisico A",ylim=c(0,0.03
lines(mu,priori_A,type='l',col=1)
lines(mu,posteriori_A,type='l',col=3)
legend("topleft",c(expression("priori: "*mu*"~"*N(900,40^2)),expression("Veross.: "*X*"~"*N(mu,40^2)),expr

plot(mu,L_mu_normalizado,type='l',xlab=expression(mu),ylab=expression(f(mu)),main="Fisico B",ylim=c(0,0.03
lines(mu,priori_B,type='l',col=1)
lines(mu,posteriori_B,type='l',col=3)
legend("topleft",c(expression("priori: "*mu*"~"*N(800,80^2)),expression("Veross.: "*X*"~"*N(mu,40^2)),expr
```

Caso 2: Quando os dados têm distribuição de Poisson

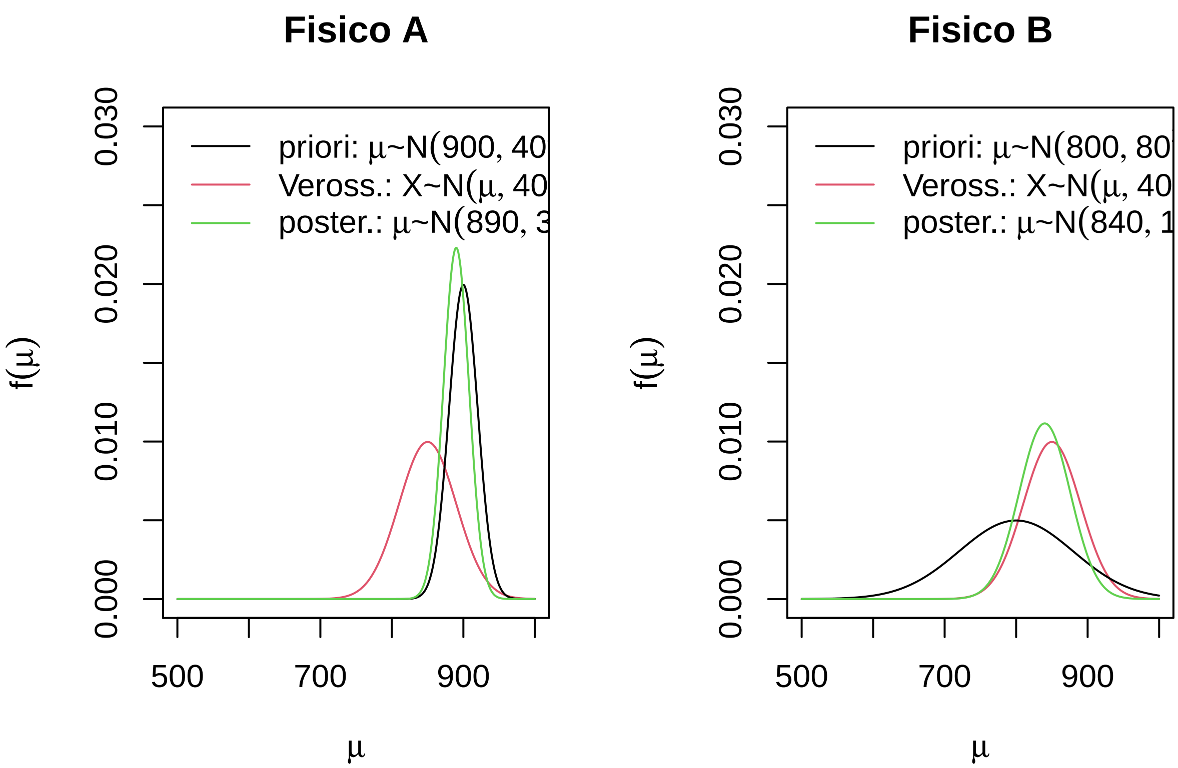


Figure 5: Figura 5: Gráficos das três funções conjuntamente: priori, verossimilhança e posteriori

- Atribui-se *priori* Gmma para a taxa λ;
- E a distribuição *a posteriori* para esta taxa será Gamma.

Resolução: Você vai precisar de

- Apêndice B do Mood - distribuições ;
- Fórmula de Bayes (da seção 1.3) para o caso contínuo (pois $0 < \lambda < \infty$);
- dispensando o termo do denominador (pois é uma constante com relação ao parâmetro λ;

$$f(\lambda|\boldsymbol{y}) \propto f(\lambda)L(\lambda|\boldsymbol{y})$$

- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Poisson), ou seja, temos uma amostra de tamanho n i.i.d. (independentes e identicamente distribuídos) de uma distribuição de Poisson com taxa λ, então $\boldsymbol{y}$ é um vetor de tamanho n.

Passo I: atribuir *priori* para λ, sendo $0 < \lambda < \infty$,

$$\begin{aligned} &\lambda \sim \text{Gamma}(a, r) \text{ então} \\ &f(\lambda) = \underbrace{\frac{a^r}{\Gamma[r]}\lambda^{r-1}e^{-a\lambda}I_{(0,\infty)}(\lambda)}_{\text{considere } a \text{ o primeiro parâmetro da Gamma, para não confundir com } \lambda \text{ na fórmula}} \end{aligned}$$

Passo II: atribuir distribuição Poisson para os dados

$$Y_1, Y_2, Y_3, \ldots, Y_n \sim \text{Poisson}(\lambda) \text{ então} \quad \begin{aligned} \underbrace{P(\boldsymbol{Y} = \boldsymbol{y})}_{\text{vetor } \boldsymbol{Y} \text{ ser igual a "vetorzinho" } \boldsymbol{y}\text{, de valores observados}} &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \times \underbrace{\prod_{i=1}^n I_{\{0,1,\ldots\}}(y_i)}_{\text{produto de funções indicado}} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} \times I_{\{0,1,\ldots\}}(\prod_{i=1}^n y_i) \\ &= L(\lambda|\boldsymbol{y}) \end{aligned}$$

Passo III: Aplicar a fórmula - pode-se dispensar os termos constantes

$$\begin{aligned} f(\lambda|\boldsymbol{y}) &\propto f(\lambda)L(\lambda|\boldsymbol{y}) \\ &\propto \tfrac{a^r}{\Gamma[r]}\lambda^{r-1}e^{-a\lambda}I_{(0,\infty)}(\lambda) \times \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!} \\ &\propto \lambda^{r+\sum_{i=1}^n y_i - 1}e^{-(a+n)\lambda}I_{(0,\infty)}(\lambda) \\ &\text{Então} \quad \lambda|\boldsymbol{y} \sim \text{Gamma}(a+n, r+\textstyle\sum_{i=1}^n y_i) \text{ é só incluir as constantes na fórmula:} \\ f(\lambda|\boldsymbol{y}) &= \frac{(a+n)^{r+\sum_{i=1}^n y_i}}{\Gamma[r+\sum_{i=1}^n y_i]}\lambda^{r+\sum_{i=1}^n y_i - 1}e^{-(a+n)\lambda}I_{(0,\infty)}(\lambda) \end{aligned}$$

- Na **inferência clássica**, nós procedemos com o método usual de maximizar $\log(L(\lambda|\boldsymbol{y}))$ com respeito a λ, pois estamos tratando de um **parâmetro fixo e desconhecido**;
- Na **inferência Bayesiana**, ao invés de encontrar a **estimativa de máxima verossimilhança**, nós estamos interessados na distribuição *a posteriori*, pois estamos tratando de um **parâmetro desconhecido** cuja distribuição *a priori* é a nossa crença *a priori* antes de observar os dados.

Caso 3: Quando os dados têm distribuição Binomial

- Atribui-se *priori* Beta para a proporção p;
- E a distribuição *a posteriori* para esta proporção será Beta.

Resolução: Você vai precisar de

- Apêndice B do Mood - distribuições ;

- Fórmula de Bayes (da seção 1.3) para o caso contínuo (pois 0<p<1);
- dispensando o termo do denominador (pois é uma constante com relação ao parâmetro p;

$$f(p|\boldsymbol{y}) \propto f(p)L(p|\boldsymbol{y})$$

- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Binomial), ou seja, quando realizados de n ensaios de Bernoulli independentes com probabilidade igual a p, verificou-se y sucessos.

Passo I: atribuir *priori* para p, sendo $0 < p < 1$,

$$\begin{array}{l} p \sim \text{Beta}(a, b) \text{ então} \\ f(p) = \frac{1}{B(a,b)} \times p^{a-1}(1-p)^{b-1} I_{(0,1)}(p) \end{array}$$

Passo II: atribuir distribuição Binomial para os dados

$$\begin{array}{ll} Y \sim \text{Binominal}(n, p) \text{ então} & \\ P(Y = y) = & \underbrace{\binom{n}{y} \times p^y (1-p)^{n-y} = L(p|\boldsymbol{y})}_{\text{distribuição dos dados = função de verossimilhança de } p \text{ dado os dados}} \end{array}$$

Passo III: Aplicar a fórmula - pode-se dispensar os termos constantes

$$\begin{array}{lll} f(p|\boldsymbol{y}) & \propto & f(p)L(p|\boldsymbol{y}) \\ & \propto & \frac{1}{B(a,b)} \times p^{a-1}(1-p)^{b-1} I_{(0,1)}(p) \times \binom{n}{y} \times p^y(1-p)^{n-y} \\ & \propto & p^{a+y-1} \times (1-p)^{b+n-y-1} \times I_{(0,1)}(p) \\ & \text{Então} & p|\boldsymbol{y} \sim \text{Beta}(a+y, b+n-y) \text{ é só incluir as constantes na fórmula:} \\ f(p|\boldsymbol{y}) & = & \frac{1}{B(a+y,b+n-y)} \times p^{a+y-1} \times (1-p)^{b+n-y-1} \times I_{(0,1)}(p) \end{array}$$

- Também pode-se mostrar que sendo $Y_1, Y_2, \ldots, Y_N \sim$ Binomial (n, p) a priori conjugada é Beta - com parâmetros atualizados para N experimentos de Bernoulli.

- Veja abaixo um exemplo no R para este cenário - olhe o código

```r
N = 20 #Número de experimentos
n = 10 #Número de ensaios de Bernoulli para cada experimento
p = 0.4 #probabilidade de sucesso para cada ensaio de Bernoulli
set.seed(18102020)
y = rbinom(N, n, p) #vetor y
y
```

```
##  [1] 4 4 5 5 7 5 5 3 5 5 4 6 6 4 4 5 3 5 7 5
```

Caso 4: Quando os dados têm distribuição normal com média conhecida e variância desconhecida

- Atribui-se *priori* Gamma invertida para a variância σ^2;
- E a distribuição *a posteriori* para esta variância será Gamma invertida.

Resolução: Você vai precisar de

- Apêndice A do material do prof. Ricardo Ehlers - distribuições ;
- Fórmula de Bayes (da seção 1.3) para o caso contínuo (pois $0 < \sigma^2 < \infty$);
- dispensando o termo do denominador (pois é uma constante com relação ao parâmetro σ^2;

$$f(\sigma^2|\boldsymbol{y}) \propto f(\sigma^2)L(\sigma^2|\boldsymbol{y})$$

- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Normal), ou seja, temos uma amostra de tamanho n i.i.d. (independentes e identicamente distribuídos) de uma distribuição de Normal com média igual a μ e variância igual a σ^2, então $\boldsymbol{y}$ é um vetor de tamanho n.

Passo I: atribuir *priori* para σ^2, sendo $0 < \sigma^2 < \infty$,

$$\begin{array}{l} \sigma^2 \sim \text{Gamma Invertida}(\alpha, \beta) \text{ então} \\ f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha+1} \exp\left[-\frac{\beta}{\sigma^2}\right] \end{array}$$

com $\alpha > 0$ e $\beta > 0$ parâmetros conhecidos.

Passo II: atribuir distribuição Normal para os dados

$$Y_1, Y_2, Y_3, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2) \text{ então} \quad \begin{array}{rcl} f_{\boldsymbol{Y}}(\boldsymbol{y}) & = & \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\ & = & \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right] \text{ distribuição dos dados = função de v} \\ & = & L(\sigma^2|\boldsymbol{y}) \end{array}$$

- A função de verossimilhança nos traz toda a informação disponível na amostra (nos dados); - com μ conhecido e σ^2 desconhecido;

Passo III: Aplicar a fórmula - pode-se dispensar os termos constantes

$$\begin{array}{lll} f(\sigma^2|\boldsymbol{y}) & \propto & f(\sigma^2)L(\sigma^2|\boldsymbol{y}) \\ & \propto & \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha+1} \exp\left[-\frac{\beta}{\sigma^2}\right] \times \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right] \\ & \propto & \left(\sigma^2\right)^{-\left(\alpha+\frac{n}{2}\right)+1} \exp\left[-\frac{\beta + \frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2}{\sigma^2}\right] \\ & \text{Então} & \sigma^2|\boldsymbol{y} \sim \text{Gamma Invertida}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right), \text{ é só incluir as constantes na fórmula} \\ f(\sigma^2|\boldsymbol{y}) & = & \frac{\left(\beta + \frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2\right)^{\alpha+\frac{n}{2}}}{\Gamma[\alpha+\frac{n}{2}]} \left(\sigma^2\right)^{-\left(\alpha+\frac{n}{2}\right)+1} \exp\left[-\frac{\beta + \frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2}{\sigma^2}\right] \end{array}$$

Caso 5: Quando os dados têm distribuição normal com média conhecida e variância desconhecida

- Atribui-se *priori* Gamma para a precisão $\tau = \frac{1}{\sigma^2}$;
- E a distribuição *a posteriori* para esta precisão será Gamma.

Resolução: Você vai precisar de

- Apêndice B do Mood - distribuições ;
- Fórmula de Bayes (da seção 1.3) para o caso contínuo (pois $0 < \tau < \infty$);
- dispensando o termo do denominador (pois é uma constante com relação ao parâmetro τ;

$$f(\tau|\boldsymbol{y}) \propto f(\tau)L(\tau|\boldsymbol{y})$$

- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Normal), ou seja, temos uma amostra de tamanho n i.i.d. (independentes e identicamente distribuídos) de uma distribuição de Normal com média igual a μ e precisão igual a τ, então $\boldsymbol{y}$ é um vetor de tamanho n.

Passo I: atribuir *priori* para τ, sendo $0 < \tau < \infty$,

$$\begin{array}{l} \tau \sim \text{Gamma}(\lambda, r) \text{ então} \\ f(\tau) = \frac{\lambda^r}{\Gamma(r)} \tau^{r-1} e^{-\lambda\tau} I_{(0,\infty)}(\tau) \end{array}$$

com $\lambda > 0$ e $r > 0$ parâmetros conhecidos (chamados de hiperparâmetros no contexto Bayesiano)

Passo II: atribuir distribuição Normal para os dados

$Y_1, Y_2, Y_3, \ldots, Y_n \sim \text{Normal}(\mu, \sigma^2)$ então $\begin{array}{rcl} f_{\boldsymbol{Y}}(\boldsymbol{y}) & = & \prod_{i=1}^{n} \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left[-\frac{\tau}{2}(y_i-\mu)^2\right] \\ & = & \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\ & = & L(\tau|\boldsymbol{y}) \end{array}$ distribuição dos dados = função de veross

- A função de verossimilhança nos traz toda a informação disponível na amostra (nos dados); - com μ conhecido e τ desconhecido;

Passo III: Aplicar a fórmula - pode-se dispensar os termos constantes

$$\begin{array}{rcl} f(\tau|\boldsymbol{y}) & \propto & f(\tau)L(\tau|\boldsymbol{y}) \\ & \propto & \frac{\lambda^r}{\Gamma(r)}\tau^{r-1}e^{-\lambda\tau}I_{(0,\infty)}(\tau) \times \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\ & \propto & \tau^{r+\frac{n}{2}-1}\exp\left[-\left(\lambda+\sum_{i=1}^{n}(y_i-\mu)^2\right)\tau\right]I_{(0,\infty)}(\tau) \\ & \text{Então} & \tau|\boldsymbol{y} \sim \text{Gamma}\left(\lambda+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2, r+\frac{n}{2}\right), \text{ é só incluir as constantes na fórmula} \\ f(\tau|\boldsymbol{y}) & = & \frac{\left(\lambda+\frac{1}{2}\sum\limits_{i=1}^{n}(y_i-\mu)^2\right)^{r+\frac{n}{2}}}{\Gamma[r+\frac{n}{2}]}\tau^{r+\frac{n}{2}-1}\exp\left[-\left(\lambda+\sum_{i=1}^{n}(y_i-\mu)^2\right)\tau\right]I_{(0,\infty)}(\tau) \end{array}$$

- E se dizemos que a distribuição *a posteriori* para τ é gamma:
- É equivalente a dizer que a distribuição *a posteriori* para σ^2 é gamma invertida:

Exemplo 2 Abaixo temos 10 valores provenientes de uma distribuição normal com média μ e variância σ^2.

```
set.seed(05062017) #cria uma semente única
n=10
mu=2 #este é o valor da média mu
sigma2=3 #este é o valor da variancia sigma2 para a criação dos dados
y=rnorm(n,mean=mu,sd=sqrt(sigma2))
y=round(y,4)
y
```

```
##  [1]  1.2156  1.2000  2.1362  2.1139  2.6546  0.0135 -0.0007  0.2131  3.3849
## [10]  4.9196
```

- a) Obtenha a estimativa de máxima verossimilhança para σ^2;
- b) Considere a média conhecida e igual a 2, e a variância desconhecida. Considere a distribuição *a priori* Gamma com média 0.5 e variância 0.5 para a precisão $\tau = \frac{1}{\sigma^2}$, obtenha a distribuição *a posteriori* para τ;
- c) Segundo o item b), qual é a média *a posteriori* para τ?
- d) Segundo o item b), qual é a varância *a posteriori* para τ?

Solução: Você vai precisar de

- Apêndice B do Mood - distribuições .
- a) A estimativa de máxima verossimilhança para σ^2 é $\widehat{\sigma^2} = 2.2837$, pois

  $$Y_1, Y_2, \ldots, Y_n \text{ variáveis aleatórias i.i.d.} \sim N(\mu, \sigma^2), \text{ então} f_{Y_i}(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \text{ e aplicando o produt}$$

  e também pode ser verificado que a segunda derivada avaliada na e.m.v. tem sinal negativo.

- Passo a passo na página 27 ;
- Fazendo os cálculos no R:

```
s2=var(y)
sigma2_hat=s2*(n-1)/n
print(paste0("sigma2_hat= ",sigma2_hat))
```

```
## [1] "sigma2_hat= 2.2837365641"
```

- Outro modo:

```
sigma2_hat=sum((y-mean(y))^2)/n
print(paste0("Modo alternativo: sigma2_hat= ",sigma2_hat))
```

```
## [1] "Modo alternativo: sigma2_hat= 2.2837365641"
```

- Visualizando a log-verossimilhança:

```
logL=function(sigma2){
L=(1/sqrt(2*pi*sigma2))^n*exp(-1/(2*sigma2)*sum((y-mean(y))^2))
log(L)
}
optimize(logL, interval=c(0.01,10), maximum=TRUE) #encontra o ponto de máximo
```

```
## $maximum
## [1] 2.283739
##
## $objective
## [1] -18.31845
```

```
sigma2=seq(0.01,20,0.01) #só assume valores positivos
temp <- data.frame(sigma2=sigma2, logL = logL(sigma2))
datatable(temp)
```

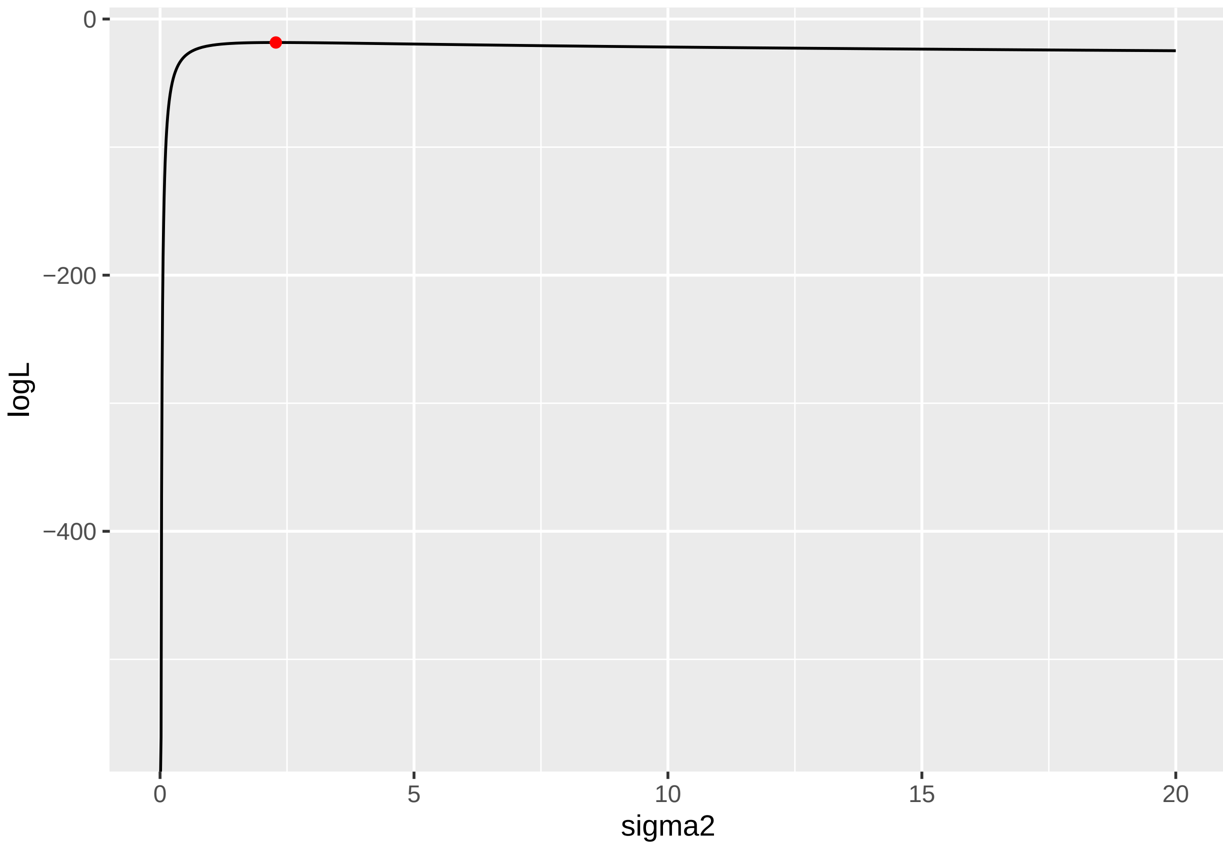
Show 10 entries Search:

	sigma2	logL
1	0.01	
2	0.02	-560.563411329906
3	0.03	-372.27935652878
4	0.04	-278.562076720206
5	0.05	-222.584380374277
6	0.06	-185.433712089913
7	0.07	-159.01712544024
8	0.08	-139.294277366755
9	0.09	-124.023910849898
10	0.1	-111.863288072076

Showing 1 to 10 of 2,000 entries Previous 1 2 3 4 5 … 200 Next

```
ggplot(data = temp, aes(x = sigma2, y = logL)) + geom_line() + stat_peaks(col = "red")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_peaks).
```



- b) Seja $\tau \sim \text{Gamma}(\lambda, r)$, então podemos encontrar os valores de λ e r baseados na média e variância *a priori*:

$$f(\tau) = \frac{\lambda^r}{\Gamma[r]}\tau^{r-1}e^{-\lambda\tau}I_{(0,\infty)}(\tau) \text{ com } \begin{cases} \text{E}(\tau) = \frac{r}{\lambda} = 0.5 \\ \text{VAR}(\tau) = \frac{r}{\lambda^2} = 0.5 \end{cases} \Rightarrow \lambda = 1 \text{ e } r = \frac{1}{2}$$

  E a *posteriori* para τ vem do caso 5 de *prioris* conjugadas (veja seção 2.2):

$$\tau|\boldsymbol{y} \sim \text{Gamma}\left(\lambda + \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2, r + \frac{n}{2}\right), \text{ ou seja, Como } \lambda = 1, r = \frac{1}{2} \text{ e } \mu = 2 : \tau|\boldsymbol{y} \sim \text{Gamma}\left(\frac{2+\sum_{i=1}^{n}(y_i-2)^2}{2}, \frac{1+n}{2}\right)$$

- Fazendo os cálculos no R: $\tau|\boldsymbol{y} \sim \text{Gamma}(12.6496, 5.5)$

```
par1=(2+sum((y-2)^2))/2
par2=(1+n)/2
print(paste0("par_1= ",par1," e par_2= ",par2))
```

```
## [1] "par_1= 12.649657345 e par_2= 5.5"
```

- c) e d) Média e variância *a posteriori*:

$$\begin{aligned} \text{E}(\tau|\boldsymbol{y}) &= \tfrac{5.5}{12.6497} \approx 0.4351, \text{ próxima da e.m.v: } \widehat{\tau} = \tfrac{1}{2.2837} = 0.4379, \\ \text{VAR}(\tau|\boldsymbol{y}) &= \tfrac{5.5}{12.6497^2} = 0.0344 \end{aligned}$$

- Graficamente:

```
tau=seq(0,1.5,0.01)
lambda=1
r=1/2
mu=2
constante=exp(19)

priori=dgamma(tau,shape=r, scale=1/lambda)
L_tau=(1/sqrt(2*pi))^n*tau^(n/2)*exp(-tau/2*sum((y-mu)^2))

#Aproxima a verossimilhança para o formato de uma densidade (com soma igual a 1, apenas para fins ilustrat
intervalos=0.01
L_tau_normalizado=L_tau/sum(L_tau*intervalos)
posteriori=dgamma(tau,shape=par2, scale=1/par1)

plot(tau,priori,type='l',xlab=expression(tau),ylab=expression(f(tau)))
lines(tau,L_tau_normalizado,type='l',col=2)
lines(tau,posteriori,type='l',col=3)
legend("topright",c(expression("priori "*f(tau)),expression("verossimilhança "*L(tau*"|"*bold(y))),express
```

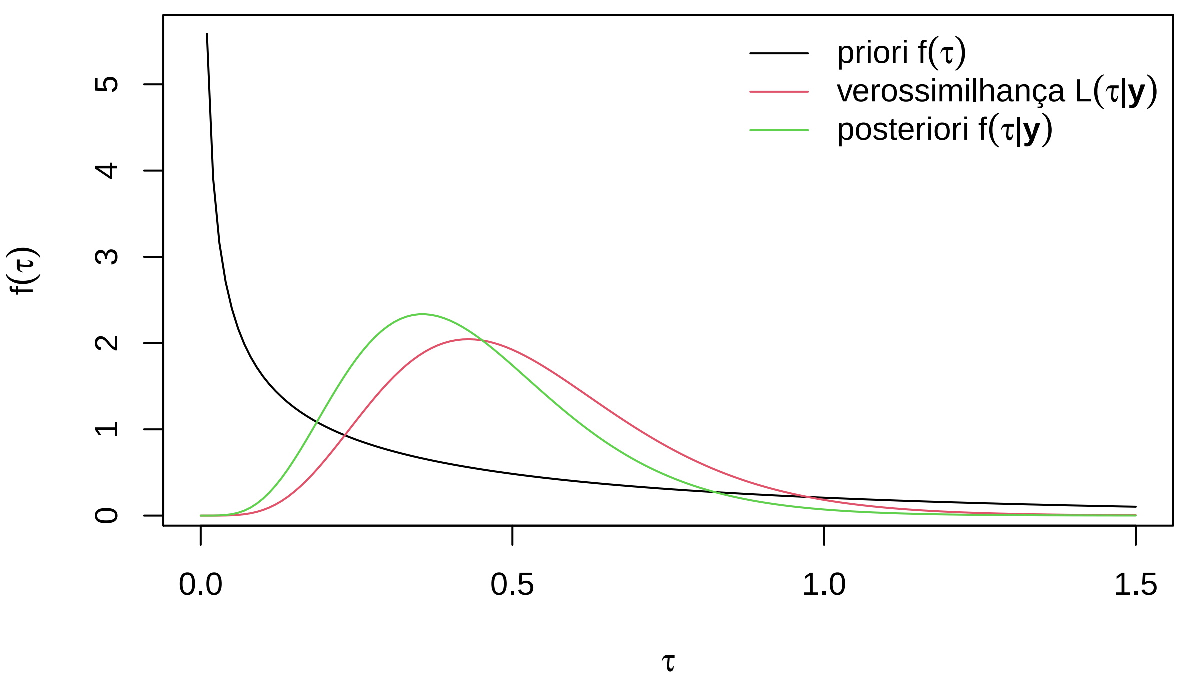


Figure 6: Figura 6: Gráficos do exemplo 2

Caso 6: Vemos que a família de distribuições Beta é conjugada a distribuição Geométrica:

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias i.i.d. com distribuição geométrica com probabilidade de sucesso igual a p.

- Distribuição geométrica (segundo apêndice do Mood): Cada Y_i é igual ao número de tentativas anteriores ao primeiro sucesso em um experimento com ensaios independentes de Bernoulli com probabilidade de sucesso igual a p.

- Atribui-se *priori* Beta para a proporção p;
- E a distribuição *a posteriori* para esta proporção será Beta.

Resolução:

- Pela fórmula de Bayes, dispensando o termo do denominador (pois é uma constante com relação ao parâmetro p;
$$f(p|\boldsymbol{y}) \propto f(p)L(p|\boldsymbol{y})$$
- o termo $\boldsymbol{y}$ se refere aos dados observados (Modelo Geométrico), ou seja, quando realizados ensaios de Bernoulli independentes com probabilidade igual a p, verificou-se os valores observados $y_1, y_1, \ldots, y_n$ das tentativas anteriores ao primeiro sucesso.

Passo I: atribuir *priori* para p, sendo $0 < p < 1$,

$$\begin{array}{l} p \sim \text{Beta}(a, b) \text{ então} \\ f(p) = \frac{1}{B(a,b)} \times p^{a-1}(1-p)^{b-1} I_{(0,1)}(p) \end{array}$$

Passo II: atribuir distribuição geométrica para os dados

$$\begin{array}{lll} Y_1, Y_2, \ldots, Y_n \sim \text{Geom}(p) \text{ então} & & \\ P(\boldsymbol{Y} = \boldsymbol{y}) & = & \underbrace{\prod_{i=1}^{n} p(1-p)^{y_i}}_{\text{os experimentos são independentes}} \quad \text{distribuição dos dados = função de verossimilhança de } p \text{ da} \\ & = & p^n (1-p)^{\sum_{i=1}^{n} y_i} \\ & = & L(p|\boldsymbol{y}) \end{array}$$

Passo III: Aplicar a fórmula - pode-se dispensar os termos constantes

$$\begin{array}{lll} f(p|\boldsymbol{y}) & \propto & f(p)L(p|\boldsymbol{y}) \\ & \propto & \frac{1}{B(a,b)} \times p^{a-1}(1-p)^{b-1} I_{(0,1)}(p) \times p^n (1-p)^{\sum_{i=1}^{n} y_i} \\ & \propto & p^{a+n-1} \times (1-p)^{b+\sum_{i=1}^{n} y_i - 1} \times I_{(0,1)}(p) \\ & \text{Então} & p|\boldsymbol{y} \sim \text{Beta}(a+n, b+\sum_{i=1}^{n} y_i) \text{ é só incluir as constantes na fórmula:} \\ f(p|\boldsymbol{y}) & = & \frac{1}{B\left(a+n, b+\sum_{i=1}^{n} y_i\right)} \times p^{a+n} \times (1-p)^{b+\sum_{i=1}^{n} y_i - 1} \times I_{(0,1)}(p) \end{array}$$

Implementação numérica do Caso 6:

```r
set.seed(10102020) #cria uma semente única
p0=0.3 #este é o valor de p utilizado para gerar os dados
n=10 #tamanho da amostra
y=rgeom(n,prob=p0)
y
```

```
##  [1] 4 6 0 1 0 1 7 7 0 0
```

```r
logL=function(p){
L=p^n* (1-p)^sum(y)
log(L)
}
optimize(logL, interval=c(0.001, 0.999), maximum=TRUE) #encontra o ponto de máximo
```

```
## $maximum
## [1] 0.2777811
##
## $objective
## [1] -21.27032
```

```r
require(DT)
require(tidyverse)
require(ggpmisc)
```

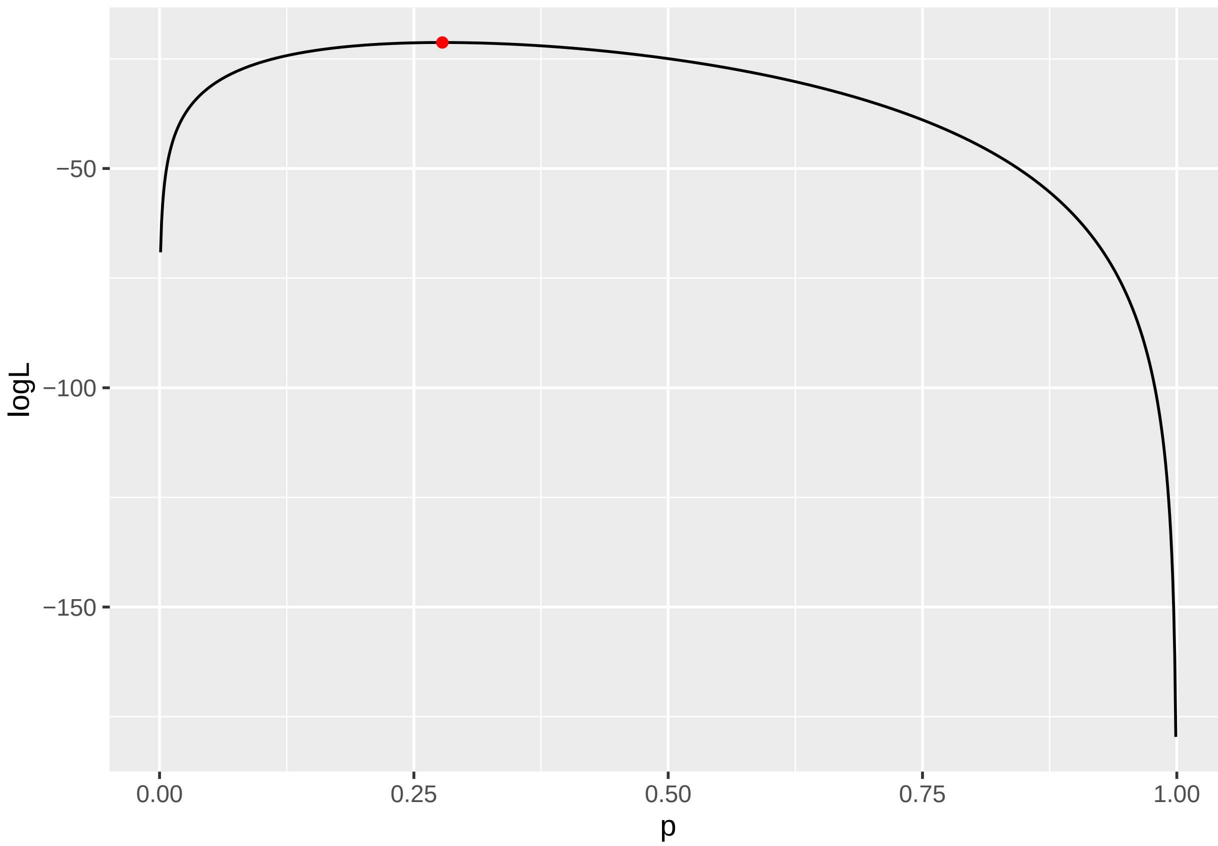
```r
p=seq(0.001,0.999,0.001)
temp <- data.frame(p, logL = logL(p))
datatable(temp)
```

Show 10 entries Search:

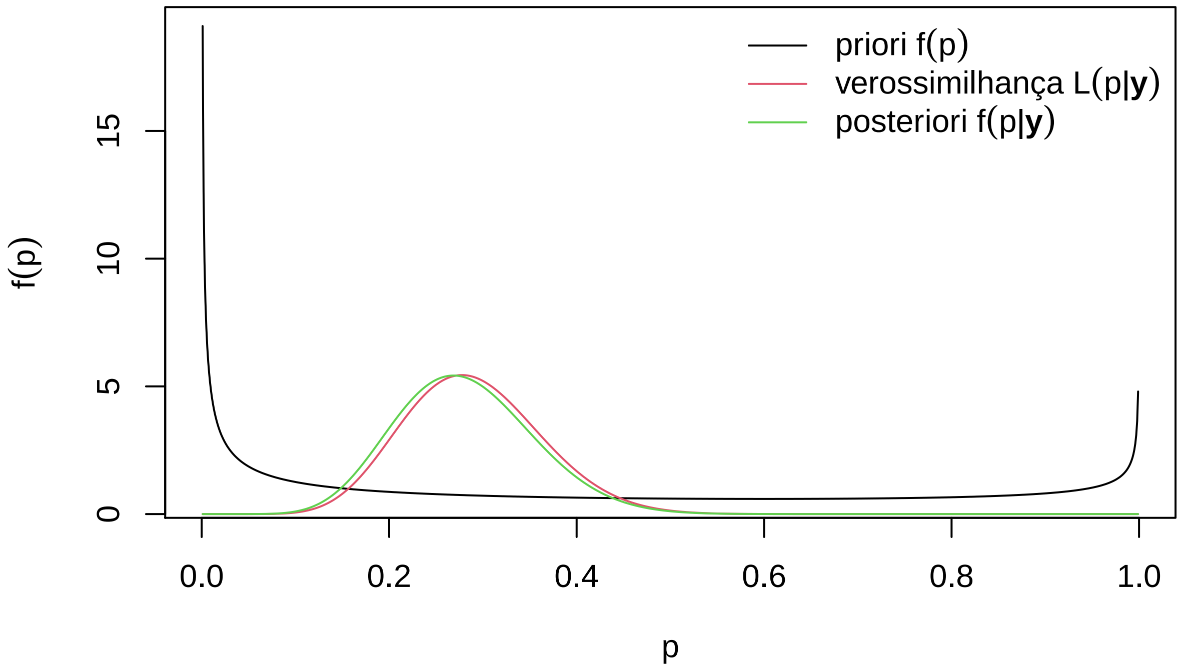
	p	logL
1	0.001	-69.1035657984945
2	0.002	-62.1981330536594
3	0.003	-58.169547137668
4	0.004	-55.3188177349585
5	0.005	-53.1134997528925
6	0.006	-51.3164279780055
7	0.007	-49.8010912876293
8	0.008	-48.4919738371519
9	0.009	-47.3403663774151
10	0.01	-46.3130105920719

Showing 1 to 10 of 999 entries Previous 1 2 3 4 5 … 100 Next

```r
ggplot(data = temp, aes(x = p, y = logL)) + geom_line() + stat_peaks(col = "red")
```




```
a=0.4
b=0.6
pri=dbeta(p,shape1=a,shape2=b)

#Aproxima a verossimilhança para o formato de uma densidade (com soma igual a 1, apenas para fins ilustrat
intervalos=0.001
vero=p^n* (1-p)^sum(y)
vero_normalizado=vero/sum(vero*intervalos)
pos=dbeta(p,shape1=a+n,shape2=b+sum(y))
plot(p,pri,type='l',xlab=expression(p),ylab=expression(f(p)))
lines(p,vero_normalizado,type='l',col=2)
lines(p,pos,type='l',col=3)
legend("topright",c(expression("priori "*f(p)),expression("verossimilhança "*L(p*"|"*bold(y))),expression(
```



Síntese de *prioris* conjugadas

Cenários	Distribuição a priori	Distribuição dos dados	Distribuição a posteriori
Caso 1	$\mu \sim \mathrm{N}(m_0, \sigma_0^2)$	$X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, com σ^2 conhecido	$\mu \mid \boldsymbol{x} \sim \mathrm{N}\left(\frac{\frac{m_0}{s_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{s_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{s_0^2} + \frac{n}{\sigma^2}}\right)$
Caso 2	$\lambda \sim$ Gamma (a, r)	$X_1, \ldots, X_n \sim$ Poisson (λ)	$\lambda \mid \boldsymbol{x} \sim \mathrm{Gamma}(a + n, r + \sum_{i=1}^n x_i)$
Caso 3	$p \sim \mathrm{Beta}(a, b)$	$X \sim \mathrm{Binomial}(n, p)$	$p \mid \boldsymbol{x} \sim \mathrm{Beta}(a + y, b + n - x)$
Caso 4	$\sigma^2 \sim$ Gamma Inv. (α, β)	$X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, com μ conhecido	$\sigma^2 \mid \boldsymbol{x} \sim \text{Gamma Inv.}(\alpha + \frac{n}{2}, \beta + \frac{n}{2} \sum_{i=1}^n (x$
Caso 5	$\tau \sim \mathrm{Gamma}(\lambda, r)$	$X_1, \ldots, X_n \sim \mathrm{N}(\mu, \frac{1}{\tau})$, com μ conhecido	$\tau \mid \boldsymbol{x} \sim \mathrm{Gamma}(\lambda + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2, r + \frac{n}{2}$
Caso 6	$p \sim \mathrm{Beta}(a, b)$	$X_1, \ldots, X_n \sim \mathrm{Geom}(p)$	$p \mid \boldsymbol{x} \sim \mathrm{Beta}(a + n, b + \sum_{i=1}^n x_i)$

Exercícios

- 1. Mostre que a família de distribuições Beta é conjugada em relação à binomial, geométrica e binomial negativa.

- 2. Para uma amostra aleatória $X_1, \ldots, X_n$ tomada da distribuição $U(0, \theta)$, mostre que a família de distribuições de Pareto com parâmetros a e b, cuja função de densidade é $f(\theta) = \frac{ab^a}{\theta^{a+1}}$, é conjugada à uniforme.

- 3. Suponha que o tempo, em minutos, para atendimento a clientes segue uma distribuição exponencial com parâmetro θ desconhecido. Com base na experiência anterior assume-se uma distribuição a priori Gamma com média 0.2 e desvio-padrão 1 para θ. Se o tempo médio para atender uma amostra aleatória de 20 clientes foi de 3.8 minutos, determine a distribuição a posteriori de θ.

- 4. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ. Determine os parâmetros da priori conjugada de θ sabendo que $\mathrm{E}(\theta) = 4$ e o coeficiente de variação *a priori* é igual a 0.5.
- 5. O número médio de defeitos por 100 metros de uma fita magnética é desconhecido e denotado por θ. Atribui-se uma distribuição a priori Gamma $(2, 10)$ para θ. Se um rolo de 1200 metros desta fita foi inspecionado e encontrou-se 4 defeitos, qual é a distribuição *a posteriori* de θ?

Princípio da Verossimilhança

Exemplo: Suponha que desejamos estimar θ, a probabilidade de observar cara (C) no lançamento de uma moeda e que, para um determinado experimento, observou-se:

$$\{C, \bar{C}, \bar{C}, C, C, \bar{C}, \bar{C}, \bar{C}, \bar{C}, C\}$$

Entre outras possibilidades, os dados acima podem ter sido gerados a partir dos seguintes experimentos:

- Seja X o número de caras em 10 lançamentos da moeda:

$$X \sim \text{ Binomial } (10, \theta), \text{ e os resultados poderiam ser: } X = 0, 1, 2, \ldots, 10.$$

- Seja Y o número de lançamentos da moeda até a obtenção de 4 caras:

$$Y \sim \text{ Binomial Negativa } (4, \theta), \text{ e os resultados poderiam ser: } Y = 4, 5, 6, \ldots$$

- Considerando os resultados do experimento no modelo Binomial:

$$P(X = 4 \mid \theta) = \left(\begin{array}{c} 10 \\ 4 \end{array} \right) \theta^4 (1 - \theta)^{10-4},$$

de modo que a função de verossimilhança será: $L(\theta \mid \text{ x}) \propto \theta^4 (1 - \theta)^6$;

- E no modelo Binomial Negativa:

$$P(Y = 10 \mid \theta) = \left(\begin{array}{c} 10 - 1 \\ 4 - 1 \end{array} \right) \theta^4 (1 - \theta)^{10-4},$$
de modo que a função de verossimilhança será: $L(\theta \mid \text{ y}) \propto \theta^4 (1 - \theta)^6$;

- X Sob a mesma *priori* para θ, a *posteriori* obtida a partir do modelo Binomial é igual à posteriori obtida a partir do modelo Binomial-Negativa.
- Porém, as estimativas de máxima verossimilhança sob cada um dos modelos são diferentes. **Tarefa:** justificar esta afirmação;
- **Formalmente:** Se temos dois vetores aleatórios pertencentes a um mesmo espaço amostral, que dependem do mesmo parâmetro θ e que possuem verossimilhancas distintas, diferindo apenas por uma constante que não depende de θ, Então as *posterioris* obtidas a partir destes dois vetores são iguais.
- Em outras palavras, a **inferência Bayesiana** é a mesma quando a condição de proporcionalidade das verossimilhanças é satisfeita.

prioris **não informativas**

- As *prioris* não informativas estão presentes quando se espera que a informação dos dados seja dominante, significa que a informação *a priori* é vaga, então temos o conceito de "conhecimento vago", "não informação" ou "ignorância *a priori*".
- Referências sobre *prioris* não informativas estão em [**?**], [**?**] e [**?**].

priori Uniforme

É uma *priori* intuitiva porque todos os possíveis valores do parâmetro θ são igualmente prováveis:

$$f(\theta) \propto k,$$

com θ variando em um subconjunto da reta de modo que nenhum valor particular tem preferência (Bayes, 1763). A *priori* uniforme, no entanto, apresenta algumas dificuldades:

- Se o intervalo de variação de θ for a reta real então a distribuição é imprópria:

$$\int_{-\infty}^{\infty} f(\theta) d\theta = \infty,$$

  mas este não chega a ser um impedimento para a escolha de *prioris*, como veremos mais adiante.
- Se $\phi = g(\theta)$ é uma reparametrização não linear monótona de θ então a *priori* para o parâmetro ϕ será:

$$f(\phi) = f(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right| \propto \left| \frac{d\theta}{d\phi} \right|,$$

  e vemos pelo **Teorema de transformação de variáveis** que a *priori* para ϕ não é uniforme.

priori de Jeffreys

É uma *priori* construída a partir da **medida de informação esperada de Fisher**, proposta por Jeffreys (1961).

- é uma *priori* **imprópria**;
- é **invariante** a transformações 1 a 1.

Definição: Medida de informação esperada de Fisher Considere uma única observação X com f.d.p. indexada pelo parâmetro θ: $f(x|\theta)$. A medida de informação esperada de Fisher de θ através de X é definida como

$$I(\theta) = E\left[-\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right],$$

em que a esperança matemática é tomada em relação à distribuição amostral $f(x|\theta)$ (a esperança é com respeito a X e não com respeito a θ). - A informação esperada de Fisher $I(\theta)$ é uma **medida de informação global**.

Extendendo esta definição para uma amostra i.i.d. $X_1, X_2, \ldots, X_n$, temos: $f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ e

$$I(\theta) = E\left[-\frac{\partial^2 \log f(\boldsymbol{x}|\theta)}{\partial \theta^2}\right],$$

é a informação esperada de Fisher de θ através do vetor $\boldsymbol{x}$.

Definição: *priori* de Jeffreys A *priori* de Jeffreys é dada por:

$$\sqrt{I(\theta)}.$$

No caso multiparamétrico (mais de um parâmetro), a medida de Informação de Fisher é dada de forma matricial, então temos:

$$\sqrt{|\det\left[I(\boldsymbol{\theta})\right]|}.$$

Exemplo: Sejam $X_1, \ldots, X_n \sim$ Poisson(θ).

$$\begin{array}{lcl}
\log f(\boldsymbol{x}|\theta) & = & -n\theta + \sum_{i=1}^{n} x_i \log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right) \\
\frac{\partial \log f(\boldsymbol{x}|\theta)}{\partial \theta} & = & -n + \frac{\sum_{i=1}^{n} x_i}{\theta} \\
\frac{\partial^2 \log f(\boldsymbol{x}|\theta)}{\partial \theta^2} & = & -\frac{\sum_{i=1}^{n} x_i}{\theta^2} \\
I(\theta) & = & \frac{n}{\theta} \\
 & \propto & \frac{1}{\theta}
\end{array}$$

- A *priori* de Jeffreys para θ no modelo Poisson é $f(\theta) \propto \theta^{-1/2}$;
- Esta priori também pode ser obtida tomando-se a *priori* conjugada Gamma(α, β) com $\alpha = \frac{1}{2}$ e $\beta \to 0$. Note que o parâmetro β é sempre positivo, por isso a noção de "tender a zero". **Tarefa: verificar**;
- Em geral, quando o modelo admite *priori* conjugada, basta fixar um dos parâmetros da *priori* conjugada e o outro parâmetro "tender a zero", resultando na *priori* de Jeffreys;
- A *priori* de Jeffreys não satisfaz o **princípio da verossimilhança**, pois a informação esperada de Fisher depende da distribuição amostral (o cálculo das esperanças matemáticas podem ser diferentes se os modelos forem diferentes como no exemplo **??** modelos Binomial e Binomial-Negativa).
- A *priori* de Jeffreys apresenta algumas particularidades nos modelos de locação-escala, como veremos a seguir.

Modelos de locação-escala

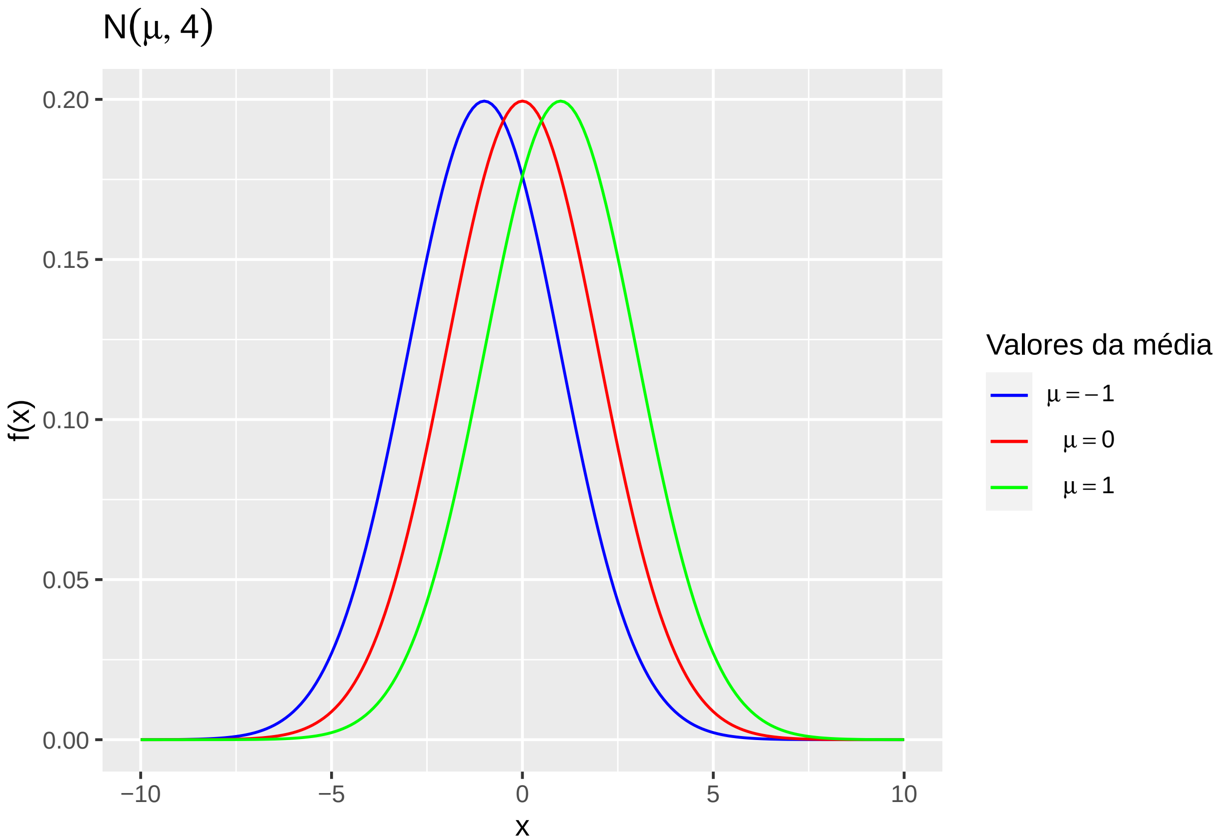
Modelo de Locação X tem um modelo de locação se existem uma função g e uma quantidade μ tais que:

$$f(x|\mu) = g(x - \mu),$$

logo μ é o parâmetro de locação.

- A definição se extende para o caso multiparamétrico;
- **Exemplo:** distribuição normal com variância conhecida é um modelo de locação:

```
mu=c(-1,0,1)  #cria um vetor de médias de -1, 0 e 1
sigma=2 #fixa o desvio padrao em 2, ou seja, a variancia é igual a 4
x=seq(-10,10,0.1)
y_1=dnorm(x,mean=mu[1],sd=sigma)
y_2=dnorm(x,mean=mu[2],sd=sigma)
y_3=dnorm(x,mean=mu[3],sd=sigma)
dados=data.frame(x,y_1,y_2,y_3)

colors <- c("y_1"="blue", "y_2"="red", "y_3"="green")

ggplot(dados, aes(x = x,y = y_1, color = "y_1")) +
    geom_line() +
    geom_line(aes(x = x,y = y_2, color = "y_2")) +
    geom_line(aes(x = x,y = y_3, color = "y_3")) +
    labs(x="x",
         y="f(x)",
         title=expression(N(mu,4)),
         color = "Valores da média") +
    scale_color_manual(labels=c(expression(mu==-1),expression(mu==0),expression(mu==1)),values = colors)
```



- È possivel calcular valores de uma normal não central a partir da normal centrada no zero

```r
x=seq(4,6,.5)
f=  dnorm(x,mean=5,sd=2)
g=  dnorm(x-5,mean=0,sd=2)
temp=data.frame(x,f,g)

require(kableExtra)
names(temp) = c("$x$","$f(x)$","$g(x-5$)")
temp %>%
   kbl(escape = FALSE) %>%
 kable_classic(full_width = F, html_font = "Cambria") %>%
kable_styling(bootstrap_options = c("striped","hold_position")) %>%
 add_header_above(c("","$f(.)$ vem da N(5,4)","$g(.)$ vem da N(0,4)"))
```

	$f(.)$ vem da N(5,4)	$g(.)$ vem da N(0,4)
x	$f(x)$	$g(x-5)$
4.0	0.1760327	0.1760327
4.5	0.1933341	0.1933341
5.0	0.1994711	0.1994711
5.5	0.1933341	0.1933341
6.0	0.1760327	0.1760327

- **Propriedade:** A *priori* de Jeffreys para o parâmetro de locação μ é:

$$f(\mu) \propto k,$$

onde k é uma constante.

Modelo de Escala X tem um modelo de escala se existem uma função g e uma quantidade σ tais que:

$$f(x|\sigma) = \frac{1}{\sigma} g\left(\frac{x}{\sigma}\right),$$

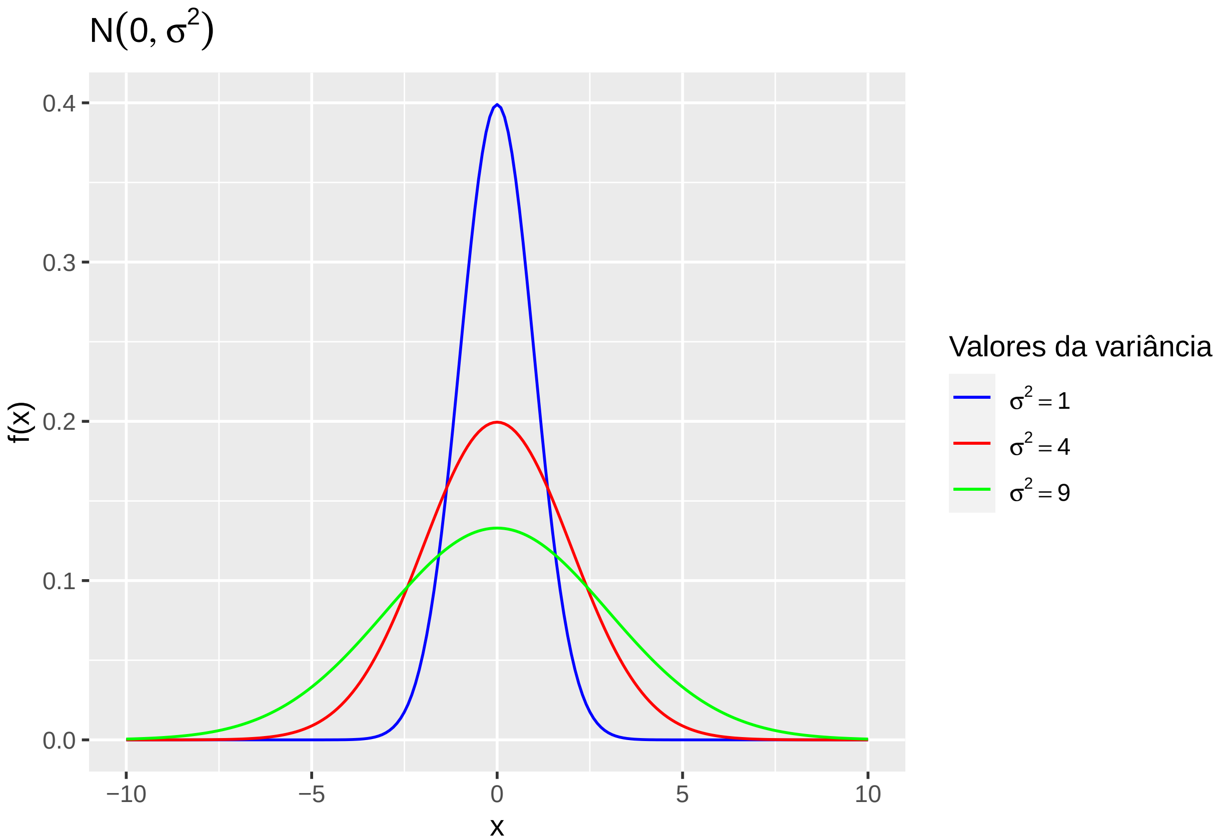
logo σ é o parâmetro de escala.

- **Exemplos:** Na distribuição Exp(θ) o parâmetro de escala é $\sigma = \frac{1}{\theta}$, e na distribuição N($\mu, \sigma^2$) com média conhecida o parâmetro de escala é σ;
- **Propriedade:** A *priori* de Jeffreys para o parâmetro de escala σ é:

$$f(\sigma) \propto \frac{1}{\sigma}.$$

- Mostrando que o Modelo normal com media conhecida é modelo de escala

```
sigma=c(1,2,3)  #cria um vetor de desvios padroes de 1, 2 e 3
media=0 #fixa a média em zero
x=seq(-10,10,0.1)
y_1=dnorm(x,mean=media,sd=sigma[1])
y_2=dnorm(x,mean=media,sd=sigma[2])
y_3=dnorm(x,mean=media,sd=sigma[3])
dados=data.frame(x,y_1,y_2,y_3)

colors <- c("y_1"="blue", "y_2"="red", "y_3"="green")

ggplot(dados, aes(x = x,y = y_1, color = "y_1")) +
    geom_line() +
    geom_line(aes(x = x,y = y_2, color = "y_2")) +
    geom_line(aes(x = x,y = y_3, color = "y_3")) +
    labs(x="x",
         y="f(x)",
         title=expression(N(0,sigma^2)),
         color = "Valores da variância") +
    scale_color_manual(labels=c(expression(sigma^2==1),expression(sigma^2==4),expression(sigma^2==9)),valu
```



- È possivel calcular valores de uma normal com desvio padrão diferente de 1 a partir de uma normal com desvio padrão igual a 1:

```
x=seq(4,6,.5)
mu=5
sigma=2
f=dnorm(x,mean=5,sd=sigma)
g=1/sigma*dnorm(x/sigma,mean=mu/sigma,sd=1)

temp=data.frame(x,f,g)
names(temp) = c("$x$","$f(x)$","$\\frac{1}{2}g(\\frac{x}{2})$")
temp %>%
   kbl(escape = FALSE) %>%
 kable_classic(full_width = F, html_font = "Cambria") %>%
kable_styling(bootstrap_options = c("striped","hold_position")) %>%
 add_header_above(c("","$f(.)$ vem da N(5,4)","$g(.)$ vem da N(5,1)"))
```

	$f(.)$ vem da N(5,4)	$g(.)$ vem da N(5,1)
x	$f(x)$	$\frac{1}{2}g(\frac{x}{2})$
4.0	0.1760327	0.1760327
4.5	0.1933341	0.1933341
5.0	0.1994711	0.1994711
5.5	0.1933341	0.1933341
6.0	0.1760327	0.1760327

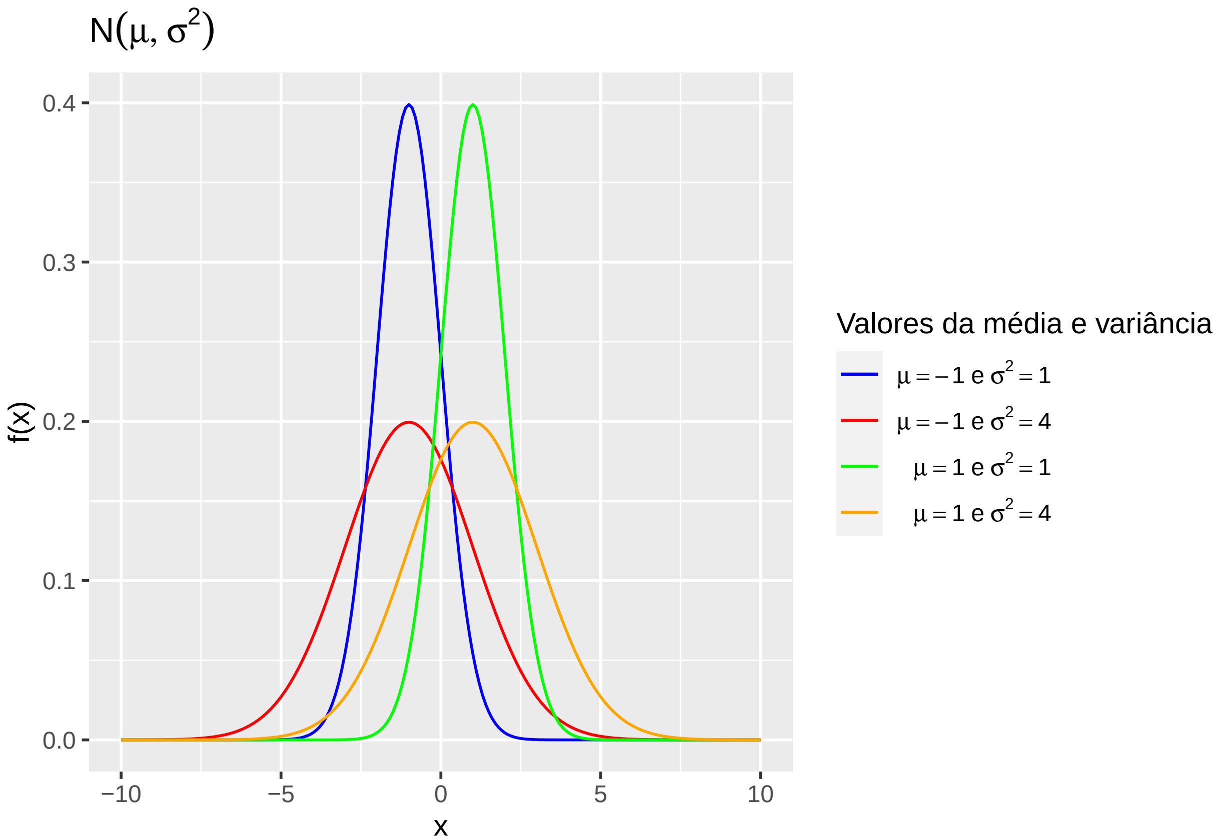
Definição: Modelo de Locação-escala X tem um modelo de locação-escala se existem uma função g e as quantidades μ e σ tais que

$$f(x|\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right),$$

logo μ é o parâmetro de locação e σ é o parâmetro de escala.

- **Exemplos:** Na distribuição $\mathrm{N}(\mu,\sigma^2)$ o parâmetro de locação é μ e o parâmetro de escala é σ, e a distribuição de Cauchy também é um modelo de locação-escala.

```
    #é modelo de locação-escala
media=c(-1,1) #a média assume os valores -1 ou 1
sigma=c(1,2)  #o desvio padrão assume os valores 1 ou 2
x=seq(-10,10,0.1)
y_1= dnorm(x,media[1],sigma[1])
y_2= dnorm(x,media[1],sigma[2])
y_3= dnorm(x,media[2],sigma[1])
y_4= dnorm(x,media[2],sigma[2])
dados=data.frame(x,y_1,y_2,y_3,y_4)

colors <- c("y_1"="blue", "y_2"="red", "y_3"="green","y_4"="orange")

ggplot(dados, aes(x = x,y = y_1, color = "y_1")) +
    geom_line() +
    geom_line(aes(x = x,y = y_2, color = "y_2")) +
    geom_line(aes(x = x,y = y_3, color = "y_3")) +
    geom_line(aes(x = x,y = y_4, color = "y_4")) +
    labs(x="x",
         y="f(x)",
         title=expression(N(mu,sigma^2)),
         color = "Valores da média e variância") +
    scale_color_manual(labels=c(
      expression(mu==-1~"e"~sigma^2==1),
      expression(mu==-1~"e"~sigma^2==4),
      expression(mu==1~"e"~sigma^2==1),
      expression(mu==1~"e"~sigma^2==4)),
      values = colors)
```



- È possivel calcular valores de uma normal genérica a partir de uma normal padrão:

```
x=seq(4,6,.5)
mu=5
sigma=2
f=dnorm(x,mean=5,sd=sigma)
g=1/sigma*dnorm((x-mu)/sigma,mean=0,sd=1)

temp=data.frame(x,f,g)
names(temp) = c("$x$","$f(x)$","$\\frac{1}{2}g(\\frac{x-5}{2})$")
temp %>%
   kbl(escape = FALSE) %>%
 kable_classic(full_width = F, html_font = "Cambria") %>%
kable_styling(bootstrap_options = c("striped","hold_position")) %>%
 add_header_above(c("","$f(.)$ vem da N(5,4)","$g(.)$ vem da N(0,1)"))
```

	\$f(.)\$ vem da N(5,4)	\$g(.)\$ vem da N(0,1)
x	$f(x)$	$\frac{1}{2}g(\frac{x-5}{2})$
4.0	0.1760327	0.1760327
4.5	0.1933341	0.1933341
5.0	0.1994711	0.1994711
5.5	0.1933341	0.1933341
6.0	0.1760327	0.1760327

- **Propriedade** A *priori* conjunta de Jeffreys para os parâmetros de locação μ e escala σ é:

$$f(\mu,\sigma)=f(\mu)f(\sigma)\propto\frac{1}{\sigma},$$

  onde nós assumimos independência *a priori* (a *priori* conjunta é o produto das *prioris*).

Exemplo: Sejam $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ com μ e σ^2 desconhecidos, temos:

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

logo μ é o parâmetro de locação e σ é o parâmetro de escala.

- A *priori* não informativa de Jeffreys para o vetor (μ,σ) é:

$$f(\mu,\sigma)\propto\frac{1}{\sigma}$$

- Pela **propriedade da invariância**, a *priori* não informativa de Jeffreys para o vetor (μ,σ^2) é:

$$f(\mu,\sigma^2)\propto\frac{1}{\sigma^2}$$

Síntese de *prioris* de Jeffreys

```
tabela <- data.frame(
      col0 =  c("Cenários","Caso 1","Caso 2","Caso 3","Caso 4","Caso 5","Caso 6"),
      col1 =  c("priori de Jeffreys (é não informativa)","$\\mu \\sim$ N$(m_0,\\sigma_0^2$)","$\\lambda \\
      col2 =  c("Distribuição dos dados","$X_1,\\ldots,X_n \\sim$ N$(\\mu,\\sigma^2)$, com $\\sigma^2$ con
      col3 =  c("Distribuição dos dados","$X_1,\\ldots,X_n \\sim$ N$(\\mu,\\sigma^2)$, com $\\sigma^2$ con


 tabela %>%
  kbl(col.names = NULL,escape = FALSE) %>%
  kable_classic(full_width = F, html_font = "Cambria") %>%
     kable_styling(bootstrap_options = c("striped","hold_position"),
                full_width = F)
```

Cenários	priori de Jeffreys (é não informativa)	Distribuição dos dados	Distribuição dos dados
Caso 1	$\mu \sim$ N(m_0,σ_0^2)	$X_1,\ldots,X_n \sim$ N(μ,σ^2), com σ^2 conhecido	$X_1,\ldots,X_n \sim$ N(μ,σ^2), com σ
Caso 2	$\lambda \sim$ Gamma (a,r)	$X_1,\ldots,X_n \sim$ Poisson (λ)	$X_1,\ldots,X_n \sim$ Poisson (λ)
Caso 3	$p \sim$ Beta(a,b)	$X \sim$ Binomial(n,p)	$X \sim$ Binomial(n,p)
Caso 4	$\sigma^2 \sim$ Gamma Inv. (α,β)	$X_1,\ldots,X_n \sim$ N(μ,σ^2), com μ conhecido	$X_1,\ldots,X_n \sim$ N(μ,σ^2), com μ
Caso 5	$\tau \sim$ Gamma(λ,r)	$X_1,\ldots,X_n \sim$ N$(\mu,\frac{1}{\tau})$, com μ conhecido	$X_1,\ldots,X_n \sim$ N$(\mu,\frac{1}{\tau})$, com μ
Caso 6	$p \sim$ Beta(a,b)	$X_1,\ldots,X_n \sim$ Geom(p)	$X_1,\ldots,X_n \sim$ Geom(p)

Exercícios

- 1. Considerando o modelo normal média conhecida e variância desconhecida:
  - a) Mostre que este modelo é de escala, sendo o desvio padrão o parâmetro de escala;

- b) Mostre que a *priori* de Jeffreys para a o desvio padrão σ é $f(\sigma) \propto \frac{1}{\sigma}$. Primeiro encontre pela informação esperada de Fisher, depois verifique se satisfaz a propriedade dos modelos de locação-escala.

- 2. Para cada uma das distribuições abaixo verifique se o modelo é de locação, escala ou locação-escala e obtenha a *priori* não informativa de Jeffreys para os parâmetros desconhecidos.
  - a) Cauchy$(0, \beta)$;
  - b) $t_\nu(\mu, \sigma^2)$, com ν conhecido;
  - c) Pareto(a, b), com b conhecido;
  - d) Uniforme$(\theta - 1, \theta + 1)$;
  - e) Uniforme$(-\theta, \theta)$.
- 3. Mostre que a dist. Cauchy é um modelo de locação-escala onde α é o parâmetro de locação e β é o parametro de escala.
- 4. Mostrar que a *priori* de Jeffreys no modelo Normal com variancia conhecida é dada por uma constante, como diz a fórmula COLOCAR.

Unidade III - Inferência Bayesiana

Exemplo 3.1: Regressão linear simples

O problema envolve as variáveis X: a dose de um medicamento anti-alérgico em estudo, e Y: tempo de duração do efeito (alívio dos sintomas alérgicos). Abaixo temos a representação gráfica dos dados observados Pelo gráfico, nós concluímos que uma relação linear (reta) é satisfatória para os dados. Também iremos supor que X é uma variável controlada pelo pesquisador (sem a presença de erros).

```
x=c(3,3,4,5,6,6,7,8,8,9)
y=c(9,5,12,9,14,16,22,18,24,22)
a=cbind(x,y)
tab_nums <- captioner(prefix = "Tabela")
tab_cap = tab_nums("tabela1","Dados do Exemplo")
datatable(as.data.frame(a),caption=tab_cap)
```

Show 10 entries

Search:

Tabela 1: Dados do Exemplo

	x	y
1	3	9
2	3	5
3	4	12
4	5	9
5	6	14
6	6	16
7	7	22
8	8	18
9	8	24
10	9	22

Showing 1 to 10 of 10 entries

Previous 1 Next

```
plot(x,y,lwd=2,xlim=c(0,10),ylim=c(0,25))
```

Modelo Estatístico O modelo estatístico será o modelo de regressão simples com erros i.i.d. normais:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n,$$

onde β_0: intercepto da linha de regressão com o eixo y; β_1: coeficiente de inclinação da reta: é o nº de unidades em y que mudam para cada unidade da variável independente x. ϵ_i: erros aleatórios com distribuição normal: $\epsilon_i \sim N(0, \sigma^2)$.

Estimadores de mínimos quadrados da regressão Encontrar $\widehat{\beta_0}$ e $\widehat{\beta_1}$ que minimizam a soma de quadrados dos erros: $S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$. Então temos as **equações normais:**

$$\frac{\partial S}{\partial \beta_0} = 0 \text{ e } \frac{\partial S}{\partial \beta_1} = 0.$$

A solução é dada por:

$$\begin{aligned}
\widehat{\beta_1} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
\widehat{\beta_0} &= \bar{y} - \widehat{\beta_1}\bar{x}.
\end{aligned}$$

Com respeito a variância σ_2:

$$\widehat{\sigma_2} = \frac{SQR}{n-2},$$

ou seja, a estimativa da variância é igual à soma dos quadrados dos resíduos sobre o número graus de liberdade do modelo. Os intervalos de confiança e testes de hipóteses para β_0 e β_1 são baseados na distribuição t-student.

Voltando ao R: - Método dos mínimos quadrados: cálculo das estimativas passo a passo

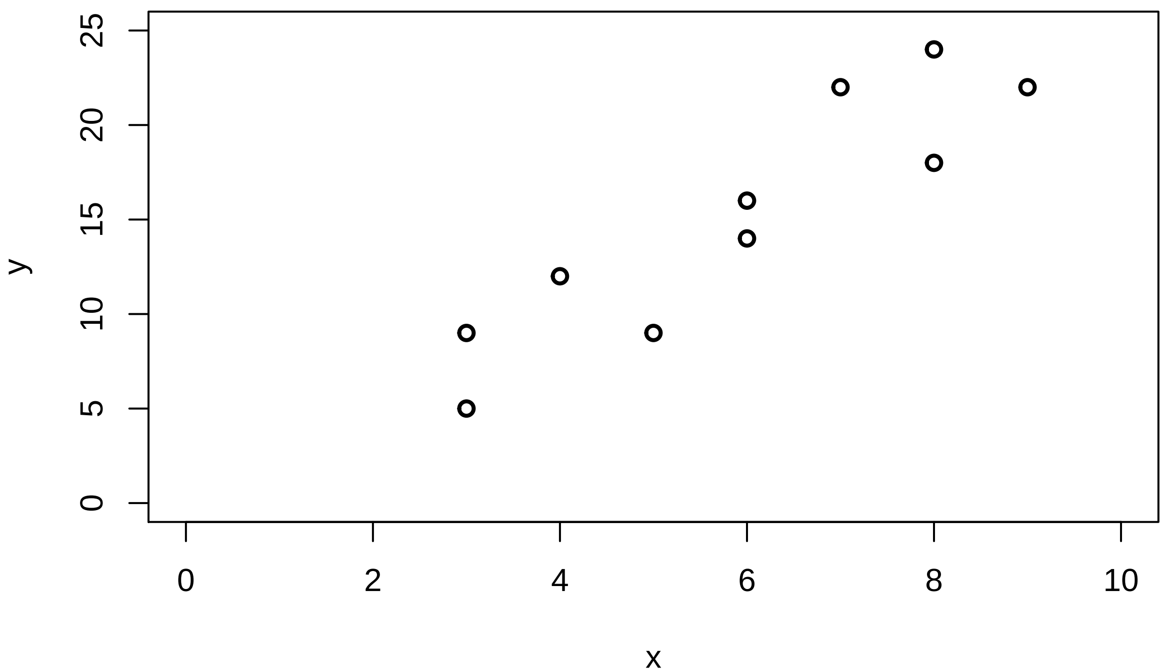


Figure 7: Figura 7: Diagrama de dispersão dos dados

```r
soma_xx=sum((x-mean(x))^2)
soma_xy=sum((x-mean(x))*(y-mean(y)))
beta1_hat=soma_xy/soma_xx
beta0_hat=mean(y)-beta1_hat*mean(x)
print(paste0("beta0_hat=",beta0_hat," & beta1_hat=",beta1_hat))
```

```
## [1] "beta0_hat=-1.07090464547677 & beta1_hat=2.74083129584352"
```

- Método dos mínimos quadrados: utilizando a função lm:linear model
  - Resíduos aproximadamente normais com variância constante.

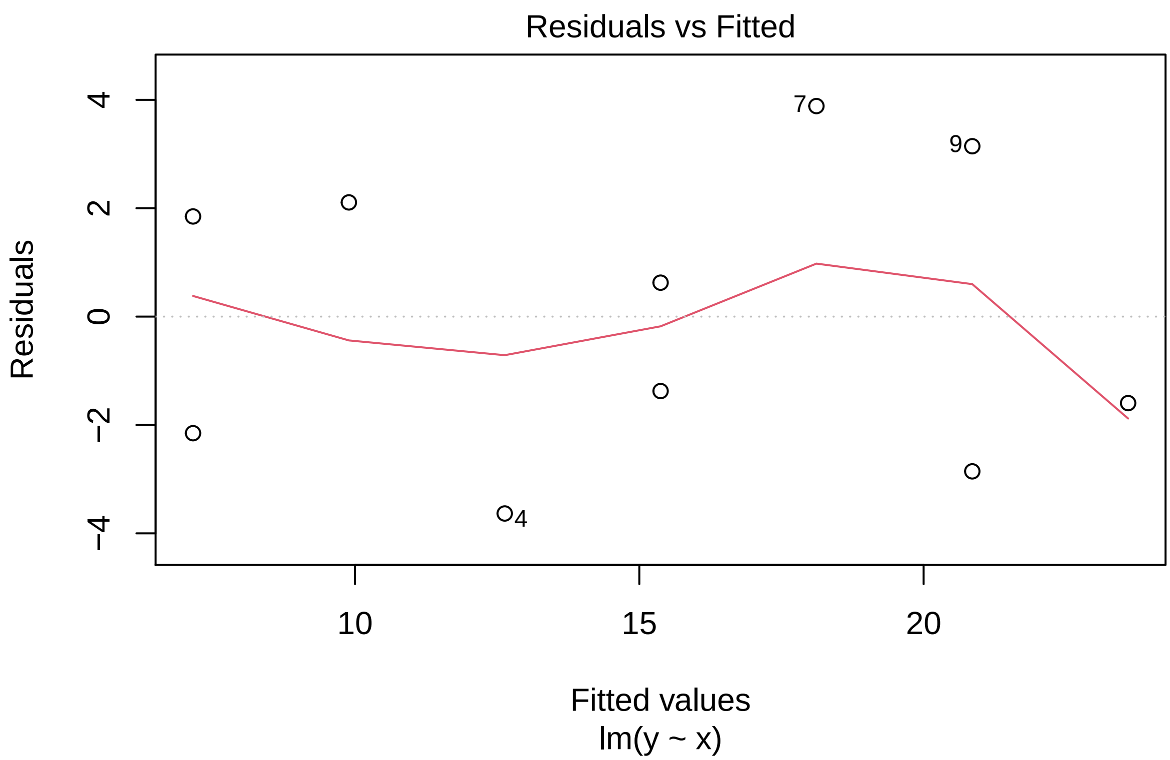
```r
a=lm(y ~ x)
summary(a)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6333 -2.0128 -0.3741  2.0428  3.8851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0709     2.7509  -0.389 0.707219
## x             2.7408     0.4411   6.214 0.000255 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.821 on 8 degrees of freedom
## Multiple R-squared:  0.8284, Adjusted R-squared:  0.8069
## F-statistic: 38.62 on 1 and 8 DF,  p-value: 0.0002555
```

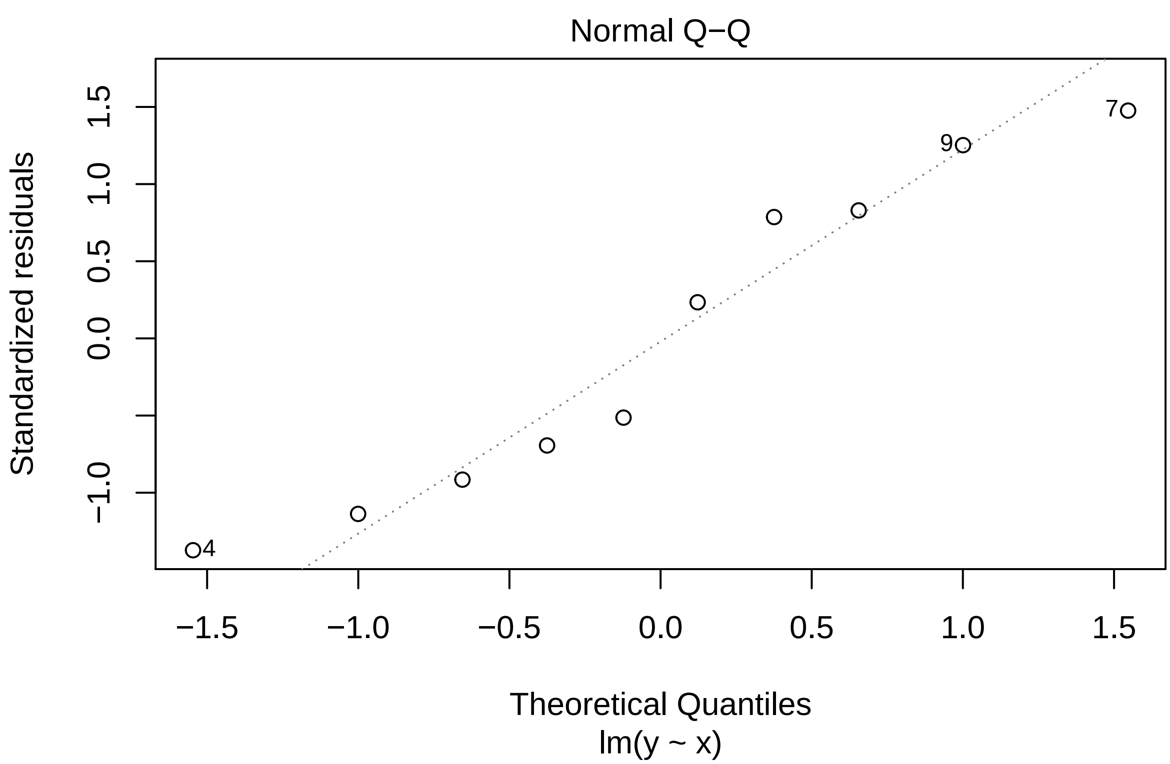
```r
anova(a)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 307.247 307.247  38.615 0.0002555 ***
## Residuals  8  63.653   7.957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

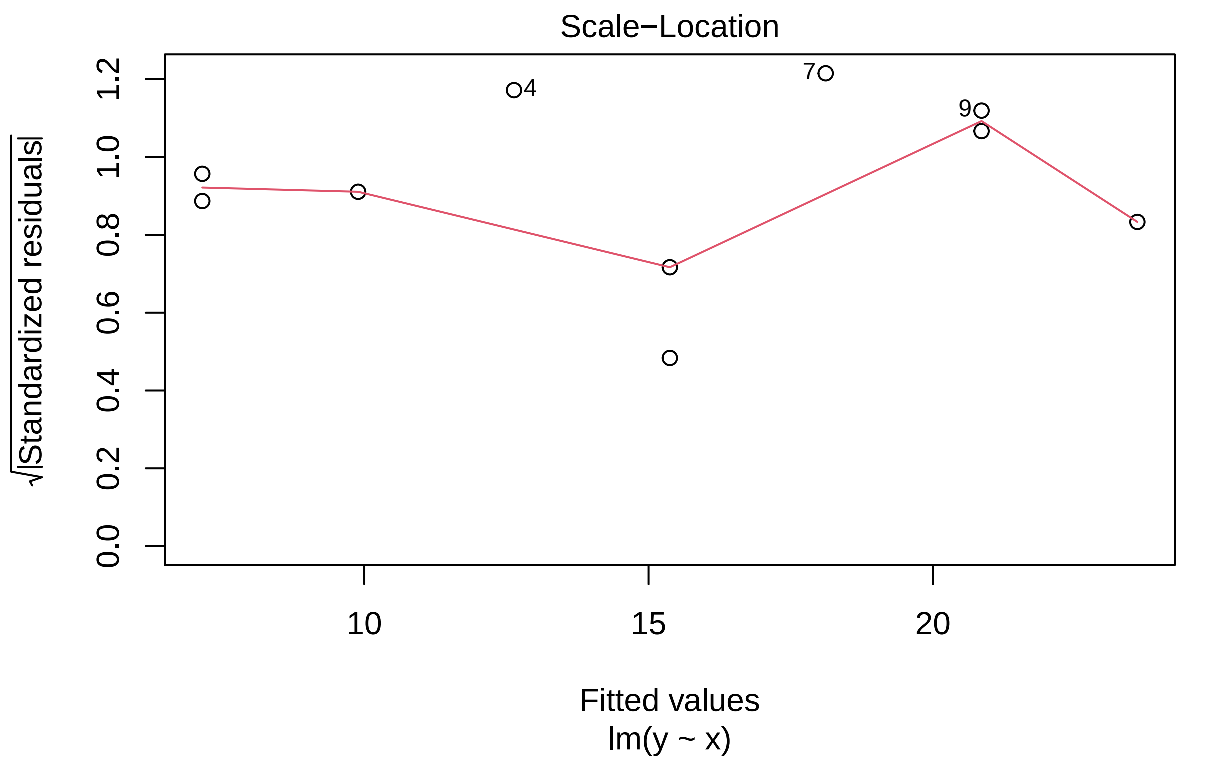
```r
plot(a)
```



Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x)
4
2
0
−2
−4
10
15
20
7
9
4



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
1.5
1.0
0.5
0.0
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
7
9
4


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
10
15
20
4
7
9
Fitted values
lm(y ~ x)


Residuals vs Leverage
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30
9
2
8
0.5
0.5
Cook's distance
Leverage
lm(y ~ x)

```r
shapiro.test(residuals(a))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(a)
## W = 0.93227, p-value = 0.4706
```

Inferência Bayesiana no modelo de regressão linear simples

Assumimos as seguintes distribuições *a priori*: $\beta_0 \sim N(0, a_0^2)$ com a_0 conhecido $\beta_0 \sim N(0, a_1^2)$ com a_1 conhecido $\sigma_2 \sim IG(b, d)$ com b e d conhecidos Iremos assumir independência *a priori* entre os parâmetros.

- Função de verossimilhança:

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \tfrac{1}{2\pi\sigma^2} \exp\left[-\tfrac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\ &= (\sigma^2)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right] \end{aligned}$$

- Distribuição *a posteriori* conjunta para β_0,β_1 e σ^2 é dada por:

$$\begin{aligned} f(\beta_0, \beta_1, \sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) &\propto \exp\left[-\tfrac{\beta_0^2}{2a_0^2}\right] \exp\left[-\tfrac{\beta_1^2}{2a_1^2}\right] (\sigma^2)^{-(b+1)} \exp\left[-\tfrac{d}{\sigma^2}\right] (\sigma^2)^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\ &\propto \exp\left[-\tfrac{\beta_0^2}{2a_0^2}\right] \exp\left[-\tfrac{\beta_1^2}{2a_1^2}\right] (\sigma^2)^{-(b+\frac{n}{2}+1)} \exp\left[-\tfrac{d}{\sigma^2}\right] \exp\left[-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right] \end{aligned}$$

- Distribuições *a posteriori* marginais - é necessário integrar com respeito aos outros parâmetros, respeitando os limites de integração: ✓ A conjunta de β_0 e β_1 é obtida da integração com respeito à variância σ^2:

$$f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) = \int_{0}^{\infty} f(\beta_0, \beta_1, \sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) d\sigma^2$$

✓ A marginal de σ^2 é obtida da integração com respeito à β_0 e β_1:

$$f(\sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\beta_0, \beta_1, \sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) d\beta_0 d\beta_1$$

✓ A marginal de β_0 é obtida da integração com respeito à β_1:

$$f(\beta_0 \mid \boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{\infty} f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) d\beta_1$$

✓ A marginal de β_1 é obtida da integração com respeito à β_0:

$$f(\beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) = \int_{-\infty}^{\infty} f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) d\beta_0$$

✓ A conjunta de β_0 e β_1 pode ser obtida analiticamente:

$$\begin{aligned} f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) &\propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \int\limits_0^\infty (\sigma^2)^{-(b+\frac{n}{2}+1)} \exp\left[-\frac{1}{\sigma^2}\left(d + \frac{1}{2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right)\right] d\sigma^2 \\ &\propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \left[d + \frac{1}{2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right]^{-\left(b+\frac{n}{2}\right)}. \end{aligned}$$

Tarefa: Provar este resultado. Dica: envolve a integral da distribuição Gamma Invertida $\left(b + \frac{n}{2}, d + \frac{1}{2}\sum\limits_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right)$ e o fato de que a integral de uma função de densidade sempre é igual a 1. Observe que mesmo tendo obtido a integral, ela não tem forma conhecida - não conseguimos identificar esta na tábua de distribuições com suporte de $-\infty$ a ∞.

Já as outras marginais não têm forma fechada - não são obtidas analiticamente - integrais analíticas não são possíveis.

Devido à este inconveniente com respeito às integrais, nós recorremos às distribuições *a posteriori* condicionais. Este tópico está relacionado com métodos MCMC (método de Monte Carlo com cadeias de Markov) e o amostrador de GIBBS que veremos mais adiante. As distribuições As *posterioris} condicionais são facilmente obtidas: ✓ A condicional de σ^2 dado β_0, β_1 e os dados:

$$\begin{aligned} & f(\sigma^2 \mid \beta_0, \beta_1, \boldsymbol{x}, \boldsymbol{y}) \propto (\sigma^2)^{-(b+\frac{n}{2}+1)} \exp\left[-\frac{1}{\sigma^2}\left(d + \frac{1}{2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right)\right], \\ \text{ou seja,} \qquad & \sigma^2 \mid \beta_0, \beta_1, \boldsymbol{x}, \boldsymbol{y} \sim IG\left(b + \frac{n}{2}, d + \frac{1}{2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right). \end{aligned}$$

a idéia de condicional nos diz que β_0 e β_1 são tratadas como constantes com respeito a σ^2.

✓ A condicional de β_0 dado β_1, σ^2 e os dados:

$$f(\beta_0 \mid \beta_1, \sigma^2, \boldsymbol{x}, \boldsymbol{y}) \propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

Ideia: expandir os termos e identificar uma distribuição normal! Tome $\mu_i^{(1)} = y_i - \beta_1 x_i$

$$\propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\beta_0 - \mu_i^{(1)})^2\right]$$

$$\propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{1}{2\sigma^2}\left(n\beta_0^2 + 2\beta_0\sum_{i=1}^{n}\mu_i^{(1)} + \sum_{i=1}^{n}\mu_i^{(1)^2}\right)\right]$$

$$\propto \exp\left[-\frac{\beta_0^2}{2}\left(\frac{1}{a_0^2} + \frac{n}{\sigma^2}\right) + \frac{\beta_0\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2}\right], \text{ o termo } \sum_{i=1}^{n}\mu_i^{(1)^2} \text{ "caiu" pois é constante com respeito a } \beta_0$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{1}{a_0^2}+\frac{n}{\sigma^2}\right)^{-1}}\left(\beta_0^2 - \frac{2\beta_0\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2\left(\frac{1}{a_0^2}+\frac{n}{\sigma^2}\right)}\right)\right]. \text{ Mas } \left(\frac{1}{a_0^2} + \frac{n}{\sigma^2}\right)^{-1} = \left(\frac{\sigma^2+a_0^2 n}{a_0^2\sigma^2}\right)^{-1} = \left(\frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)}\left(\beta_0^2 - \frac{2\beta_0\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2\left(\frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)}\right)\right], \text{ e simplificando um pouco mais:}$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)}\left(\beta_0^2 - \frac{2\beta_0 a_0^2\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2+a_0^2 n}\right)\right] \text{ e completando quadrados:}$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)}\left(\beta_0^2 - \frac{2\beta_0 a_0^2\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2+a_0^2 n} + \left(\frac{a_0^2\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2+a_0^2 n}\right)^2\right)\right]$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)}\left(\beta_0 - \frac{a_0^2\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2+a_0^2 n}\right)^2\right]$$

ou seja, $\beta_0 \mid \beta_1, \sigma^2, \boldsymbol{x}, \boldsymbol{y} \sim N\left(\frac{a_0^2\sum_{i=1}^{n}\mu_i^{(1)}}{\sigma^2+a_0^2 n}, \frac{a_0^2\sigma^2}{\sigma^2+a_0^2 n}\right)$

✓ A condicional de β_1 dado β_0, σ^2 e os dados: Note que o parâmetro β_1 acompanha o termo x_i:

$$f(\beta_1 \mid \beta_0, \sigma^2, \boldsymbol{x}, \boldsymbol{y}) \propto \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]. \text{ Tome } \mu_i^{(2)} = y_i - \beta_0$$

$$\propto \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\beta_1 x_i - \mu_i^{(2)})^2\right]$$

$$\propto \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\beta_1^2 x_i^2 - 2\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i + \sum_{i=1}^{n}\mu_i^{(2)^2}\right)\right]$$

$$\propto \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\beta_1^2 x_i^2 - 2\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i\right)\right]$$

$$\propto \exp\left[-\frac{\beta_1^2}{2}\left(\frac{1}{a_1^2} + \frac{\sum_{i=1}^{n}x_i^2}{\sigma^2}\right) + \frac{\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2}\right]$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{1}{a_1^2} + \frac{\sum_{i=1}^{n}x_i^2}{\sigma^2}\right)^{-1}}\left(\beta_1^2 - \frac{2\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2\left(\frac{1}{a_1^2} + \frac{\sum_{i=1}^{n}x_i^2}{\sigma^2}\right)}\right)\right] \text{ Mas } \left(\frac{1}{a_1^2} + \frac{\sum_{i=1}^{n}x_i^2}{\sigma^2}\right)^{-1} = \frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)}\left(\beta_1^2 - \frac{2\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)}\right)\right]$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)}\left(\beta_1^2 - \frac{2a_1^2\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)\right]$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)}\left(\beta_1^2 - \frac{2a_1^2\beta_1\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2} + \left(\frac{a_1^2\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)^2\right)\right]$$

$$\propto \exp\left[-\frac{1}{2\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)}\left(\beta_1 - \frac{a_1^2\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)^2\right]$$

ou seja, $\beta_1 \mid \beta_0, \sigma^2, \boldsymbol{x}, \boldsymbol{y} \sim N\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}, \frac{a_1^2\sum_{i=1}^{n}\mu_i^{(2)}x_i}{\sigma^2 + a_1^2\sum_{i=1}^{n}x_i^2}\right)$

✓ Por fim, a condicional de σ^2 dado β_1, β_0 e os dados:

$$f(\sigma^2 \mid \beta_0, \beta_1, \boldsymbol{x}, \boldsymbol{y}) \propto (\sigma^2)^{-(b+\frac{n}{2}+1)} \exp\left[-\frac{d}{\sigma^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$\text{ou seja, } \sigma^2 \mid \beta_0, \beta_1, \boldsymbol{x}, \boldsymbol{y} \sim \text{ IG } \left(b + \frac{n}{2}, d + \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

✓ Esta metodologia de encontrar as *posterioris* condicionais para os coeficientes da regressão se extende ao modelo de regressão linear múltipla de maneira análoga. ✓ Uma outra alternativa a este problema é utilizar uma *priori} conjunta conjugada. Veja o texto: Exemplo Regressao.pdf.

Aplicação da Metodologia

✓ Atribuir *prioris* não informativas para β_0 e β_1: Normais com média igual a zero e variância grande, por exemplo 10^6. ✓ Atribuir *prioris* não informativas para σ^2. A distribuição $IG(0.001, 0.001)$ é não informativa, e é muito próxima à *priori* de Jeffreys para o modelo normal com média e variância desconhecidos. **Tarefa: Verificar!** ✓ Utilizar o algoritmo de Gibbs. Veja descrição em sala com aplicação no R e Winbugs.