# 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

Tábua de distribuições

Atividades síncronas no google meet (dia & horários no Guia de Estudos)

# 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 $\Omega$ 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.

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 $\Omega$, também chamada de *classe de subconjuntos do espaço amostral* $\Omega$, ou *Conjunto das partes de* $\Omega$, é o conjunto que contém todos os subconjuntos de $\Omega$ 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 $\Omega$) 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 $\Omega$ 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 $\Omega$, 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:

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

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

## 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 $\theta$;
- Verossimilhança : relaciona todas as variáveis num modelo de probabilidade completo, então temos $L(\theta|\boldsymbol{y})$: função de verossimilhança de $\theta$ 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 $\theta$.

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 $\theta$ 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 $\theta$ pode assumir, ou seja, **o espaço paramétrico de $\theta$ é discreto**,

- b) **Caso contínuo:** neste caso, assumimos que $\theta$ 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 $\Theta$ é o espaço paramétrico de $\theta$, **o espaço paramétrico de $\theta$ 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 $\int_\Theta f(\theta)L(\theta|\boldsymbol{y})d\theta$ pois se trata de uma constante que não depende de $\theta$, então temos

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

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 $\theta$?

**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 $p$;
- b) Considerando a distribuição *a priori* Gamma com média 1 e variância 5, obtenha a distribuição *a posteriori* para $\lambda$;
- c) Segundo o item $b$), qual é a média *a posteriori* para $\lambda$?
- d) Segundo o item $b$), qual é a variância *a posteriori* para $\lambda$? Veremos mais tarde que os exercícios 4 e 5 envolvem distribuições *a priori* conjugadas.

**Resolução:** - Estimativa de máxima verossimilhança

```
lambda_hat=mean(x) #estimativa de maxima verossimilhança
lambda_hat
```

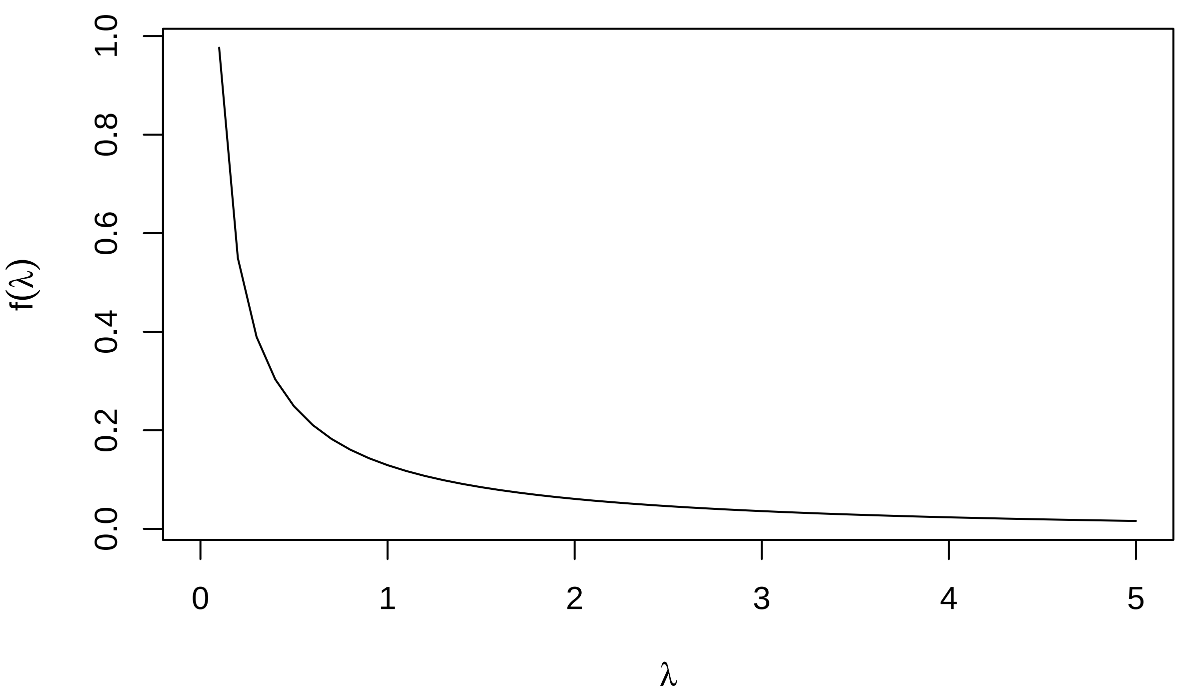
```
## [1] 2.3
```

- Fazendo os graficos da priori, verossimilhança e posteriori

```
lambda=seq(0,5,0.1) #só assume valores positivos
```

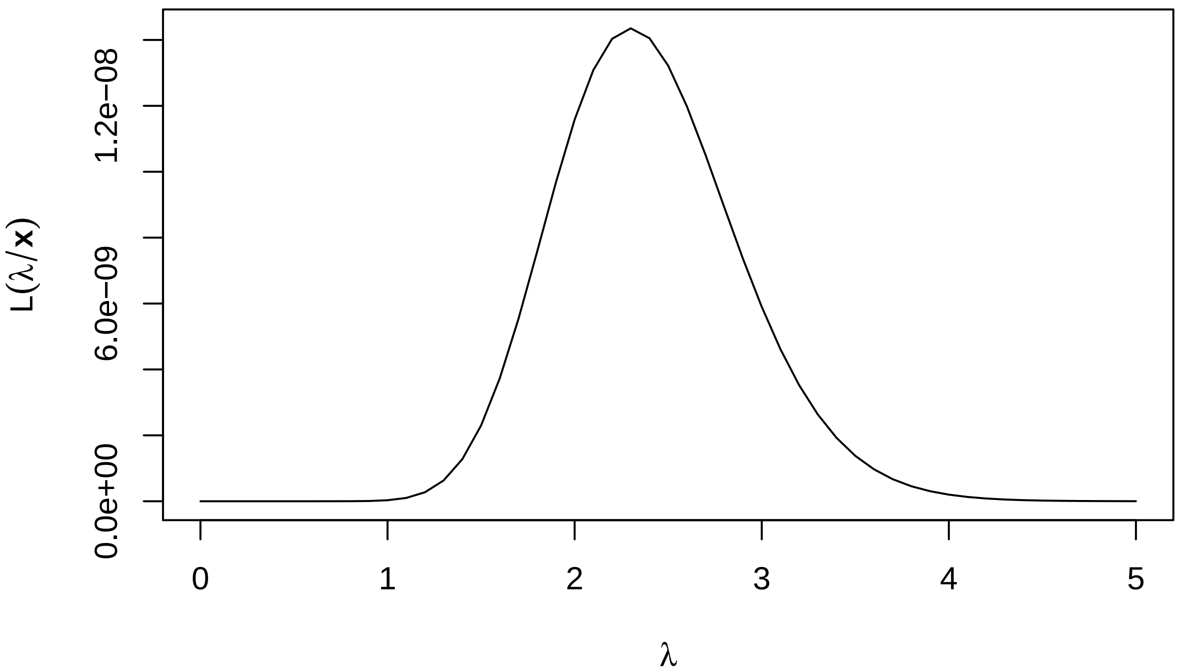
- Priori

```
priori_lambda=dgamma(lambda,shape=1/5, scale=5) #na parametrização do R, o parâmetro de   escala beta é
plot(lambda,priori_lambda,type='l',xlab=expression(lambda),ylab=expression(f(lambda)))
```



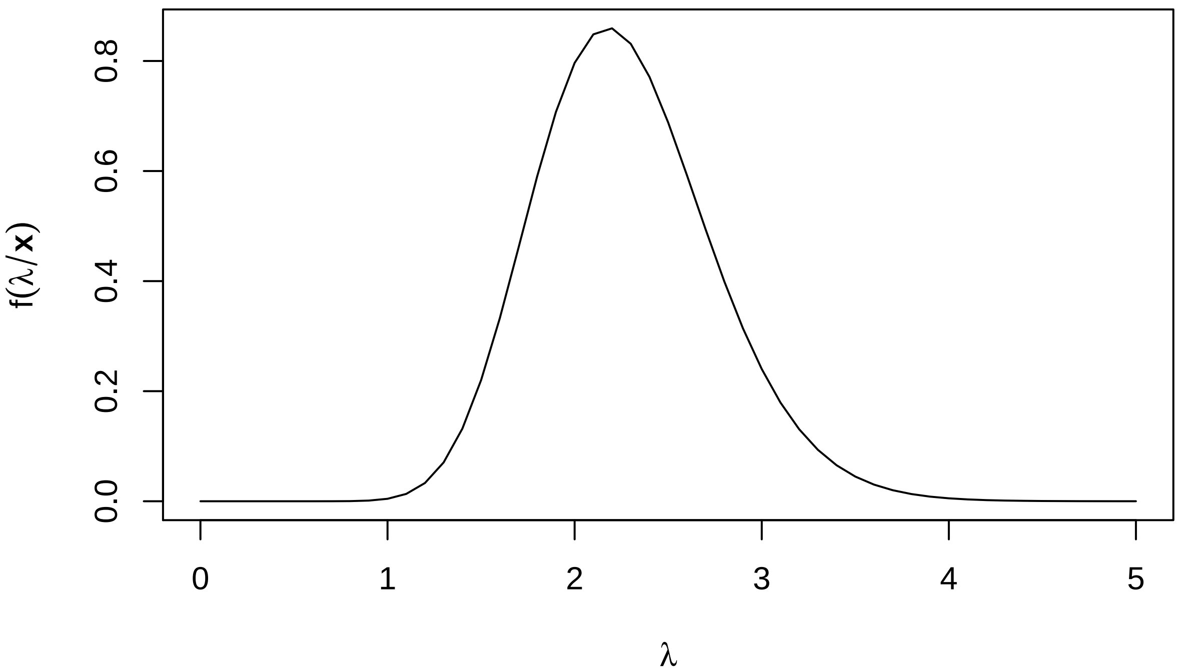

- Verossimilhança

```
L_lambda=exp(-n*lambda)*lambda^(sum(x))*1/(prod(factorial(x)))
plot(lambda,L_lambda,type='l',ylab=expression(L(lambda/bold(x))),xlab=expression(lambda))
```




- Posteriori

```
posteriori_lambda=dgamma(lambda,shape=sum(x)+1/5, scale=1/(n+1/5)) #na parametrização do R, o parâmetro
plot(lambda,posteriori_lambda,type='l',xlab=expression(lambda),ylab=expression(f(lambda/bold(x))))
```




- **Exemplo 1.1:** Ensaios de Bernoulli com distribuição *a priori* discreta. Uma determinada droga tem taxa de resposta $\theta$ 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.

# 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

### Priori normal e verossimilhança normal (Caso 1)

✓ Temos uma amostra de tamanho $n$ i.i.d. com distribuição normal, com média $\mu$ e variância $\sigma^2$: $Y_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n$:

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right],$$

com $\mu$ desconhecido e $\sigma^2$ conhecido;

✓ A função de verossimilhança será

$$
\begin{aligned}
L(\mu|\boldsymbol{y}) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\
&= \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[ -\frac{1}{2\sigma^2}(y_i - \mu)^2 \right] \right\} \\
&\propto \exp\left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right],
\end{aligned}
$$

donde o símbolo $\propto$ significa "é proporcional a", ou seja, todos os termos multiplicativos que não dependem de $\mu$ podem ser desconsiderados na fórmula.

A função de verossimilhança nos traz toda a informação disponível na amostra (nos dados).

✓ A distribuição *a priori* para $\mu$ é normal: $\mu \sim N(m_0, s_0^2)$:

$$
f(\mu) = \frac{1}{\sqrt{2\pi} s_0} \exp\left[ -\frac{1}{2s_0^2}(\mu - m_0)^2 \right],
$$

com $m_0$ e $s_0^2$ conhecidos. A distribuição *a priori* nos traz o conhecimento *a priori* sobre a média $\mu$.

- Se temos pouca informação a respeito de $\mu$, podemos fixar a média $m_0$ e atribuir uma variância $s_0^2$ grande;
- Se temos muita informação a respeito de $\mu$, podemos fixar a média $m_0$ e atribuir uma variância $s_0^2$ pequena. ✓ A distribuição *a posteriori* para $\mu$ será:

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

donde temos:

$$
\mu|\boldsymbol{y} \sim N\left( \frac{\frac{m_0}{s_0^2} + \frac{n\overline{y}}{\sigma^2}}{\frac{1}{s_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{s_0^2} + \frac{n}{\sigma^2}} \right)
$$

**Demonstração** Veja demonstração em aula.

✓ Então concluímos que dado um modelo normal com média desconhecida e variância conhecida, então a *priori* conjugada para a média é normal.

## Prioris Conjugadas - continuação

**Exemplo 1** Box & Tiao (1973) Os físicos $A$ e $B$ desejam determinar uma constante física $\theta$. O físico $A$ tem mais experiência nesta área e especifica sua priori como $\theta \sim N(900, 20^2)$. O físico $B$ tem pouca experiência e especifica uma priori muito mais incerta em relação à posição de $\theta$, $\theta \sim N(800, 80^2)$. Assim, nós verificamos que:

- Para o físico $A$: $P(860 < \theta < 940) \approx 0,95$;
- Para o físico $B$: $P(640 < \theta < 960) \approx 0,95$.

Faz-se então uma medição $X$ de $\theta$ em laboratório com um aparelho calibrado com distribuição amostral $X|\theta \sim N(\theta, 40^2)$ e observou-se $X = 850$. Aplicando o "Caso 1)" de *prioris* conjugadas, temos:

- para o físico $A$: - a variância de $\theta$ era igual a 400, e passou a ser igual a 320 $\Rightarrow$ significa que ganhamos informação com os dados observados; - a **precisão** de $\theta$ passou de $0,0025$ para $0,00312$ (aumento de 25% na precisão), - a **precisão** é usualmente representada pela letra grega $\tau$ e é igual ao inverso da variância;

- para o físico $B$: - a variância de $\theta$ era igual a 6400, e passou a ser igual a 1280; - precisão de $\theta$ passou de $0,000156$ para $0,000781$ (aumento de 400%).

- Abaixo temos a representação gráfica da distribuição *a priori*, verossimilhança e distribuição *a posteriori*.

- 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.

**Tarefa** Verifique o desenvolvimento em R para este exemplo.

- Distribuição a priori para o físico A: $\theta$ $N(900, 40^2)$.
- Qual é o intervalo de valores que corresponde a 95% da área sob a curva da normal? - quantis dos limites inferior e superior

```
q1=qnorm(0.025,mean=900,sd=20)
q2=qnorm(0.975,mean=900,sd=20)
print(paste0("q1=",q1," & q2=",q2))
```

```
## [1] "q1=860.800720309199 & q2=939.199279690801"
```

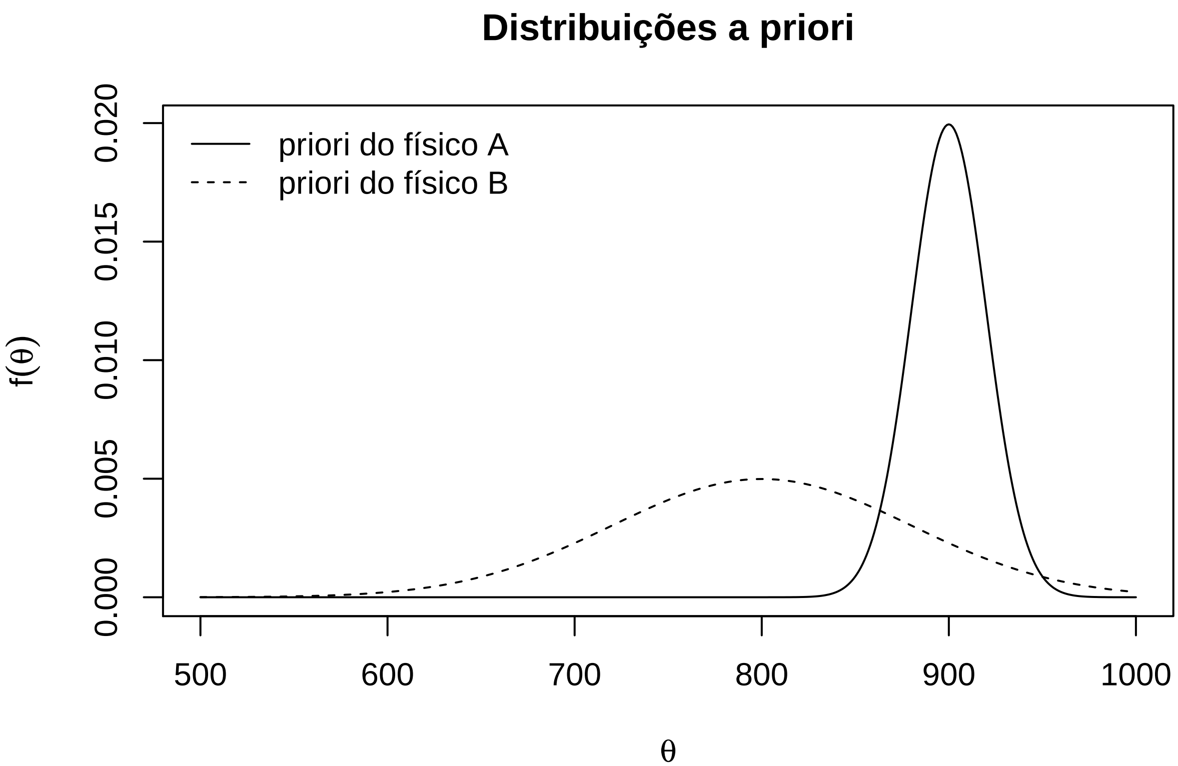
- Distribuição a priori para o físico B: $\theta$ $N(900, 40^2)$ - quantis dos limites inferior e superior

```
q1=qnorm(0.025,mean=800,sd=80)
q2=qnorm(0.975,mean=800,sd=80)
print(paste0("q1=",q1," & q2=",q2))
```

```
## [1] "q1=643.202881236796 & q2=956.797118763204"
```

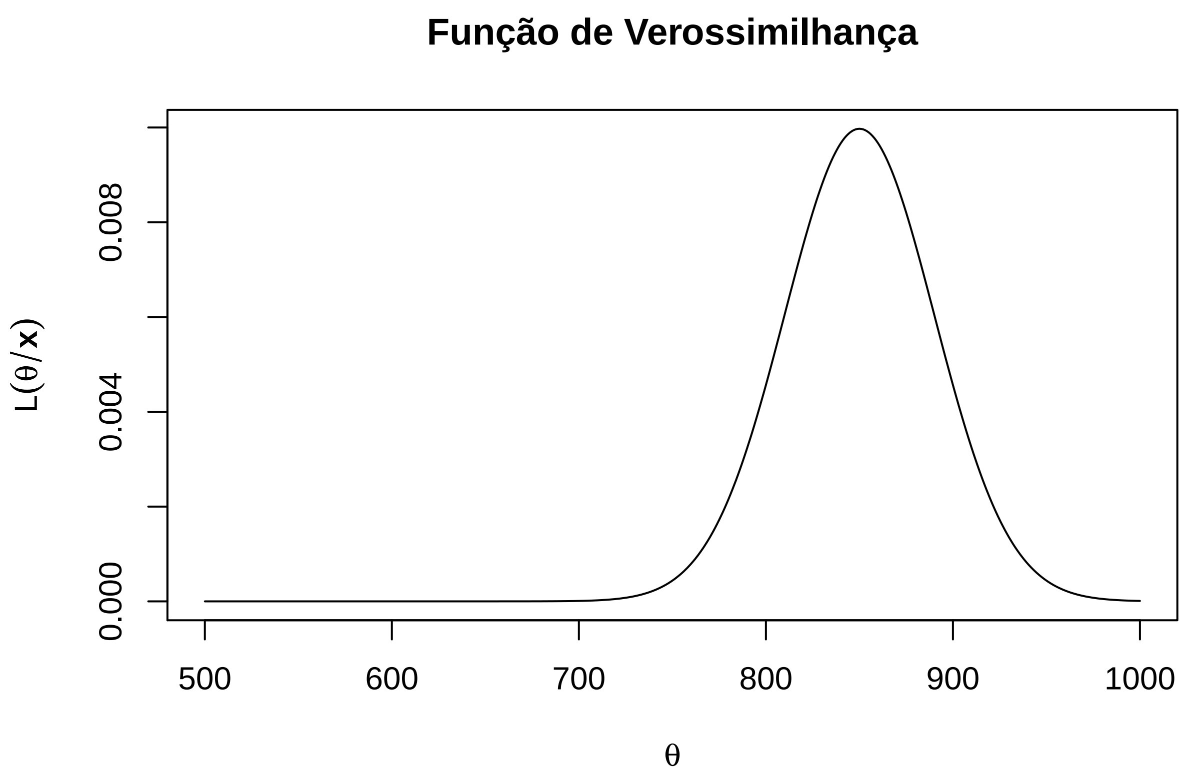
- Gráficos das distribuições a priori - atribui valores no eixo para theta, para plotar no gráfico

```
theta=seq(500,1000)
priori_theta_A=dnorm(theta,mean=900,sd=20)
priori_theta_B=dnorm(theta,mean=800,sd=80)
plot(theta,priori_theta_A,type='l',xlab=expression(theta),ylab=expression(f(theta)),
main="Distribuições a priori")
lines(theta,priori_theta_B,type='l',lty=2)
legend("topleft",c("priori do físico A","priori do físico B"),lty=c(1,2),bty = "n")
```



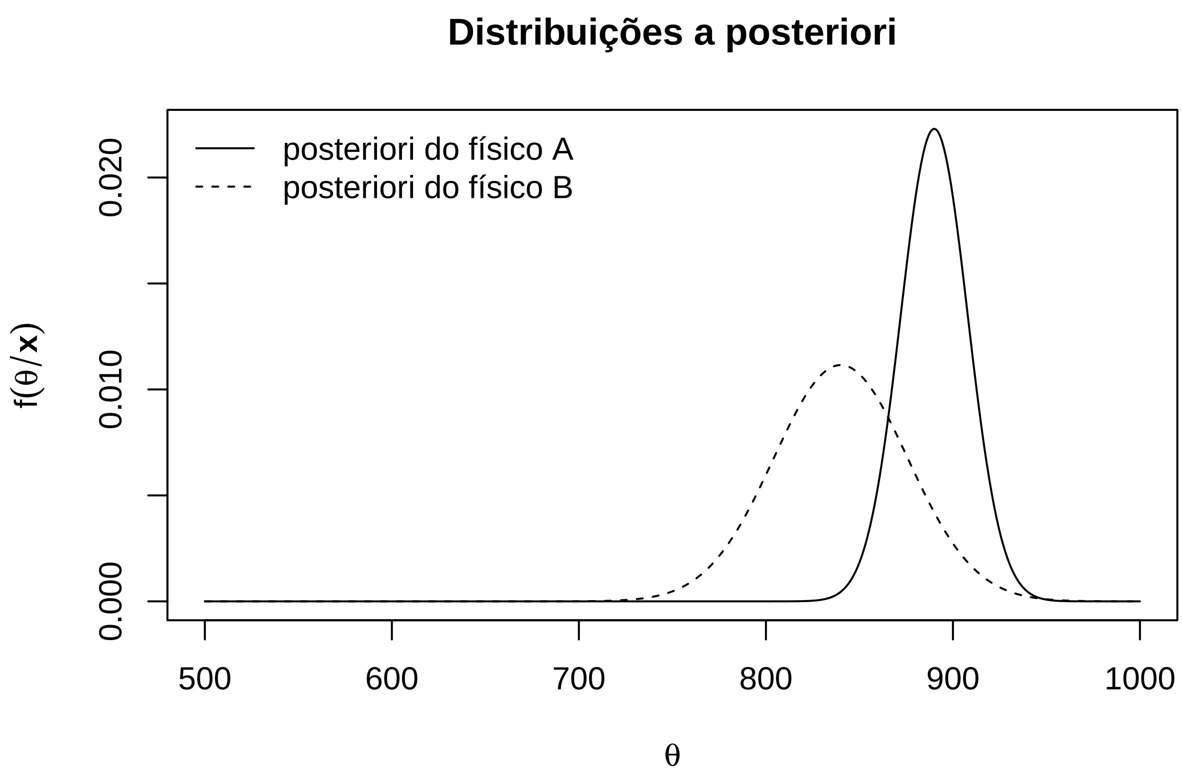

- Gráfico da função de verossimilhança

```
x=850
L_theta=dnorm(x,mean=theta,sd=40)
plot(theta,L_theta,type='l',xlab=expression(theta),ylab=expression(L(theta/bold(x))),
main="Função de Verossimilhança")
```




- Gráfico das distribuições a posteriori. As distribuições à posteriori vêm do "Caso 1)" de prioris conjugadas

```
n=1 #tamanho da amostra
y_bar=850 #é a média amostral dos dados
sigma2=1600 #é a variância dos dados (no caso 1 a variancia é conhecida)
posteriori=function(m_0,sigma2_0){ #m_0 é a média a priori e sigma2_0 é a variância a priori
#(varia para os físicos A e B)
media=(m_0/sigma2_0+n*y_bar/sigma2)/(1/sigma2_0+n/sigma2)
#onde m_0 é a média a priori e sigma2_0 é a variância a priori (varia para os físicos A e B)
variancia=1/(1/sigma2_0+n/sigma2)
dnorm(theta,mean=media,sd=sqrt(variancia))
}
plot(theta,posteriori(900,20^2),type='l',xlab=expression(theta),ylab=expression(f(theta/bold(x))),main=
lines(theta,posteriori(800,80^2),type='l',lty=2)
legend("topleft",c("posteriori do físico A","posteriori do físico B"),lty=c(1,2),bty = "n")
```

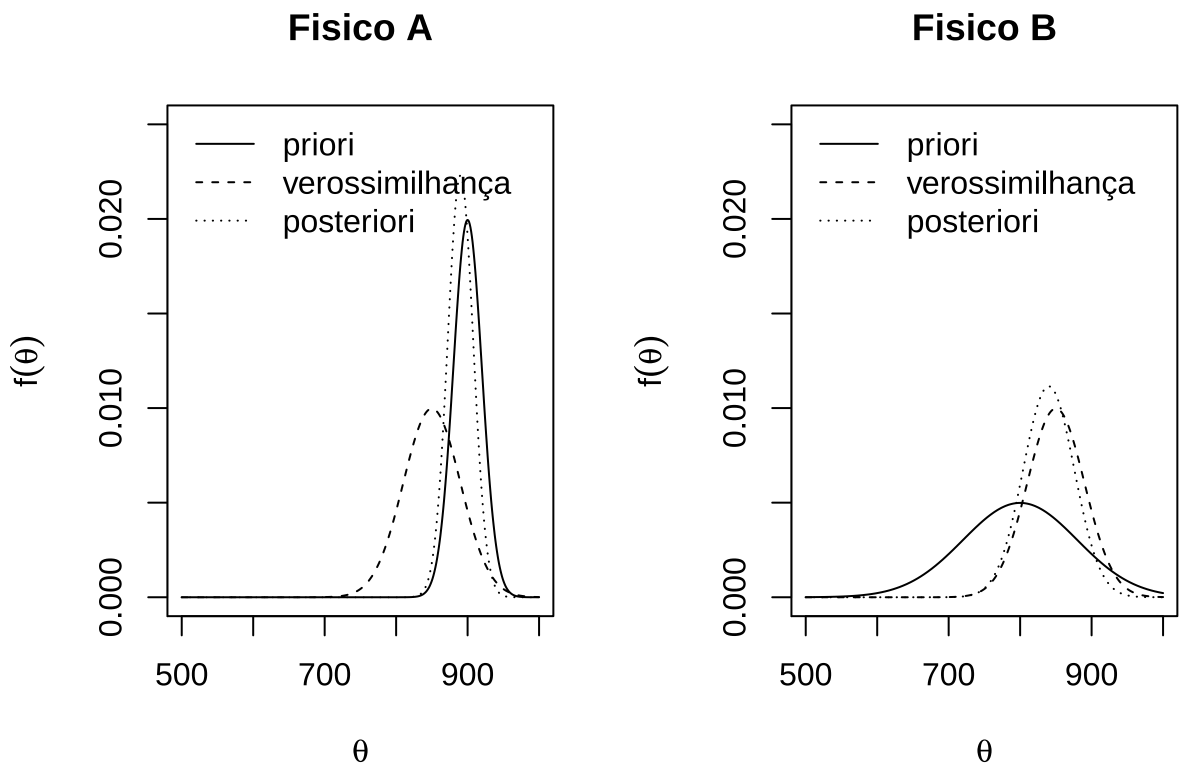



- Gráficos das três funções conjuntamente: priori, verossimilhança e posteriori. - cria uma janela para dois gráficos (1 linha por 2 colunas)

```
par(mfrow=c(1,2))

#Para o físico A
plot(theta,priori_theta_A,type='l',xlab=expression(theta),ylab=expression(f(theta)),
main="Fisico A",ylim=c(0,0.025))
lines(theta,L_theta,type='l',lty=2)
lines(theta,posteriori(900,20^2),type='l',lty=3)
legend("topleft",c("priori","verossimilhança","posteriori"),lty=c(1,2,3),bty = "n")

#Para o físico B
plot(theta,priori_theta_B,type='l',xlab=expression(theta),ylab=expression(f(theta)),
main="Fisico B",ylim=c(0,0.025))
lines(theta,L_theta,type='l',lty=2)
lines(theta,posteriori(800,80^2),type='l',lty=3)
legend("topleft",c("priori","verossimilhança","posteriori"),lty=c(1,2,3),bty = "n")
```




**Distribuição de Poisson($\lambda$) com *priori* gamma para a taxa $\lambda$ (Caso 2)**

(Caso 2) - Temos uma amostra de tamanho $n$ i.i.d. com distribuição **Poisson** com taxa $\lambda$: $Y_i \sim \text{Poisson}(\lambda), i = 1, 2, \ldots, n$:

$$P(Y_i = y_i) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!},$$

com $\lambda$ desconhecido;

- A função de verossimilhança será

$$\begin{aligned} L(\lambda|\boldsymbol{y}) &= \prod_{i=1}^{n} P(Y_i = y_i) \\ &= \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \\ &\propto \exp(-n\lambda)\lambda^{\sum_{i=1}^{n} y_i}, \end{aligned}$$

em que todos os termos multiplicativos que não dependem de $\lambda$ foram desconsiderados na fórmula.

$L(\lambda|\boldsymbol{y})$ é uma função de $\lambda$ (nosso parâmetro desconhecido).

- Na **inferência clássica**, nós procedemos com o método usual de maximizar $\log(L(\lambda|\boldsymbol{y}))$ com respeito a $\lambda$, 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.
- A distribuição *a priori* para $\lambda$ é **Gamma**: $\lambda \sim \text{Gamma}(\alpha, \beta)$,

$$f(\lambda) = \frac{\beta^{\alpha}}{\Gamma[\alpha]}\lambda^{\alpha-1}\exp(-\beta\lambda),$$

  com $\alpha > 0$ e $\beta > 0$ parâmetros conhecidos.
- A distribuição *a posteriori* para $\lambda$ será:

$$f(\lambda|\boldsymbol{y}) \quad \propto \quad \exp\left(-\lambda(n+\beta)\right)\lambda^{\sum\limits_{i=1}^{n} y_i+\alpha-1}$$

donde temos:

$$\lambda|\boldsymbol{y} \sim \text{Gamma}\left(\sum_{i=1}^{n} y_i + \alpha, n+\beta\right)$$

## Distribuição de Binomial($\theta$) com *priori* beta para a proporção $\theta$ (Caso 3)

Demonstração em aula

## Distribuição Normal para os dados, com média conhecida e variância desconhecida (Caso 4)

- Temos uma amostra de tamanho $n$ i.i.d. com distribuição **normal** com média $\mu$ e variância $\sigma^2$: $Y_i \sim N(\mu, \sigma^2), i = 1, 2, \ldots, n$:

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right],$$

  com $\mu$ conhecido e $\sigma^2$ desconhecido;
- A função de verossimilhança será

$$\begin{aligned}
L(\sigma^2|\boldsymbol{y}) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\
&= \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right]\right\} \\
&\propto \left(\frac{1}{\sigma}\right)^{n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right],
\end{aligned}$$

donde todos os termos multiplicativos que não dependem de $\sigma^2$ podem ser desconsiderados na fórmula e colocamos o sinal de $\propto$: "proporcional a"

Neste problema, temos que enxergar o parâmetro $\tau = \frac{1}{\sigma^2}$, então a função de verossimilhança passa a ser em função da precisão $\tau$:

$$L(\tau|\boldsymbol{y}) \propto \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right],$$

e observamos que o núcleo da verossimilhança possui a mesma forma do núcleo de uma distribuição gamma. Por isso faz sentido atribuirmos uma distriubição *a priori* gamma para a precisão $\tau$.

- A distribuição *a priori* para $\tau$ é **gamma**: $\tau \sim$ Gamma $(\alpha, \beta)$:

$$f(\tau) = \frac{\beta^\alpha}{\Gamma[\alpha]}\tau^{\alpha-1}\exp(-\beta\tau),$$

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

- A distribuição *a posteriori* para $\tau$ será:

$$f(\tau|\boldsymbol{y}) \propto \tau^{\alpha+\frac{n}{2}-1}\exp\left[-\tau\left(\beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)\right],$$

  donde temos:

$$\tau|\boldsymbol{y} \sim \text{Gamma}\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

- Devido à relação entre as distribuições gamma e gamma invertida, temos:

- Se atribuímos uma distribuição *a priori* gamma para a precisão $\tau$: $\tau \sim$ Gamma $(\alpha, \beta)$ tal que

$$f(\tau) = \frac{\beta^\alpha}{\Gamma[\alpha]}\tau^{\alpha-1}\exp(-\beta\tau),$$

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

- É equivalente a atribuir uma distribuição *a priori* gamma invertida para a variância $\sigma^2$: $\sigma^2 \sim$ Gamma Invertida $(\alpha, \beta)$ tal que

$$f(\sigma^2) = \frac{\beta^\alpha}{\Gamma[\alpha]}\left(\sigma^2\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right).$$

**Tarefa:** Demonstrar esta relação utilizando transformação de variáveis.

- E se dizemos que a distribuição *a posteriori* para $\tau$ é gamma: $\tau|\boldsymbol{y} \sim \text{Gamma}\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$,

- É equivalente a dizer que a distribuição *a posteriori* para $\sigma^2$ é gamma invertida: $\sigma^2|\boldsymbol{y} \sim \text{Gamma Invertida}\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$.

**Exemplo 2** Abaixo temos 10 valores provenientes de uma distribuição normal com média $\mu$ e variância $\sigma^2$.

```
set.seed(05062017) #cria uma semente única
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(10,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 $\sigma^2$;
- 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 $\tau$;
- c) Segundo o item b), qual é a média *a posteriori* para $\tau$?
- d) Segundo o item b), qual é a varância *a posteriori* para $\tau$?

**Solução**

- a) A estimativa de máxima verossimilhança para $\sigma^2$ é $\widehat{\sigma^2} = 2.2837$,
- b) Distribuição *a priori* para $\tau$ é Gamma: $\tau \sim$ Gamma $(\alpha, \beta)$, tais que $E(\tau) = 0.5$ e $VAR(\tau) = 0.5$, logo $\alpha = 0,5$ e $\beta = 1$. Pelo "Caso 4)",

$$\tau|\boldsymbol{y} \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

  E como $n = 10$ e $\sum_{i=1}^{n}(y_i - \mu)^2 = 23.2993$, então

$$\tau|\boldsymbol{y} \sim \text{Gamma}\,(5.5, 12.6497)$$

- c) A média *a posteriori* será:

$$E(\tau|\boldsymbol{y}) = \frac{5.5}{12.6497} \approx 0.4351, \text{ próxima da estimativa de máxima verossimilhança: } \widehat{\tau} = 0.4379.$$

- d) A variância *a posteriori* será:

$$VAR(\lambda|\boldsymbol{y}) = \frac{5.5}{160.0149} \approx 0.0344.$$

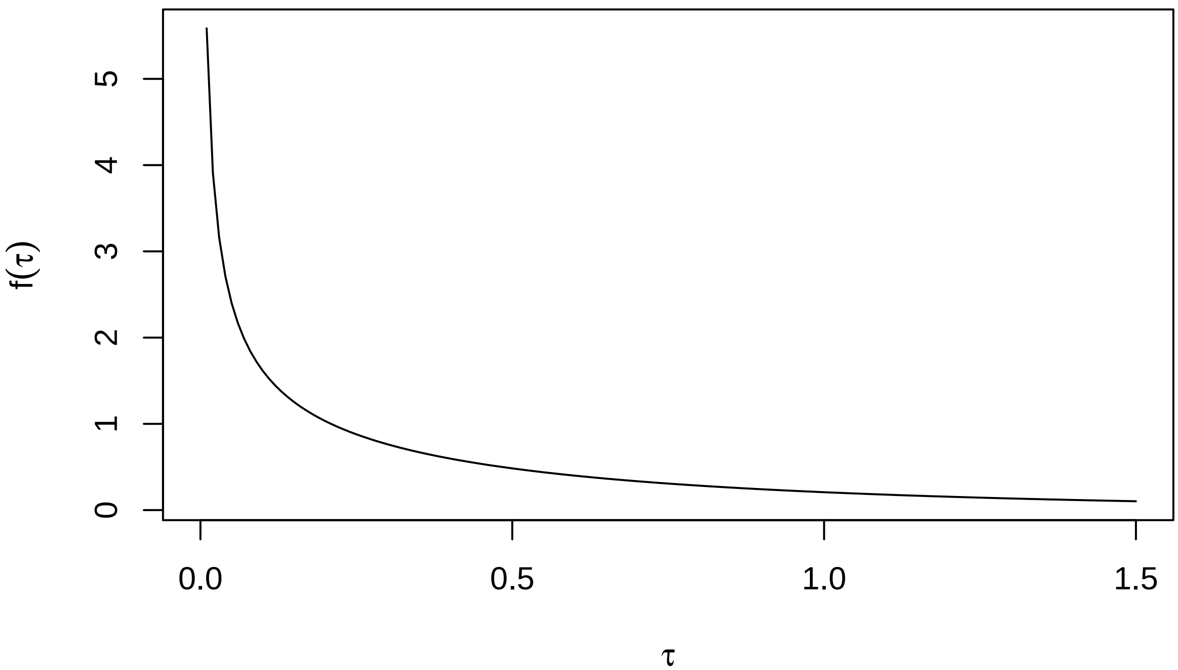
**Tarefa** Verificar os cálculos e o código em R para estes dados.

- Estimativa de máxima verossimilhança de $\sigma^2$

```
y=c(1.2156,1.2000,2.1362,2.1139,2.6546,0.0135,-0.0007,0.2131,3.3849,4.9196)
n=10
sigma2_hat=sum((y-mean(y))^2)/n
```

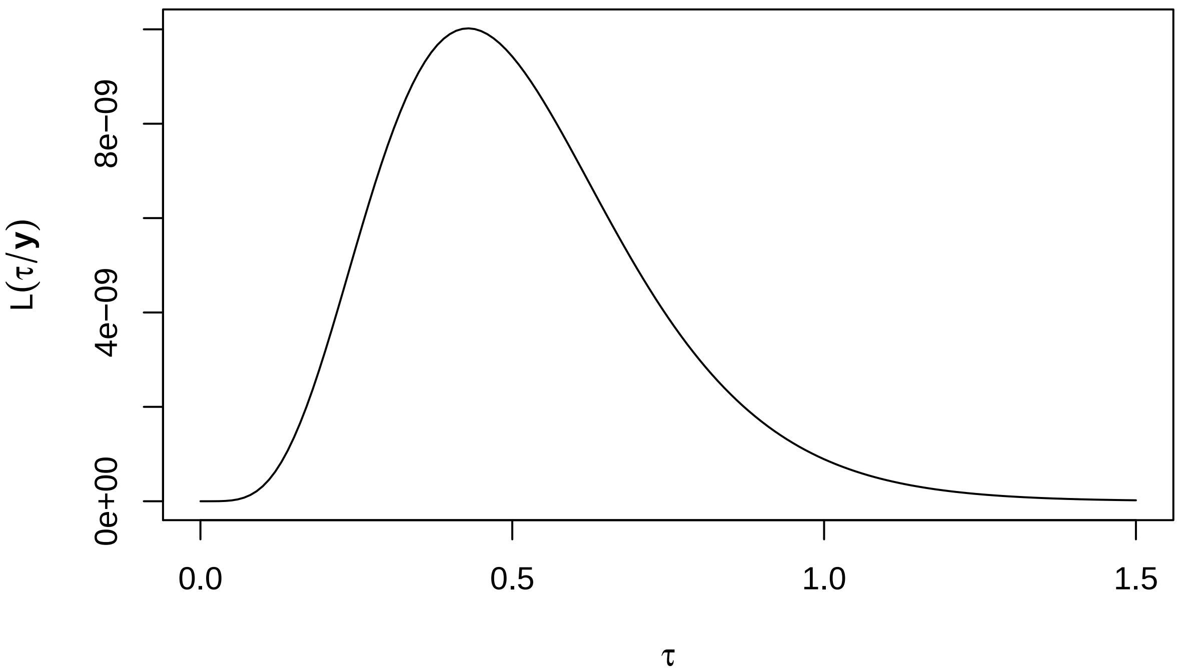
- Gráfico da função de densidade a priori - intervalo de 0.01 para a curva ficar bem suavizada - parametrização da dist. gamma no R

```
tau=seq(0,1.5,0.01)
alpha=0.5
beta=1
priori=dgamma(tau,shape=alpha, scale=1/beta)
plot(tau,priori,type='l',xlab=expression(tau),ylab=expression(f(tau)))
```



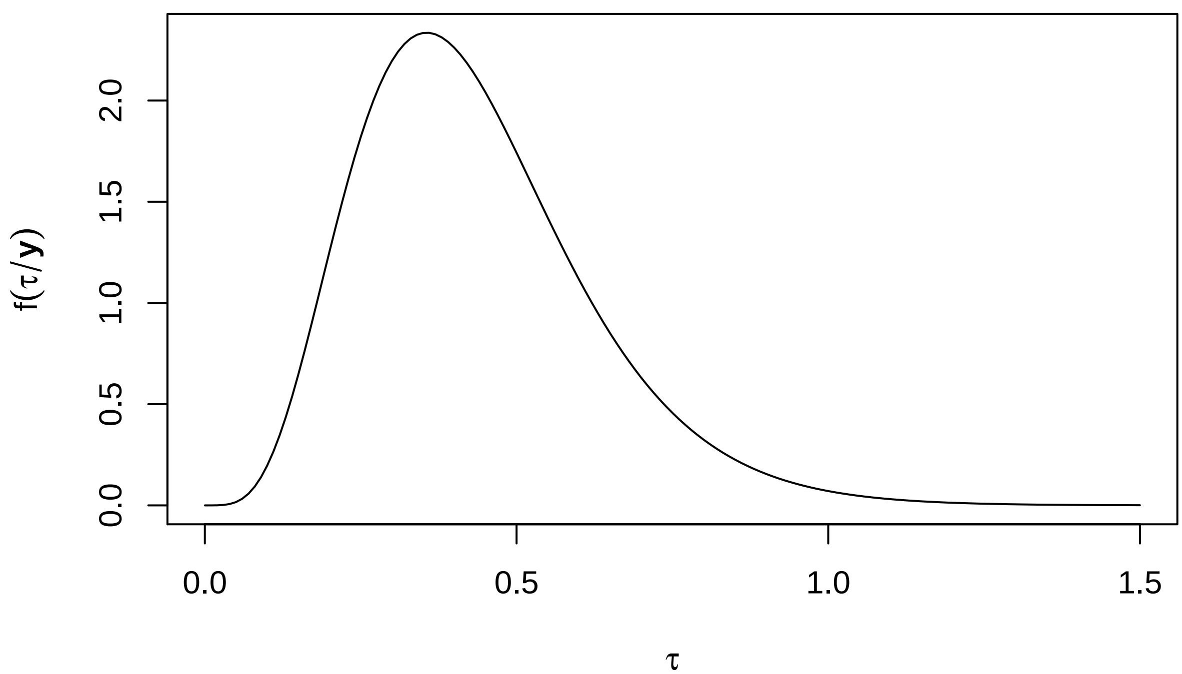

- Gráfico da função de verossimilhança - vamos considerar a média conhecida

```
mu=2
L_tau=(1/sqrt(2*pi))^n*tau^(n/2)*exp(-tau/2*sum((y-mu)^2))
plot(tau,L_tau,xlab=expression(tau),ylab=expression(L(tau /bold(y))),type='l')
```




- Posteriori

```
posteriori=dgamma(tau,shape=alpha+n/2, scale=1/(beta+1/2*sum((y-mu)^2)))
plot(tau,posteriori,xlab=expression(tau),ylab=expression(f(tau /bold(y))),type='l')
```




### 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 $\theta$ desconhecido. Com base na experiência anterior assume-se uma distribuição a priori Gamma com média 0.2 e desvio-padrão 1 para $\theta$. Se o tempo médio para atender uma amostra aleatória de 20 clientes foi de 3.8 minutos, determine a distribuição a posteriori de $\theta$.

- 4. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro $\theta$. Determine os parâmetros da priori conjugada de $\theta$ sabendo que $E(\theta) = 4$ 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 $\theta$. Atribui-se uma distribuição a priori Gamma $(2, 10)$ para $\theta$. Se um rolo de 1200 metros desta fita foi inspecionado e encontrou-se 4 defeitos, qual é a distribuição *a posteriori* de $\theta$?

## Princípio da Verossimilhança

**Exemplo:** Suponha que desejamos estimar $\theta$, 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 $\theta$, 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 $\theta$ e que possuem verossimilhancas distintas, diferindo apenas por uma constante que não depende de $\theta$, 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 $\theta$ são igualmente prováveis:

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

com $\theta$ 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 $\theta$ 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 $\theta$ então a *priori* para o parâmetro $\phi$ 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 $\phi$ 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 $\theta$: $f(x|\theta)$. A medida de informação esperada de Fisher de $\theta$ 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 $\theta$). - 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 $\theta$ 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 $X_1, \ldots, X_n \sim \text{Poisson}(\theta)$.

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

- A *priori* de Jeffreys para $\theta$ no modelo Poisson é $f(\theta) \propto \theta^{-1/2}$;
- Esta priori também pode ser obtida tomando-se a *priori* conjugada Gamma$(\alpha, \beta)$ com $\alpha = \frac{1}{2}$ e $\beta \to 0$. Note que o parâmetro $\beta$ 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 $\mu$ tais que:

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

logo $\theta$ é o parâmetro de locação.

- A definição se extende para o caso multiparamétrico;
- **Exemplos:** distribuição normal com variância conhecida e distribuição normal multivariada com matriz de variância-covariância conhecida.
- **Propriedade:** A *priori* de Jeffreys para o parâmetro de locação $\mu$ 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 $\sigma$ tais que:

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

logo $\sigma$ é o parâmetro de escala.

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

- **Propriedade:** A *priori* de Jeffreys para o parâmetro de escala $\sigma$ é:

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

**Definição: Modelo de Locação-escala** $X$ tem um modelo de locação-escala se existem uma função $g$ e as quantidades $\mu$ e $\sigma$ tais que

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

logo $\mu$ é o parâmetro de locação e $\sigma$ é o parâmetro de escala.

- **Exemplos:** Na distribuição $\mathrm{N}(\mu,\sigma^2)$ o parâmetro de locação é $\mu$ e o parâmetro de escala é $\sigma$, e a distribuição de Cauchy também é um modelo de locação-escala.
- **Propriedade** A *priori* conjunta de Jeffreys para os parâmetros de locação $\mu$ e escala $\sigma$ 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 $\mu$ e $\sigma^2$ 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 $\mu$ é o parâmetro de locação e $\sigma$ é o parâmetro de escala.

- A *priori* não informativa de Jeffreys para o vetor $(\mu,\sigma)$ é:

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

- Pela **propriedade da invariância**, a *priori* não informativa de Jeffreys para o vetor $(\mu,\sigma^2)$ é:

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

## 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 $\sigma$ 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 $\nu$ 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 $\alpha$ é o parâmetro de locação e $\beta$ 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

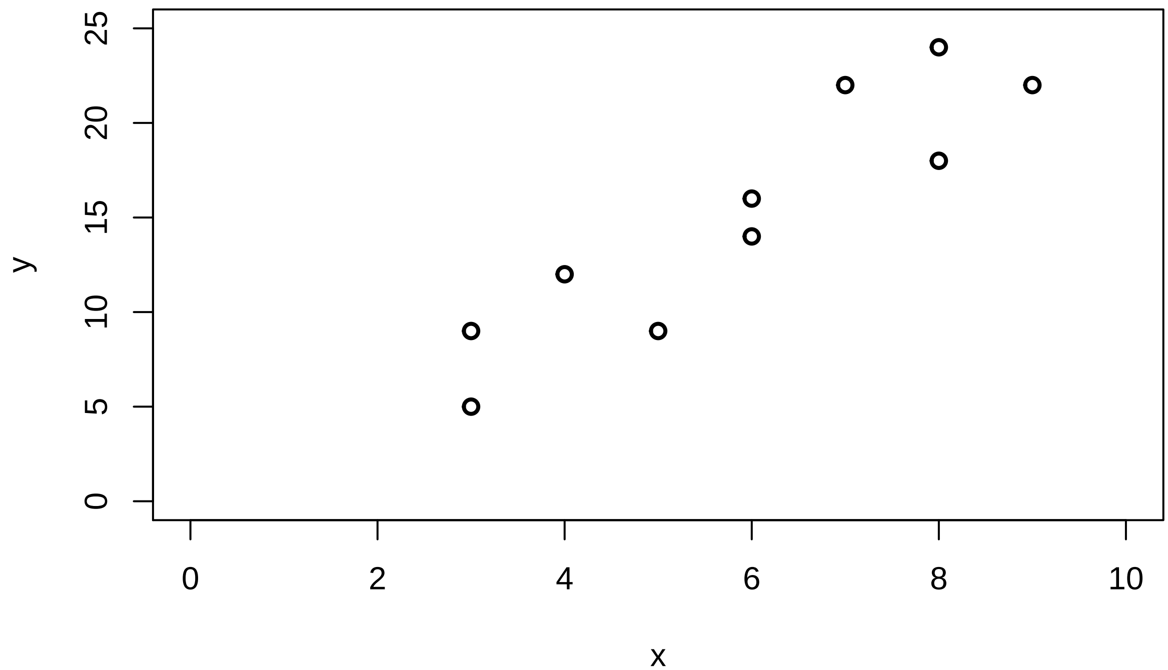
O problema envolve as variáveis $X$: a dose de um medicamento anti-alérgico em estudo, e $Y$: tempo de duração do efeito (alívio dos sintomas alérgicos). Abaixo temos a representação gráfica dos dados observados Pelo gráfico, nós concluímos que uma relação linear (reta) é satisfatória para os dados.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

```
##       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
```

```
plot(x,y,lwd=2,xlim=c(0,10),ylim=c(0,25),main="Figura 5: Diagrama de dispersão dos dados")
```

**Figura 5: Diagrama de dispersão dos dados**




**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 $\beta_0$: intercepto da linha de regressão com o eixo $y$; $\beta_1$: coeficiente de inclinação da reta: é o nº de unidades em $y$ que mudam para cada unidade da variável independente $x$. $\epsilon_i$: 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 $\sigma_2$:

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

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 $\beta_0$ e $\beta_1$ são baseados na distribuição t-student.

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

```
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

```
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
```

# 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:

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

- Distribuição *a posteriori* conjunta para $\beta_0$,$\beta_1$ e $\sigma^2$ é dada por:

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

- Distribuições *a posteriori* marginais - é necessário integrar com respeito aos outros parâmetros, respeitando os limites de integração: ✓ A conjunta de $\beta_0$ e $\beta_1$ é obtida da integração com respeito à variância $\sigma^2$:

$$
f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) = \int_0^\infty f(\beta_0, \beta_1, \sigma^2 \mid \boldsymbol{x}, \boldsymbol{y}) d\sigma^2
$$

✓ A marginal de $\sigma^2$ é obtida da integração com respeito à $\beta_0$ e $\beta_1$:

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

✓ A marginal de $\beta_0$ é obtida da integração com respeito à $\beta_1$:

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

✓ A marginal de $\beta_1$ é obtida da integração com respeito à $\beta_0$:

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

✓ A conjunta de $\beta_0$ e $\beta_1$ pode ser obtida analiticamente:

$$
\begin{aligned}
f(\beta_0, \beta_1 \mid \boldsymbol{x}, \boldsymbol{y}) &\propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \int_0^\infty (\sigma^2)^{-(b+\frac{n}{2}+1)} \exp\left[-\frac{1}{\sigma^2}\left(d + \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)\right] d\sigma^2 \\
&\propto \exp\left[-\frac{\beta_0^2}{2a_0^2}\right] \exp\left[-\frac{\beta_1^2}{2a_1^2}\right] \left[d + \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]^{-\left(b+\frac{n}{2}\right)}.
\end{aligned}
$$

**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 $\infty$.

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 $\sigma^2$ dado $\beta_0$, $\beta_1$ e os dados:

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

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

a idéia de condicional nos diz que $\beta_0$ e $\beta_1$ são tratadas como constantes com respeito a $\sigma^2$.

✓ A condicional de $\beta_0$ dado $\beta_1$, $\sigma^2$ 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 $\beta_1$ dado $\beta_0$, $\sigma^2$ e os dados: Note que o parâmetro $\beta_1$ 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, $\beta_1 \mid \beta_0, \sigma^2, \boldsymbol{x}, \boldsymbol{y} \sim N\left(\frac{a_1^2\sigma^2}{\sigma^2 + a_1^2 \sum x_i^2}, \frac{a_1^2 \sum_{i=1}^{n}\mu_i^{(2)} x_i}{\sigma^2 + a_1^2 \sum x_i^2}\right)$

✓ Por fim, a condicional de $\sigma^2$ dado $\beta_1$, $\beta_0$ e os dados:

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

✓ 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 $\beta_0$ e $\beta_1$: Normais com média igual a zero e variância grande, por exemplo $10^6$. ✓ Atribuir *prioris* não informativas para $\sigma^2$. 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.