Métodos computacionais para aproximar a distribuição a posteriori

Prof. Vinícius D. Mayrink
EST088 - Inferência Bayesiana

Sala: 4073 Email: vdm@est.ufmg.br

1^o semestre de 2025

5.1 - Distribuições a posteriori sem forma fechada

Algum tempo atrás (rever conjunto de slides 3), estudamos as famílias conjugadas. Vimos que se a *priori* $f_\theta(\theta)$ é membro de uma família de distribuições $\mathcal{F}$ que conjuga com o modelo estabelecido pela verossimilhança $f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$, então a *posteriori* $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ também será membro de $\mathcal{F}$.

Este resultado é bastante útil e estabelece um protocolo a ser seguido para simplificar as contas do Teorema de Bayes e garantir a identificação de uma distribuição *a posteriori* tendo **forma fechada** da f.d.p. ou f.m.p.

O termo **forma fechada** significa que a formulação da f.d.p./f.m.p. *a posteriori* é completamente conhecida (sua integração no espaço paramétrico Θ dará 1).

Porém, nem todo modelo estatístico permite utilizar a conjugação para obter **forma fechada** a *posteriori*. Existem várias situações em que apenas o **núcleo** da distribuição alvo pode ser calculado. Faltará uma constante normalizadora que completa a fórmula da f.d.p. ou f.m.p.

Para uma exposição mais concreta deste problema, reconsidere o estudo do modelo $Y_i \sim N(\mu, 1/\phi)$ com ambos os parâmetros desconhecidos. Reveja este caso no conjunto 3 de slides deste curso.

Admita independência condicional de $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_n\}$ dado $\{\mu, \phi\}$.

Além disso, adote independência *a priori*: $f_{\mu,\phi}(\mu, \phi) = f_\mu(\mu)\ f_\phi(\phi)$.

Espaço paramétrico: $\mu \in \mathbb{R}$ e $\phi \in \mathbb{R}^+$.

Especifique: $\mu \sim N(m, v)$ e $\phi \sim \text{Ga}(a, b)$.

$$f_{\mu,\phi}(\mu, \phi) = (2\pi v)^{-1/2} \exp\{-\tfrac{1}{2v}(\mu - m)^2\} \quad \tfrac{b^a}{\Gamma(a)}\ \phi^{a-1} \exp\{-b\phi\}.$$

Verossimilhança: $f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y}) = \prod_{i=1}^n (2\pi/\phi)^{-1/2} \exp\{-\tfrac{\phi}{2}(y_i - \mu)^2\}$.

$$= (2\pi)^{-n/2}\ \phi^{n/2}\ \exp\{-\tfrac{\phi}{2}(\textstyle\sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2)\}.$$

Distribuição *a posteriori*: $f_{\mu,\phi|\boldsymbol{Y}}(\mu,\phi|\boldsymbol{y}) \propto f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y})\, f_{\mu,\phi}(\mu,\phi)$.

Após algums contas (rever slides), chegou-se ao seguinte resultado

$$f_{\mu,\phi|\boldsymbol{Y}}(\mu,\phi|\boldsymbol{y}) \propto (v_\phi^*)^{-1/2} \exp\left\{-\frac{1}{2v_\phi^*}(\mu - m_\phi^*)^2\right\} \times$$

$$\times (v_\phi^*)^{1/2} \exp\left\{\frac{(m_\phi^*)^2}{2v_\phi^*}\right\} \times \phi^{\tilde{a}-1} \exp\{-\phi\tilde{b}\},$$

sendo $v_\phi^* = (n\phi + \frac{1}{v})^{-1}$, $m_\phi^* = v_\phi^*(\phi\sum_{i=1}^n y_i + \frac{m}{v})$,

$\tilde{a} = a + n/2$ e $\tilde{b} = b + \sum_{i=1}^n y_i^2/2$.

O subscrito "ϕ" é um lembrete de que as expressões dependem de ϕ.

Temos o núcleo de uma Normal em **vermelho** e o núcleo de uma Gama em **verde**. O termo (em **preto**) depende de ϕ e não pode ser incorporado na constante normalizadora. Ele também não pode ser combinado com as demais partes para formar uma distribuição conhecida.

Não é possível identificar a distribuição conjunta *a posteriori* $f_{\mu,\phi|\boldsymbol{Y}}(\mu, \phi|\boldsymbol{y})$.

Consequentemente, estimativas pontuais (média, moda e mediana) e intervalares (intervalo de credibilidade) para μ e ϕ não estão disponíveis. A ausência de forma fechada da distrbuição *a posteriori* configura-se um obstáculo para a inferência.

A solução para este tipo de problema é considerar estratégias computacionais que permitem uma **amostragem indireta** da distribuição alvo desconhecida.

O termo **amostragem indireta** indica que valores serão gerados de uma distribuição alvo sem a necessidade de conhecer sua f.d.p. ou f.m.p. completa. O máximo que conhecemos é um núcleo (expressão sem forma fechada, faltando a constante normalizadora) obtido via Teorema de Bayes.

Se pudermos gerar indiretamente um número suficiente de valores provenientes da distribuição alvo, seremos capazes de avaliar estatísticas descritivas amostrais que fornecerão as estimativas pontuais e intervalares desejadas na inferência.

Antes de prosseguirmos, é importante estabelecer um conceito chave para a situação em que amostras geradas a partir de uma distribuição alvo estão disponíveis para análise.

Definição: Suponha m valores $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}$ gerados independentemente a partir de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. O **estimador Monte Carlo** estabelece uma aproximação do resultado da integral $E_{\theta|\boldsymbol{Y}}[h(\theta)|\boldsymbol{y}] = \int_{\Theta} h(\theta)\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta$ com o seguinte cálculo:

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} h(\theta^{(i)}).$$

Naturalmente, quanto maior o valor de m, melhor é a aproximação.

Este resultado indica que se há interesse em estimar um parâmetro $h(\theta)$ usando a esperança *a posteriori*, podemos simplesmente aplicar a transformação $h(\bullet)$ aos valores gerados $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}$ e computar a média aritmética.

Dois casos particulares da definição anterior são mostrados a seguir.

- Se $h(\theta) = \theta$, o estimador Monte Carlo fornecerá a média da distribuição *a posteriori*:

$$E_{\theta|\boldsymbol{Y}}(\theta) \approx \bar{\theta} = \frac{1}{m}\sum_{i=1}^{m} \theta^{(i)}.$$

- Se $h(\theta) = [\theta - E_{\theta|\boldsymbol{Y}}(\theta)]^2$, teremos a variância *a posteriori*:

$$Var_{\theta|\boldsymbol{Y}}(\theta) \approx \frac{1}{m}\sum_{i=1}^{m} (\theta^{(i)} - \bar{\theta})^2.$$

Visto que a tendência é ter m grande, tanto faz calcular a versão acima ou a não viciada abaixo

$$Var_{\theta|\boldsymbol{Y}}(\theta) \approx \frac{1}{m-1}\sum_{i=1}^{m} (\theta^{(i)} - \bar{\theta})^2.$$

Esta variância expressa a incerteza *a posteriori* sobre a estimação de θ.

Além do estimador Monte Carlo $\bar{h}$, podemos obter a variância deste estimador.

Definição: Sejam $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}$ gerados independentemente a partir de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. A **variância do estimador Monte Carlo** $\bar{h} = \sum_{i=1}^{m} h(\theta^{(i)})/m$ é dada por:

$$S_h^2 = \frac{\frac{1}{m-1}\sum_{i=1}^{m}\left[h(\theta^{(i)}) - \bar{h}\right]^2}{m} = \frac{1}{m\,(m-1)}\sum_{i=1}^{m}\left[h(\theta^{(i)}) - \bar{h}\right]^2.$$

Note que $\bar{h}$ é uma média amostral e a fórmula acima é basicamente a variância amostral dividida pelo tamanho amostral m.

Obs.: O cálculo acima fornece a variabilidade do estimador $\bar{h}$ e não a variabilidade *a posteriori* sobre θ expressa em $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

5.2 - Importance sampling

Este método é um primeiro passo para a análise dita **Monte Carlo** em que observações simuladas de uma distribuição são usadas para explorar outra distribuição (simular de uma distribuição "errada" pode ser útil).

Alguns outros métodos que também consideram esta estratégia são o *Rejection sampling* e o *Metropolis-Hastings*. Eles serão estudados mais adiante.

O *Importance sampling* é uma escolha interessante para problemas em que:

- temos baixa dimensão (poucos parâmetros);
- a densidade da distribuição de interesse pode ser obtida, mas ela não permite uma simulação direta;
- é fácil identificar e simular a partir de distribuições que aproximam a distribuição de interesse.

O *Importance sampling* aplicado no contexto da inferência Bayesiana tem a configuração descrita a seguir.

A distribuição de interesse, a partir da qual deseja-se amostrar valores, é a distribuição *a posteriori* com f.d.p. ou f.m.p. $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

A forma fechada de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ não está disponível. Após aplicar o Teorema de Bayes, obteve-se apenas o núcleo desta função (falta a constante normalizadora). Denote esse núcleo por $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

Seja $g_\theta(\theta)$ uma f.d.p. ou f.m.p. conhecida (chamada de distribuição *Importance sampling*). Ela é escolhida visando uma aproximação análitica para $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Algumas diretrizes para selecionar $g_\theta(\theta)$ são:

- facilidade para gerar valores de θ a partir de $g_\theta(\theta)$;
- para qualquer θ no domínio de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ é possível calcular $g_\theta(\theta)$ e $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

Quanto maior a semelhança entre $g_\theta(\theta)$ e $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$, mais preciso será o resultado do *Importance sampling*.

Algumas vezes $g_\theta(\theta)$ é apenas uma f.d.p. ou f.m.p. conveniente e de fácil simulação que não foi escolhida com a preocupação de se ter grande semelhança com $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Nestes casos, uma amostra Monte Carlo grande ajuda a superar o problema de aproximação entre $g_\theta(\theta)$ e $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

No *Importance sampling* estamos interessados no seguinte cálculo de esperança:

$$H = \int_\Theta h(\theta)\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta.$$

Podemos escrever:

$$H = \int_\Theta h(\theta)\, \frac{f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})}{g_\theta(\theta)}\, g_\theta(\theta)\, d\theta = \int_\Theta h(\theta)\, \omega(\theta)\, g_\theta(\theta)\, d\theta.$$

Obs.: Não conhecemos a constante normalizadora de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Usaremos o núcleo $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ em substituição como segue $\omega(\theta) = \pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})/g_\theta(\theta)$.

Este novo $\omega(\theta)$ requer um procedimento de normalização para contrapor a ausência da constante normalizadora. Isto será visto mais adiante, por enquanto mantenha $\omega(\theta) = f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})/g_\theta(\theta)$.

Temos $\int_\Theta h(\theta)\, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta = \int_\Theta h(\theta)\, \omega(\theta)\, g_\theta(\theta)\, d\theta$.

O resultado acima indica que a esperança de $h(\theta)$ sob $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ foi trocada pela esperança de $h(\theta)\, \omega(\theta)$ sob $g_\theta(\theta)$.

Sejam m valores $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}$ gerados independentemente a partir de $g_\theta(\theta)$. O **estimador Monte Carlo** que aproxima $\int_\Theta h(\theta)\, \omega(\theta)\, g_\theta(\theta)\, d\theta$ será:

$$\bar{h} = \frac{1}{m} \sum_{i=1}^{m} h(\theta^{(i)})\, \omega(\theta^{(i)}).$$

Note que estamos amostrando aqui da distribuição "errada" $g_\theta(\theta)$.

A medição do quão "errado" é o valor simulado $\theta^{(i)}$ é feita através do peso:

$$\omega(\theta^{(i)}) = f_{\theta|\boldsymbol{Y}}(\theta^{(i)}|\boldsymbol{y}) / g_\theta(\theta^{(i)}).$$

Estes pesos calibram as estimativas $h(\theta^{(i)})$ visando corrigir o fato de amostrarmos da distribuição auxiliar $g_\theta(\theta)$.

Lembrete: abordamos aqui o cenário em que a constante normalizadora de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ é desconhecida. Temos acesso apenas ao núcleo $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.
Os valores $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}$ são gerados independentemente de $g_\theta(\theta)$.

O seguinte procedimento de normalização é necessário:

$$\omega_i = \frac{\omega(\theta^{(i)})}{\sum_{i=1}^m \omega(\theta^{(i)})} \quad \text{sendo} \quad \omega(\theta^{(i)}) = \frac{\pi_{\theta|\boldsymbol{Y}}(\theta^{(i)}|\boldsymbol{y})}{g(\theta^{(i)})}.$$

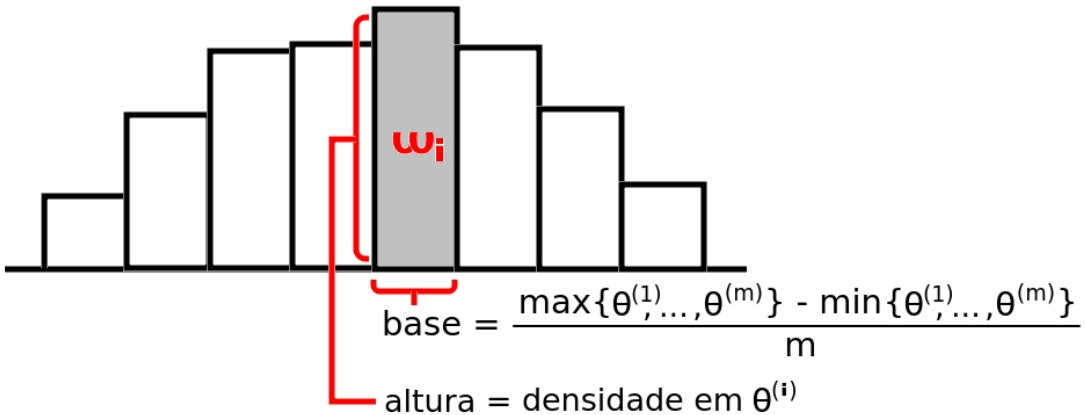
Considere $\hat{h} = \sum_{i=1}^m \omega_i \, h(\theta^{(i)})$ para aproximar $E_{\theta|\boldsymbol{Y}}[h(\theta)] = \int_\Theta h(\theta) \, f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) \, d\theta$

Use $\hat{v}_h = \sum_{i=1}^m \omega_i \, [\, h(\theta^{(i)}) - \hat{h} \,]^2$ para aproximar $Var_{\theta|\boldsymbol{Y}}[h(\theta)]$

Referência: *Geweke (1989) Bayesian inference in econometric models using Monte Carlo integration, Econometrica, 57, 6, 1317-1339*. Este artigo discute a convergência e a teoria assintótica associada ao estimador $\hat{h}$.

No resultado do último slide, estamos usando uma distribuição discreta para aproximar $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Esta distribuição discreta define probabilidade ω_i para o ponto $\theta^{(i)}$, sendo $i = 1, \cdots, m$.

Se a distribuição *a posteriori* de θ é contínua, teremos interesse em calcular o valor aproximado da densidade $f_{\theta|\boldsymbol{Y}}(\theta^{(i)}|\boldsymbol{y}) \neq \omega_i$. O raciocíno para obter esta densidade leva em conta a estrutura de um histograma.



ω_i = base x altura

altura = ω_i / base

$$\text{densidade em } \theta^{(i)} = \frac{\omega_i \; m}{\max\{\theta^{(1)},\ldots,\theta^{(m)}\} - \min\{\theta^{(1)},\ldots,\theta^{(m)}\}}$$

Exemplo: Sejam $Y_1, Y_2, \cdots, Y_n$ dado θ condicionalmente independentes com $Y_i|\theta \sim$ Bernoulli(θ). Veja que $W = \sum_{i=1}^{n} Y_i \sim$ Binomial(n, θ). Este caso já foi estudado anteriormente (rever conjunto de slides 3). Temos

- Espaço paramétrico: $\theta \in [0, 1]$.
- Verossimilhança: $f_{W|\theta}(w) = \binom{n}{w}\theta^w(1-\theta)^{n-w}$.
- Distribuição *a priori*: $\theta \sim$ Beta(a, b).
- Distribuição *a posteriori*: $\theta|W \sim$ Beta(a^*, b^*), com $a^* = a + w$ e $b^* = b + w + n$.

A f.d.p. *a posteriori* deste problema tem forma fechada, entretanto, para ilustrar o uso do *Importance sampling* iremos imaginar que a constante normalizadora (em vermelho) é desconhecida.

$$f_{\theta|W}(\theta|w) = \frac{\Gamma(a^*+b^*)}{\Gamma(a^*)\ \Gamma(b^*)}\ \theta^{a^*-1}\ (1-\theta)^{b^*-1}.$$

Denote: $\pi_{\theta|W}(\theta|w) = \theta^{a^*-1}(1-\theta)^{b^*-1}$.

O objetivo aqui é usar o *Importance sampling* para avaliar a distribuição *a posteriori* $f_{\theta|W}(\theta|w)$ sem ter acesso direto à sua forma fechada. Apenas o núcleo $\pi_{\theta|W}(\theta|w)$ acima é conhecido.

Admita nesta análise que:

- $\theta \sim \text{Beta}(a = 1, b = 1)$.
- A coleta de dados estabeleceu: $n = 10$ e $w = 3$.
- Perante as configurações acima temos: $a^* = 4$ e $b^* = 14$.
- Distribuição geradora auxiliar: $g_\theta(\theta)$ é a f.d.p. da $U(0, 1)$.

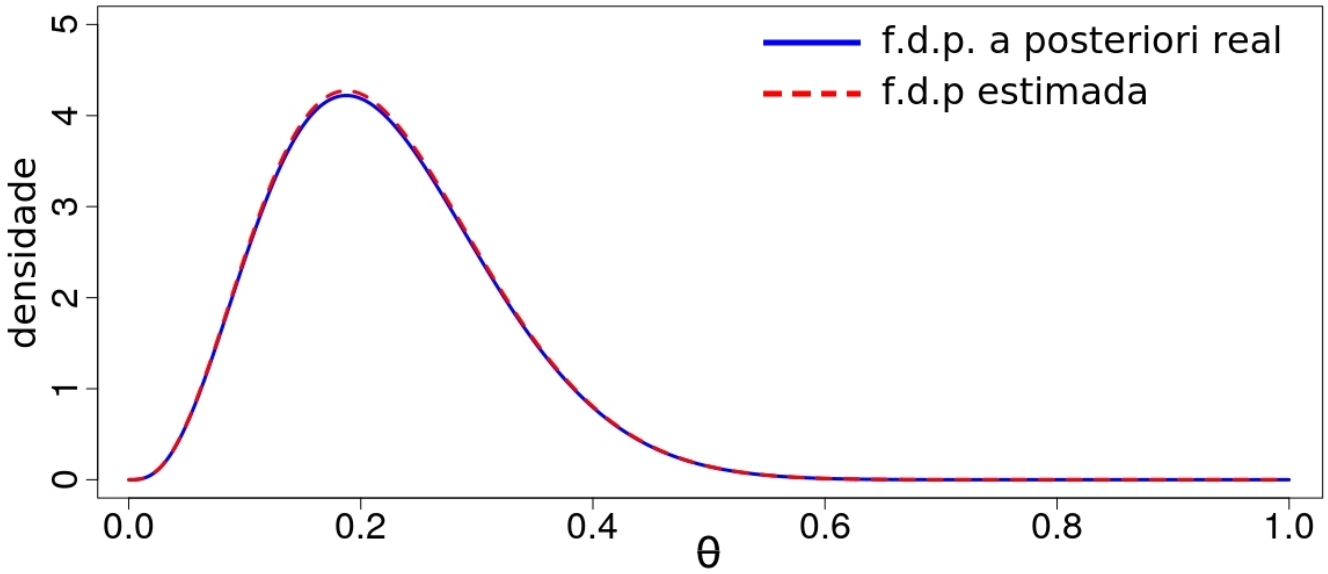
O seguinte código R aplica o método *Importance sampling*:

```
a = 1; b = 1;
n = 10; w = 3;
as = a + w;
bs = b + w + n;

m = 1000  # número pontos a serem gerados

theta = runif(m, 0, 1)  # gerando valores a partir de g(theta)

pi_theta = theta^(as-1) * (1-theta)^(bs-1) # avaliando o núcleo
w_theta = pi_theta / dunif(theta,0,1)      # calculando pesos
w_theta = w_theta / sum(w_theta)           # normalizando
me = sum(w_theta * theta)                  # média a posteriori
v = sum(w_theta * (theta - me)^2)          # variância a posteriori

d_post = (w_theta * m) / (max(theta) - min(theta))  # densidade alvo
```

A média e variância *a posteriori* são $\frac{a^*}{a^*+b^*} = 0.2222$ e $\frac{a^* b^*}{(a^*+b^*)^2(a^*+b^*+1)} = 0.0091$.

O método estimou média `me = 0.2237` e variância `v = 0.0094`.

A figura abaixo compara a f.d.p. *a posteriori* real (em azul), dada pela Beta(4,14), e a densidade estimada (vermelho trecejado) via *Importance sampling*. Foi usado $m = 1000$ gerações de valores a partir da f.d.p. $U(0,1)$ para $g_\theta(\theta)$.



Exercício: Siga o raciocínio deste exemplo Binomial, suponha desconhecimento da forma fechada da f.d.p. *a posteriori*, e implemente o *Importance sampling* admitindo $Y_1, Y_2, \cdots, Y_n$ dado θ i.i.d. tal que:

- $Y_i|\theta \sim N(\theta, \sigma^2)$ com σ^2 conhecido.
- $Y_i|\theta \sim N(\mu, \theta)$ com μ conhecido.

5.3 - Método da rejeição (Rejection sampling)

Este é outro método útil para gerar valores *a posteriori* em problemas com distribuições univariadas ou com baixa dimensão.

Ele utiliza ideias do *Importance sampling*, mas considera algumas correções ditas "exatas" para fornecer amostras "exatas" da distribuição de interesse.

O método da rejeição pode ser complicado de implementar, visto que requer o conhecimento de um limite superior M para o chamado "*importance ratio*".

Outras metodologias alternativas já existem na literatura e elas oferecem vantagens em relação ao método da rejeição até mesmo para problemas simples.

Contudo, a estratégia de aceitação/rejeição, presente no método da rejeição, baseia-se em princípios essenciais que ajudarão a compreender adiante um algorítmo mais robusto conhecido como Metropolis-Hastings.

Seja $\omega(\theta) = f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})/g_\theta(\theta) < M$ uma razão de importância (*importance ratio*) delimitada pela constante $M > 1$ que deve ser escolhida.

- $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ é a distribuição *a posteriori* da qual queremos amostrar.
- $g_\theta(\theta)$ é uma distribuição auxiliar para gerar candidados θ^*.
- $\omega(\theta) < M$ pode ser rescrito como $0 < \omega(\theta)/M < 1$.
- Se $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}) = g_\theta(\theta)$, teremos $\omega(\theta) = 1 < M$.
- Se $g_\theta(\theta)$ é uma boa aproximação para $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$, então a escolha de M não precisa ser muito maior que 1.

O método tem a seguinte lógica:

- Gere um candidato θ^* a partir de $g_\theta(\theta)$.
- $\omega(\theta^*)$ grande $\Rightarrow$ θ^* é compatível com $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$, pois numerador tem alta importância na razão $\omega(\theta^*)$.
- $\omega(\theta^*)$ pequeno $\Rightarrow$ θ^* é incompatível com $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$, pois numerador tem baixa importância na razão $\omega(\theta^*)$.

No teste de aceitação/rejeição, o candidato θ^* é aceito com probabilidade $\omega(\theta^*)/M$. A implementação do teste é simples:

- gere $\mathcal{U} \sim U(0,1)$ independente de θ^*.
- se $\mathcal{U} \leq \omega(\theta)/M$, aceite θ^* como uma observação de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.
- se $\mathcal{U} > \omega(\theta)/M$, rejeite θ^* e gere um novo candidato.

O método da rejeição é justificado pela análise da distribuição de $\theta|\mathcal{U}$, ou seja, a distribuição de um candidato que foi aceito.

$$f_{\theta|\mathcal{U}}(\theta|\mathcal{U} < \omega(\theta)/M) = \frac{f_{\theta,\mathcal{U}}(\theta, \mathcal{U} < \omega(\theta)/M)}{f_{\mathcal{U}}(\mathcal{U} < \omega(\theta)/M)} = \frac{f_{\mathcal{U}|\theta}(\mathcal{U} < \omega(\theta)/M)\, g_\theta(\theta)}{f_{\mathcal{U}}(\mathcal{U} < \omega(\theta)/M)}$$

$$f_{\mathcal{U}|\theta}(\mathcal{U} < \omega(\theta)/M) = \omega(\theta)/M, \text{ visto que } \mathcal{U} \sim U(0,1).$$

$$f_{\mathcal{U}}(\mathcal{U} < \omega(\theta)/M) = \int_\Theta f_{\mathcal{U}|\theta}(\mathcal{U} < \omega(\theta)/M)\, g_\theta(\theta)\, d\theta = \int_\Theta \frac{\omega(\theta)}{M}\, g_\theta(\theta)\, d\theta$$

$$= \frac{1}{M} \int_\Theta \frac{f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})}{g_\theta(\theta)}\, g_\theta(\theta)\, d\theta = \frac{1}{M} \int_\Theta f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})\, d\theta = 1/M.$$

Conclusão:

$$f_{\theta|\mathcal{U}}(\theta|\mathcal{U} < \omega(\theta)/M) \;=\; \frac{g_\theta(\theta)\,\omega(\theta)/M}{1/M} \;=\; g_\theta(\theta)\,\omega(\theta) \;=\; f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y}).$$

Este resultado mostra que os valores aceitos de θ são de fato provenientes da distribuição alvo $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$.

Uma questão importante no método da rejeição é a eficiência do algorítmo tomando como base a proporção de rejeições. A escolha de $g_\theta(\theta)$ e M tem impacto sobre a taxa de rejeição (se ela for alta, vai demorar para obter a quantidade amostral desejada).

O resultado $f_{\mathcal{U}}(\mathcal{U} < \omega(\theta)/M) = 1/M$ sugere que a probabilidade marginal de aceitar um candidado θ é igual a $1/M$.

Uma escolha $g_\theta(\theta)$ que permite selecionar M perto de 1, determina uma amostragem eficiente (alta probabilidade marginal de aceitação)

Nos slides anteriores a razão de importância $\omega(\theta)$ foi definida com numerador $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Entretanto estamos em um cenário no qual apenas o núcleo $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ está disponível (falta a constante normalizadora C_N).

$$\omega_1(\theta) = \frac{f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})}{g_\theta(\theta)} = \frac{C_N\, \pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})}{g_\theta(\theta)} < M_1.$$

Podemos então escrever

$$\omega_2(\theta) = \frac{\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})}{g_\theta(\theta)} < \frac{M_1}{C_N} = M_2.$$

Substituir $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ por $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ tem impacto sobre a cota superior M. Precisamos apenas certificar que $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ é positivo para todo θ candidato, caso contrário $C_N < 0$ e a desigualdade entre $\omega_2(\theta)$ e M_2 seria invertida.

Exemplo: Reconsidere $Y_1, Y_2, \cdots, Y_n$ dado θ condicionalmente independentes com $Y_i|\theta \sim \text{Bernoulli}(\theta)$. Temos $W = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, \theta)$ e

- Espaço paramétrico: $\theta \in [0, 1]$.
- Verossimilhança: $f_{W|\theta}(w) = \binom{n}{w}\theta^w(1-\theta)^{n-w}$.
- Distribuição *a priori*: $\theta \sim \text{Beta}(a, b)$.
- Distribuição *a posteriori*: $\theta|W \sim \text{Beta}(a^*, b^*)$, com $a^* = a + w$ e $b^* = b + w + n$.

Assim como ilustrado no estudo do *Importance sampling*, considere a constante normalizadora C_N (destacada em vermelho abaixo) como desconhecida.

$$f_{\theta|W}(\theta|w) = \frac{\Gamma(a^*+b^*)}{\Gamma(a^*)\,\Gamma(b^*)}\,\theta^{a^*-1}\,(1-\theta)^{b^*-1} = C_N\,\theta^{a^*-1}\,(1-\theta)^{b^*-1}.$$

$$\pi_{\theta|W}(\theta|w) = \theta^{a^*-1}\,(1-\theta)^{b^*-1}.$$

Objetivo: usar o método da rejeição para avaliar a distribuição *a posteriori* sem conhecer sua forma fechada $f_{\theta|W}(\theta|w)$. Apenas o núcleo $\pi_{\theta|W}(\theta|w)$ está disponível.

Admita que:

- $\theta \sim \text{Beta}(a = 1, b = 1)$.
- A coleta de dados estabeleceu: $n = 10$ e $w = 3$.
- Perante as configurações acima temos: $a^* = 4$ e $b^* = 14$.
- Distribuição geradora auxiliar: $g_\theta(\theta)$ é a f.d.p. da $U(0, 1)$.

A *importance ratio* será: $\omega_2(\theta) = \dfrac{\pi_{\theta|W}(\theta|w)}{g_\theta(\theta)} = \dfrac{\theta^{a^*-1}(1-\theta)^{b^*-1}}{1}$.

Veja que o numerador é sempre positivo para todo $\theta \in (0, 1)$. Este ponto é importante para manter a constante $M_2 = M_1/C_N$ como um limite superior.

Estratégia para escolher M_2: avalie $\omega_2(\theta)$ para diversos valores do espaço paramétrico $[0, 1]$. Selecione M_2 igual ao maior $\omega_2(\theta)$ registrado.

Código R aplicando o método da rejeição no exemplo Binomial:

```
a = 1; b = 1; n = 10; w = 3;
as = a + w; bs = b + w + n;

theta = seq(0.01,0.99,0.01)
pi_theta = theta^(as-1) * (1-theta)^(bs-1) # avalie o núcleo
w_theta = pi_theta / dunif(theta,0,1)      # importance ratio
M2 = max(w_theta)                          # defina M2

n_samp = 1000          # tamanho amostral desejado
samp = rep(0,n_samp)   # vetor para salvar amostras
i = 1                  # inicialize contador

repeat{
  theta = runif(1, 0, 1)                       # gere candidato
  pi_theta = theta^(as-1) * (1-theta)^(bs-1) # avalie o núcleo
  w_theta = pi_theta / dunif(theta,0,1)        # importance ratio
  if(runif(1,0,1) < w_theta/M2){               # teste de aceitação
    samp[i] = theta; i = i+1;                  # salve theta e aumente o contador
  }
  if(i > n_samp){ break } # pare o algorítmo ao atingir o tamanho amostral
}
```

Para calcular média e variância *a posteriori* use os comandos

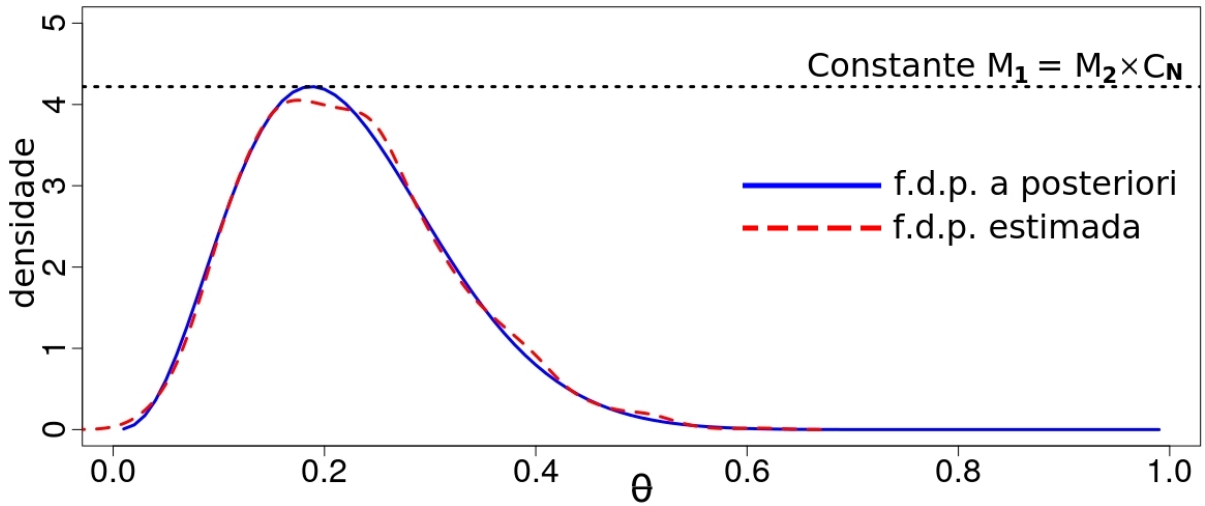
`med = mean(samp)` e `v = var(samp)`

A média e variância *a posteriori* reais são

$\frac{a^*}{a^*+b^*} = 0.2222$ e $\frac{a^* b^*}{(a^*+b^*)^2(a^*+b^*+1)} = 0.0091.$

O método estimou média `med = 0.2243` e variância `v = 0.0090`.

A figura abaixo compara a f.d.p. *a posteriori* real (azul) e a densidade estimada (vermelho trecejado) via método da rejeição. Use o comando `density(samp)` para traçar a curva vermelha.



Exercício 1: Siga o raciocínio deste exemplo Binomial, suponha desconhecimento da forma fechada da f.d.p. *a posteriori*, e implemente o método da rejeição para gerar 1000 observações provenientes da distribuição *a posteriori* de θ. Admita $Y_1, Y_2, \cdots, Y_n$ dado θ i.i.d. tal que:

- $Y_i|\theta \sim N(\theta, \sigma^2)$ com σ^2 conhecido.
- $Y_i|\theta \sim N(\mu, \theta)$ com μ conhecido.

Exercício 2: O método da rejeição foi explicado no contexto de geração de amostras a partir de uma distribuição *a posteriori* sem forma fechada. Entretanto, ele também é uma alternativa que usa uma distribuição auxiliar $g_\theta(\theta)$ para simular valores de uma distribuição com forma fechada $f_\theta(\theta)$. Suponha que a distribuição alvo é a Cauchy com f.d.p. $f_\theta(\theta) = \pi/(1+\theta^2)$. Use o método da rejeição para gerar uma amostra de tamanho 1000 da Cauchy. Assuma $g_\theta(\theta)$ sendo

- f.d.p. da $N(0, 1)$.
- f.d.p. da $N(0, 10)$

5.4 - Metropolis-Hastings (MH)

O Metropolis-Hastings (MH) é outro método que objetiva gerar amostras a partir de uma distribuição de probabilidade sem forma fechada.

A abordagem também possui um passo de aceitação/rejeição, porém a construção do algorítmo é mais sofisticada do que aquela discutida para o método da rejeição.

O MH é amplamente utilizado na literatura relacionada à modelagem Bayesiana. Ele é membro de uma classe de métodos computacionais conhecida como **Markov Chain Monte Carlo (MCMC)**.

Resumidamente, algorítmos MCMC funcionam em ciclos (iterações). Em cada iteração, um novo valor é amostrado tomando como base a última observação simulada no ciclo anterior. Essa dependência sequencial interligando a atual observação com a próxima geração estabelece a configuração de uma cadeia de Markov (Markov Chain).

Originalmente, o algorítmo foi nomeado em referência ao físico Greco-Americano Nicholas Metropolis ($1915^{\star}$-$1999^{\dagger}$), o qual publicou a ideia no artigo *Metropolis N, Rosenbluth A, Teller M, Teller E (1953) Equations of state calculations by fast computing machines. Journal of Chemistry and Physics, 21, 1087-1091*.

O artigo acima apresentou uma forma particular do método focada em distribuições geradoras de candidatos com argumentos simétricos. Nessa versão o algorítmo era chamado simplesmente de Metropolis.

O pesquisador Canadense Wilfred. K. Hastings ($1930^{\star}$-$2016^{\dagger}$) foi quem estendeu o algorítmo para o caso mais geral na publicação *Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 57, 1, 97-109*. Este caso geral passou a ser chamado de Metropolis-Hastings.

Ao contrário do *importance sampling* e do método da rejeição, o MH consegue lidar melhor com a amostragem indireta em problemas com alta dimensão (modelos probabilísticos com muitos parâmetros).

A estrutura do MH é mostrada a seguir.

Seja $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ a f.d.p. ou f.m.p. *a posteriori* a partir da qual deseja-se gerar amostras. A forma fechada de $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ não é conhecida (falta a constante normalizadora). Adote o cenário q-dimensional com $\theta = (\theta_1, \theta_2, \cdots, \theta_q)^\top$.

Através do Teorema de Bayes iremos obter o núcleo $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ que é o trecho conhecido da distribuição alvo.

Denote:

- $\theta^{(j)}$ é o valor do vetor θ na iteração j do algorítmo.
- θ^* é um vetor candidato para θ, o qual será gerado a partir de uma distribuição auxiliar com f.d.p. ou f.m.p. $g_{\theta|\mu,\nu}(\theta)$.
- $g_{\theta|\theta^{(j)},\nu}(\theta^*)$ é o valor da f.d.p. obtido com $\mu = \theta^{(j)}$ e avaliada no ponto θ^*.
- $g_{\theta|\theta^*,\nu}(\theta^{(j)})$ é o valor da f.d.p. obtido com $\mu = \theta^*$ e avaliada no ponto $\theta^{(j)}$.
- $g_{\theta|\mu,\nu}(\theta)$ estabelece a distribuição geradora de propostas (candidatos). Ela é indexada por ν, o qual é um escalar a ser escolhido pelo analista. O valor de ν afeta a taxa de rejeição/aceitação do método.

O Metropolis-Hastings tem os seguintes passos:

1. Faça $j = 0$ e escolha uma semente $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_q^{(0)})^\top$ para iniciar.
2. Gere θ^* a partir de $g_{\theta|\theta^{(j)},\nu}(\theta)$.
3. Calcule a razão $R_1 = \dfrac{\pi_{\theta|\boldsymbol{Y}}(\theta^*|\boldsymbol{y})}{\pi_{\theta|\boldsymbol{Y}}(\theta^{(j)}|\boldsymbol{y})}$.

   Por questão numérica use o log: $\ln(R_1) = \ln[\pi_{\theta|\boldsymbol{Y}}(\theta^*|\boldsymbol{y})] - \ln[\pi_{\theta|\boldsymbol{Y}}(\theta^{(j)}|\boldsymbol{y})]$.
4. Calcule a razão $R_2 = \dfrac{g_{\theta|\theta^*,\nu}(\theta^{(j)})}{g_{\theta|\theta^{(j)},\nu}(\theta^*)}$.

   Escala log: $\ln(R_2) = \ln[g_{\theta|\theta^*,\nu}(\theta^{(j)})] - \ln[g_{\theta|\theta^{(j)},\nu}(\theta^*)]$.
5. Calcule a probabilidade de aceitação: $\alpha = \min\{1\ ,\ R1 \times R2\}$.

   Escala log: $\ln(\alpha) = \min\{0\ ,\ \ln(R_1) + \ln(R_2)\}$ e $\alpha = \exp\{\ln(\alpha)\}$.
6. Gere $\mathcal{U} \sim U(0,1)$ e avalie:

   Se $\mathcal{U} \leq \alpha$, faça $\theta^{(j+1)} = \theta^*$. Caso contrário, faça $\theta^{(j+1)} = \theta^{(j)}$.
7. Faça $j = j + 1$ e retorne ao Passo (2) até obter o tamanho amostral desejado.

O cálculo da razão R_2 (Passo 4) é o aspecto que diferencia a versão particular Metropolis do caso geral Metropolis-Hastings.

Uma escolha popular para $g_{\theta|\mu,\nu}(\theta)$ é a f.d.p. da $N_q(\mu, \nu \boldsymbol{I}_q)$, ou seja, um candidato θ^* é gerado da Normal q-variada com vetor de médias μ e matriz de covariância $\nu \boldsymbol{I}_q$ ($\boldsymbol{I}_q$ = matriz identidade $q \times q$).

A opção acima é geralmente chamada de **passeio aleatório** ou **random walk**, pois o candidato θ^* é gerado na vizinhança de $\theta^{(j)}$. Este esquema tende a gerar uma trajetória que segue uma direção com oscilações frequentes para as laterais (analogia: um bêbado caminhando).

Para estabelecer a dependência Markoviana no algorítmo, adota-se $\mu = \theta^{(j)}$. Isto significa que o candidato θ^* será gerado em uma região ao redor do ponto $\theta^{(j)}$ obtido na iteração j do algorítmo.

Note que: $R_2 = \dfrac{g_{\theta|\theta^*,\nu}(\theta^{(j)})}{g_{\theta|\theta^{(j)},\nu}(\theta^*)} =$

$$= \frac{(2\pi)^{-q/2}\ |\nu \boldsymbol{I}_q|^{-1/2}\ \exp\{-\frac{1}{2\nu}(\theta^{(j)} - \theta^*)^\top\ \boldsymbol{I}_q^{-1}\ (\theta^{(j)} - \theta^*)\}}{(2\pi)^{-q/2}\ |\nu \boldsymbol{I}_q|^{-1/2}\ \exp\{-\frac{1}{2\nu}(\theta^* - \theta^{(j)})^\top\ \boldsymbol{I}_q^{-1}\ (\theta^* - \theta^{(j)})\}} = 1$$

Quando escolhemos a geradora de propostas do tipo *random walk* (Normal), o numerado e o denominador da razão R_2 serão iguais. Esta simplificação será obtida para qualquer distribuição simétrica em seus argumentos (este é o caso da Normal). No artigo de 1953, Nicholas Metropolis e coautores apresentaram esta versão particular que recebeu o nome de método Metropolis.

Em 1970, Wilfred Hastings provou que não somos obrigados a usar geradoras de propostas simétricas. Podemos escolher outras distribuições, tomando sempre o cuidado de calcular a razão R_2.

Perceba que a escolha de ν indica a liberdade que o gerador de propostas tem para sugerir valores longe do centro $\mu = \theta^{(j)}$.

- ν grande $\Rightarrow$ candidados θ^* longe de $\theta^{(j)}$.
  O algorítmo é propenso a rejeitar valores muito distantes de $\theta^{(j)}$, então ν grande tende a reduzir a taxa de aceitação.
- ν pequeno $\Rightarrow$ candidados θ^* perto de $\theta^{(j)}$.
  O algorítmo é propenso a aceitar valores próximos de $\theta^{(j)}$, então ν pequeno tende a aumentar a taxa de aceitação.

A escolha de ν deve ser feita com base em testes exploratórios (*tunagem*) do algorítmo. Recomenda-se rodar o MH algumas vezes aplicando diferentes valores de ν. Verifique a taxa de aceitação (n^o aceitações / n^o de iterações). Ela não deve ser muito alta (algorítmo pouco criterioso que aceita qualquer valor) e nem muito baixa (cadeia sem movimentação).

Na literatura existem trabalhos que recomendam taxas de aceitação do MH em torno de 40% a 60%.

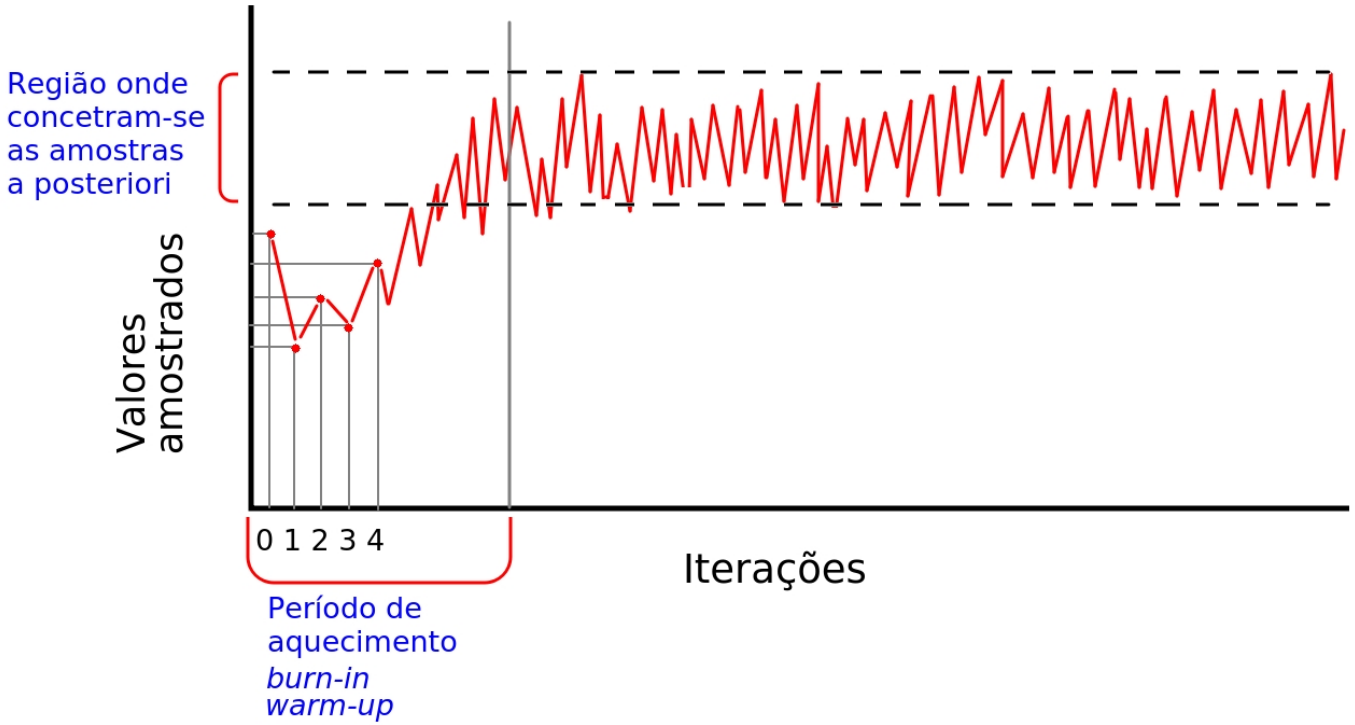
O MH inicia com valores arbitrários para θ (semente $\theta^{(0)}$). A cada ciclo, um candidato θ^* é proposto e testado. Se houver aceitação, θ^* é salvo. Se houver rejeição, θ^* é descartado e o atual valor $\theta^{(j)}$ é mantido.

Esta sequência de testes com inclusão/permanência de valores possibilita criar vetores que guardam o histórico amostral de cada parâmetro θ_k em θ. Estes vetores são chamados de **cadeias**.

No início as cadeias estão muito associadas ao chute inicial $\theta^{(0)}$, entretanto, a teoria relacionada ao MH garante que, após um certo número de iterações, as cadeias irão **convergir** (estabilizar) para uma região do espaço paramétrico no qual a massa de probabilidade *a posteriori* está concentrada.

A partir do momento em que a **convergência** das cadeias ocorrer, cada novo valor θ^* aceito dentro do MH será tratado como proveniente da distribuição alvo $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Antes da convergência, isso não é verdade.

As amostras de θ salvas durante o período inicial de execução do algorítmo (antes da convergência) devem ser eliminadas. Estas amostras formam o **período de aquecimento da cadeia** (*burn-in* ou *warm-up*) em que as observações ainda não são da distribuição alvo.



O usuário define o n^o de iterações N que deseja executar o MH. Esta escolha deve considerar que haverá um *burn-in* incluindo N_* iterações iniciais. A amostra *a posteriori* é formada pelas $N - N_*$ iterações finais.

Exemplo: Considere $Y_i \sim N(\mu, 1/\phi)$, com $\{\mu, \phi\}$ desconhecidos. Admita independência condicional de $Y_1, Y_2, \cdots, Y_n$ dado $\{\mu, \phi\}$.

- Espaço paramétrico: $\mu \in \mathbb{R}$ e $\phi \in \mathbb{R}^+$.
- Independência *a priori*: $f_{\mu,\phi}(\mu, \phi) = f_\mu(\mu)\ f_\phi(\phi)$.
  $\mu \sim N(m, v)$ e $\phi \sim \text{Ga}(a, b)$.
- Verossimilhança: $(2\pi)^{-n/2}\ \phi^{n/2}\ \exp\{-\frac{\phi}{2}(\sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2)\}$.
- $\pi_{\mu,\phi|\boldsymbol{Y}}(\mu, \phi|\boldsymbol{y}) \propto (v_\phi^*)^{-1/2}\ \exp\left\{-\frac{1}{2v_\phi^*}\ (\mu - m_\phi^*)^2\right\} \times$

  $\times\ (v_\phi^*)^{1/2} \exp\left\{\frac{(m_\phi^*)^2}{2v_\phi^*}\right\} \times \phi^{\tilde{a}-1}\ \exp\{-\phi\tilde{b}\}$,

  com $v_\phi^* = (n\phi + \frac{1}{v})^{-1}$, $m_\phi^* = v_\phi^*\ (\phi \sum_{i=1}^n y_i + \frac{m}{v})$,
  $\tilde{a} = a + n/2$ e $\tilde{b} = b + \sum_{i=1}^n y_i^2/2$.

A forma fechada de $f_{\mu,\phi|\boldsymbol{Y}}(\mu, \phi|\boldsymbol{y})$ não é conhecida. Usaremos o MH para amostragem indireta.

Código R preparatório para executar o MH:

```
# gerando dados artificiais.
n = 100          # tamanho amostral.
mu_real = 10     # mu real.
phi_real = 2     # phi real.
y = rnorm(n,mu_real,sqrt(1/phi_real))

# especificações a priori:  mu ~ N(m,v) e  phi ~ Ga(a,b).
m = 5;  v = 10;  a = 0.1;   b = 0.1;

# semente de inicialização das cadeias MCMC.
mu = 5; phi = 1;

# variância da distribuição geradora de propostas.
nu = 0.01  # cadidato gerado da N(valor_atual, nu).

# alguns termos presentes no núcleo a posteriori.
as = a + n/2;  bs = b + sum(y^2)/2;

N = 10000  # número de iterações do MH.

# objetos auxiliares para salvar cadeias e taxa de aceitação.
save_mu = mu;  save_phi = phi;   ARate = 0;
```

Código R com os passos do MH:

```
for(j in 1:N){
  mu_c = rnorm(1,mu,sqrt(nu))    # gere o candidado mu.
  phi_c = rnorm(1,phi,sqrt(nu))  # gere o candidado phi.

  # se phi_c < 0 ou phi_c = 0, rejeite mu_c e phi_c e siga para a iteração j+1

  if(phi_c > 0){
   # calcule termos presente no núcleo a posteriori.
   vs = 1/(n*phi + 1/v);        ms = vs*(phi*sum(y) + m/v);
   vs_c = 1/(n*phi_c + 1/v);    ms_c = vs_c*(phi_c*sum(y) + m/v);
   #
   # insira aqui o cálculo de log(R1) e log(R2) mostrado no próximo slide.
   #
   alpha = exp(min(0,(logR1+logR2))) # prob. de aceitação
   # teste de aceitação/rejeição.
   if(runif(1,0,1) < alpha){ mu = mu_c;  phi = phi_c; ARate = ARate + 1 }
  }

  # salvando cadeia.
  save_mu = c(save_mu, mu)
  save_phi = c(save_phi, phi)
}
```

Código R com o cálculo de $\ln(R_1)$ e $\ln(R_2)$ no MH:

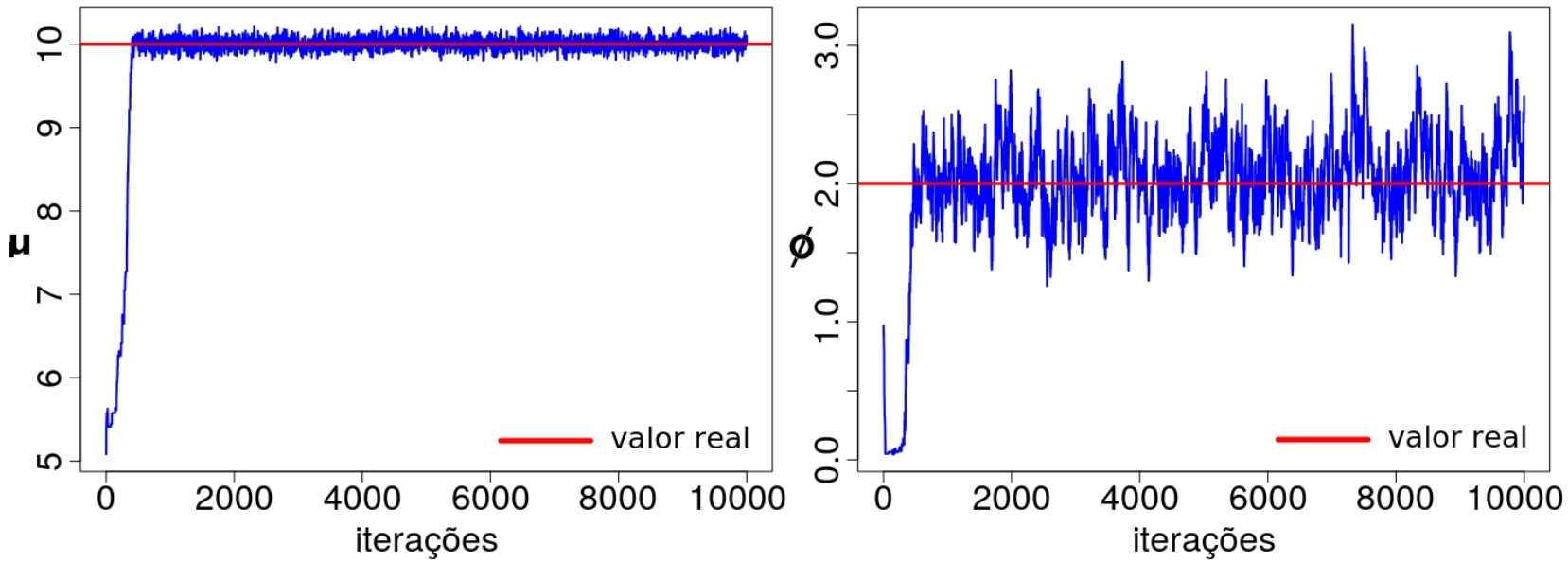
```
# calcule a razão R1 (escala log).
logR1_1 = -(0.5/vs_c) * ((mu_c - ms_c)^2) +(ms_c^2) / (2 * vs_c)
          +(as - 1) * log(phi_c) -phi_c * bs
logR1_2 = -(0.5 / vs) * ((mu - ms)^2) +(ms^2) / (2 * vs)
          +(as - 1) * log(phi) -phi * bs
logR1 = logR1_1 - logR1_2

# calcule a razão R2 (escala log).
dp = sqrt(nu)
logR2_1 = dnorm(mu, mu_c, dp, log=TRUE) + dnorm(phi, phi_c, dp, log=TRUE)
logR2_2 = dnorm(mu_c, mu, dp, log=TRUE) + dnorm(phi_c, phi, dp, log=TRUE)
logR2 = logR2_1 - logR2_2
```

A conta de R_2 é desnecessária perante a geração de propostas via *random walk*. Note também que a f.d.p. da normal bivariada, com matriz de covariâncias $\nu \times \boldsymbol{I}_2$, equivale ao produto de 2 normais univariadas com variância ν.

Uma alternativa ao *random walk* para gerar candidatos é $\phi^* \sim \text{Ga}(\phi^{(j)}\kappa, \kappa)$. Teríamos $E(\phi^*) = \phi^{(j)}\kappa/\kappa = \phi^{(j)}$ e $Var(\phi^*) = \phi^{(j)}\kappa/\kappa^2 = \phi^{(j)}/\kappa$. O parâmetro κ seria submetido à *tunagem* (rever slides) para ser escolhido.

A figura abaixo mostra as cadeias obtidas via MH para o exemplo Normal com média μ e precisão ϕ desconhecidos (independência *a priori* entre os dois parâmetros).

A linha horizontal vermelha destaca o valor real do parâmetro (usado para gerar os dados artificiais). A taxa de aceitação ficou em 56.52%.



Exercício: Execute o algorítmo MH apresentado aqui para o caso Normal com μ e ϕ desconhecidos (independência *a priori* entre os parâmetros). Use as mesmas especificações de: geração de dados, distribuições *a priori*, sementes e número de iterações. A única modificação a ser feita é na escolha do parâmetro ν. Avalie o resultado das seguintes opções: $\nu = 0.5$, $\nu = 1$ e $\nu = 10$. Em cada avaliação, faça o gráfico das cadeias de μ e σ^2. Calcule também as taxas de aceitação do MH. Comente o comportamento das cadeias e os valores das taxas.

5.5 - Gibbs Sampling (GS)

O nome *Gibbs Sampling* (GS) faz referência ao físico Americano Josiah Willard Gibbs (1839*-1903†) que em estudos de física-estatística abordou alguns conceitos matemáticos úteis para desenvolver mais tarde o amostrador que iremos aprender.

O método GS foi apresentado pelos irmãos Americanos Stuart Geman (1949*) e Donald Geman (1943*) na publicação *Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 6, 721-741*.

Apesar dos irmãos Geman terem sido pioneiros na apresentação do GS, a constatação de que o método poderia ser usado amplamente para amostrar de distribuições *a posteriori* foi divulgada pelo Americano Alan Gelfand (1945*) e o Inglês Adrian Smith (1946*) no artigo *Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities, Journal of the American Statistical Association, 85, 410, 398-409*. Este artigo revolucionou a análise de dados permitindo abordar problemas que antes eram intratatáveis.

O GS é considerado um método geral que permite amostragem indireta de distribuições com alta dimensão (muitos parâmetros). O aspecto principal do algorítmo é quebrar a amostragem conjunta em amostragens individuais ou em blocos de parâmetros.

O objetivo é usar o GS para gerar valores da distribuição $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$, a qual não tem forma fechada. Admita que

$$\theta = (\theta_1, \theta_2, \cdots, \theta_q)^\top$$

é o vetor de parâmetros do modelo estatístico. Então $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ representa uma distribuição conjunta *a posteriori* de dimensão q.

Denote: $\theta_{-k} = (\theta_1, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_q)^\top$, ou seja,
θ_{-k} é o vetor θ sem o elemento θ_k.

A f.d.p. ou f.m.p. $f_{\theta_k|\theta_{-k},\boldsymbol{Y}}(\theta_k|\theta_{-k},\boldsymbol{y})$ indica a distribuição **condicional completa a posteriori** de θ_k.

O GS usa as distribuições condicionais completas para realizar as amostragens individuais (ou em blocos) de cada parâmetro. Note que $f_{\theta_k|\theta_{-k},\boldsymbol{Y}}(\theta_k|\theta_{-k},\boldsymbol{y})$ estabelece a geração de um valor θ_k condicionado (sob influência) dos demais parâmetros em θ_{-k}.

Pode ser mostrado que as observações simuladas nesta estrutura formam uma cadeia de Markov. Partindo de um chute inicial (semente $\theta_k^{(0)}$), a cadeia passará por um período de aquecimento (*burn-in*) e após algumas iterações do GS, ela irá convergir para a região paramétrica onde a distribuição *a posteriori* coloca mais massa de probabilidade para θ_k.

Assim como o Metropolis-Hastings, o GS também é um método de amostragem indireta classificado como membro da classe de algorítmos MCMC.

As amostras coletadas após o *burn-in* para cada θ_k serão de fato amostras provenientes da distribuição conjunta $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$. Todas as observações pertencentes ao período de *burn-in* devem ser descartadas.

A aplicação do GS envolverá os seguinte aspectos:

- $f_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ é a distribuição conjunta dos q parâmetros em θ. Sua forma fechada é desconhecida.
- $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ é o núcleo *a posteriori* obtido via Teorema de Bayes.
- Use $\pi_{\theta|\boldsymbol{Y}}(\theta|\boldsymbol{y})$ para calcular $f_{\theta_k|\theta_{-k},\boldsymbol{Y}}(\theta_k|\theta_{-k},\boldsymbol{y})$. Apesar da conjunta não ter forma fechada, a conta da condicional completa é em geral mais fácil (em muitos casos será possível identificar tal distribuição).
- Se a forma fechada de alguma condicional completa não puder ser determinada, o GS permite que a amostragem do parâmetro θ_k em questão seja feita com o auxílio de outro algorítmo (MH por exemplo).

Os passos do GS são detalhados a seguir.

Passos do *Gibbs sampling*:

1. Faça $j = 0$ e escolha uma semente $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_q^{(0)})^\top$.
2. Gere $\theta_1^{(j+1)}$ de $f_{\theta_1|\theta_{-1},\boldsymbol{Y}}(\theta_1|\theta_2^{(j)}, \theta_3^{(j)}, \theta_4^{(j)}, \cdots, \theta_q^{(j)}, \boldsymbol{y})$.
3. Gere $\theta_2^{(j+1)}$ de $f_{\theta_2|\theta_{-2},\boldsymbol{Y}}(\theta_2|\theta_1^{(j+1)}, \theta_3^{(j)}, \theta_4^{(j)}, \cdots, \theta_q^{(j)}, \boldsymbol{y})$.
4. Gere $\theta_3^{(j+1)}$ de $f_{\theta_3|\theta_{-3},\boldsymbol{Y}}(\theta_3|\theta_1^{(j+1)}, \theta_2^{(j+1)}, \theta_4^{(j)}, \cdots, \theta_q^{(j)}, \boldsymbol{y})$.
5. Gere $\theta_4^{(j+1)}$ de $f_{\theta_4|\theta_{-4},\boldsymbol{Y}}(\theta_4|\theta_1^{(j+1)}, \theta_2^{(j+1)}, \theta_3^{(j+1)}, \cdots, \theta_q^{(j)}, \boldsymbol{y})$.
6. Siga a lógica acima para gerar $\theta_5^{(j+1)}, \cdots, \theta_{q-1}^{(j+1)}$.
7. Gere $\theta_q^{(j+1)}$ de $f_{\theta_q|\theta_{-q},\boldsymbol{Y}}(\theta_q|\theta_1^{(j+1)}, \theta_2^{(j+1)}, \theta_3^{(j+1)}, \cdots, \theta_{q-1}^{(j+1)}, \boldsymbol{y})$.
8. Faça $j = j + 1$ e retorne ao Passo (2) até obter o tamanho amostral desejado.

Exemplo: Considere novamente $Y_i \sim N(\mu, 1/\phi)$, com $\{\mu, \phi\}$ desconhecidos. Admita independência condicional de $Y_1, Y_2, \cdots, Y_n$ dado $\{\mu, \phi\}$.

- Espaço paramétrico: $\mu \in \mathbb{R}$ e $\phi \in \mathbb{R}^+$.
- Independência *a priori*: $f_{\mu,\phi}(\mu, \phi) = f_\mu(\mu)\, f_\phi(\phi)$.
  $\mu \sim N(m, v)$ e $\phi \sim \text{Ga}(a, b)$.
- Verossimilhança: $(2\pi)^{-n/2}\, \phi^{n/2}\, \exp\{-\frac{\phi}{2}(\sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2)\}$.
- $\pi_{\mu,\phi|\boldsymbol{Y}}(\mu, \phi|\boldsymbol{y}) \propto (v_\phi^*)^{-1/2}\, \exp\left\{-\frac{1}{2v_\phi^*}\,(\mu - m_\phi^*)^2\right\} \times$

  $\times\, (v_\phi^*)^{1/2} \exp\left\{\frac{(m_\phi^*)^2}{2v_\phi^*}\right\} \times \phi^{\tilde{a}-1}\, \exp\{-\phi\tilde{b}\}$,

  com $v_\phi^* = (n\phi + \frac{1}{v})^{-1}$, $m_\phi^* = v_\phi^*\,(\phi \sum_{i=1}^n y_i + \frac{m}{v})$,
  $\tilde{a} = a + n/2$ e $\tilde{b} = b + \sum_{i=1}^n y_i^2/2$.

A forma fechada de $f_{\mu,\phi|\boldsymbol{Y}}(\mu, \phi|\boldsymbol{y})$ não é conhecida.

Usaremos o GS para amostragem indireta.

Aplique o Teorema de Bayes para determinar a distribuição condicional completa *a posteriori* de $\mu|\phi, \boldsymbol{Y}$. Note nas contas abaixo que ϕ sempre será tratado como condicional.

$$f_{\mu|\phi,\boldsymbol{Y}}(\mu|\phi,\boldsymbol{y}) \;=\; \frac{f_{\mu,\boldsymbol{Y}|\phi}(\mu,\boldsymbol{y}|\phi)}{f_{\boldsymbol{Y}|\phi}(\boldsymbol{y}|\phi)} \;=\; \frac{f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y}|\mu,\phi) \;\times\; f_{\mu|\phi}(\mu|\phi)}{\int_{-\infty}^{\infty} f_{\mu,\boldsymbol{Y}|\phi}(\mu,\boldsymbol{y}|\phi)\; d\mu}.$$

Por independência *a priori* temos:

$$f_{\mu|\phi}(\mu|\phi) \;=\; f_{\mu}(\mu) \;=\; (2\pi v)^{-1/2}\; \exp\{-\tfrac{1}{2v}(\mu - m)^2\}.$$

A verossimilhança $f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y}|\mu,\phi)$ está escrita no último slide. Lembre-se também que a integral no denominador é uma constante normalizadora.

$$f_{\mu|\phi,\boldsymbol{Y}}(\mu|\phi,\boldsymbol{y}) \;\propto\; f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y}|\mu,\phi) \;\times\; f_{\mu}(\mu)$$

$$\propto\; (2\pi)^{-n/2}\; \phi^{n/2}\; \exp\{-\tfrac{\phi}{2}(\textstyle\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2)\} \;\times$$

$$\times\; (2\pi v)^{-1/2}\; \exp\{-\tfrac{1}{2v}(\mu^2 - 2\mu m + m^2)\}$$

Visto que μ é o elemento de interesse nesta conta, o próximo passo é incorporar na proporcionalidade todo termo que não está atrelado a μ. Tais termos estão destacados em vermelho abaixo:

$$\propto \color{red}{(2\pi)^{-n/2}\,\phi^{n/2}\,\exp\{-\tfrac{\phi}{2}\textstyle\sum_{i=1}^{n} y_i^2\}}\;\exp\{-\tfrac{\phi}{2}\left[n\mu^2 - 2\mu\textstyle\sum_{i=1}^{n} y_i\right]\}\times$$
$$\times\;\color{red}{(2\pi v)^{-1/2}}\;\exp\{-\tfrac{1}{2v}\left[\mu^2 - 2\mu m\right]\}\;\color{red}{\exp\{-\tfrac{1}{2v}m^2\}}$$

$$\propto \exp\{-\tfrac{\phi}{2}\left[n\mu^2 - 2\mu\textstyle\sum_{i=1}^{n} y_i\right]\}\;\times\;\exp\{-\tfrac{1}{2v}\left[\mu^2 - 2\mu m\right]\}$$

$$\propto \exp\left\{-\tfrac{1}{2}\left[n\phi\mu^2 + \tfrac{1}{v}\mu^2 - 2\mu\phi\textstyle\sum_{i=1}^{n} y_i - 2\mu\tfrac{m}{v}\right]\right\}$$

$$\propto \exp\left\{-\tfrac{1}{2}\left[\mu^2\left(n\phi + \tfrac{1}{v}\right) - 2\mu\left(\phi\textstyle\sum_{i=1}^{n} y_i + \tfrac{m}{v}\right)\right]\right\}$$

$$\propto \exp\left\{-\tfrac{1}{2}\left(n\phi + \tfrac{1}{v}\right)\left[\mu^2 - 2\mu\left(n\phi + \tfrac{1}{v}\right)^{-1}\left(\phi\textstyle\sum_{i=1}^{n} y_i + \tfrac{m}{v}\right)\right]\right\}.$$

A expressão acima pode ser reconhecida como o núcleo da distribuição Normal com variância $v_\phi^* = (n\phi + \frac{1}{v})^{-1}$ e média $m_\phi^* = v_\phi^*\,(\phi\sum_{i=1}^{n} y_i + \frac{m}{v})$.

Conclusão: $\mu|\phi, \boldsymbol{Y} \sim N(m_\phi^*, v_\phi^*)$.

Aplique o Teorema de Bayes para determinar a distribuição condicional completa *a posteriori* de $\phi|\mu, \boldsymbol{Y}$. Nas contas abaixo, μ sempre será tratado como condicional.

$$f_{\phi|\mu,\boldsymbol{Y}}(\phi|\mu,\boldsymbol{y}) = \frac{f_{\phi,\boldsymbol{Y}|\mu}(\phi,\boldsymbol{y}|\mu)}{f_{\boldsymbol{Y}|\mu}(\boldsymbol{y}|\mu)} = \frac{f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y}|\mu,\phi) \times f_{\phi|\mu}(\phi|\mu)}{\int_0^\infty f_{\phi,\boldsymbol{Y}|\mu}(\phi,\boldsymbol{y}|\mu)\, d\phi}.$$

Por independência *a priori* temos:

$$f_{\phi|\mu}(\phi|\mu) = f_\phi(\phi) = \frac{b^a}{\Gamma(a)}\, \phi^{a-1}\, \exp\{-b\phi\}.$$

Novamente perceba que a integral no denominador é uma constante normalizadora.

$$f_{\phi|\mu,\boldsymbol{Y}}(\phi|\mu,\boldsymbol{y}) \propto f_{\boldsymbol{Y}|\mu,\phi}(\boldsymbol{y}|\mu,\phi) \times f_\phi(\phi)$$

$$\propto (2\pi)^{-n/2}\, \phi^{n/2}\, \exp\{-\tfrac{\phi}{2}(\textstyle\sum_{i=1}^n y_i^2 - 2\mu \sum_{i=1}^n y_i + n\mu^2)\} \times$$

$$\times \frac{b^a}{\Gamma(a)}\, \phi^{a-1}\, \exp\{-b\phi\}$$

Agora ϕ é o elemento de interesse. O próximo passo é incorporar na proporcionalidade todo termo que não está atrelado a ϕ. Tais termos estão destacados em vermelho abaixo:

$$\propto \ (2\pi)^{-n/2}\ \phi^{n/2}\ \exp\{-\phi\ \tfrac{1}{2}(\textstyle\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2)\} \times$$

$$\times\ \frac{b^a}{\Gamma(a)}\ \phi^{a-1}\ \exp\{-b\phi\}$$

$$\propto\ \phi^{n/2}\ \exp\{-\phi\ \tfrac{1}{2}(\textstyle\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2)\}\ \times\ \phi^{a-1}\ \exp\{-b\phi\}$$

$$\propto\ \phi^{a+(n/2)-1}\ \exp\{-\phi\ \left[b + \tfrac{1}{2}(\textstyle\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2)\right]\}$$

A expressão acima pode ser reconhecida como o núcleo da distribuição Gama parametrizada por $a^* = a + (n/2)$ e $b_\mu^* = b + \frac{1}{2}(\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2)$.

Conclusão: $\phi|\mu, \boldsymbol{Y} \sim \text{Ga}(a^*, b_\mu^*)$.

As duas f.d.p.'s condicionais completas *a posteriori* possuem forma fechada. A implementação em R do algorítmio GS é mostrado a seguir.

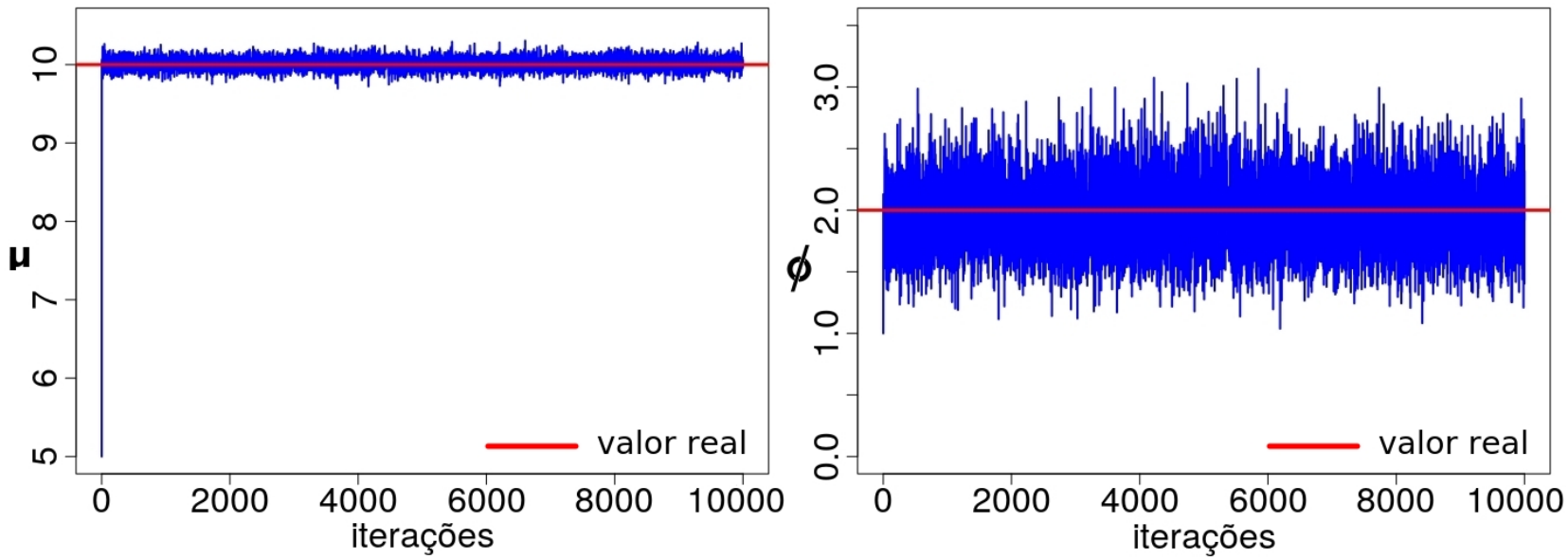
Código preparatório antes do ciclo de iterações:

```
# gerando dados artificiais.
n = 100         # tamanho amostral.
mu_real = 10    # mu real.
phi_real = 2    # phi real.
y = rnorm(n,mu_real,sqrt(1/phi_real))

# especificações a priori.
m = 5; v = 10;      # mu ~ N(m,v).
a = 0.1; b = 0.1;   # phi ~ Ga(a,b).

# semente de inicialização da cadeia MCMC.
mu = 5; phi = 1;

# número total de iterações.
N = 10000

# objetos auxiliares para salvar as cadeias.
save_mu = mu;   save_phi = phi;
```

Código R com os passos do GS:

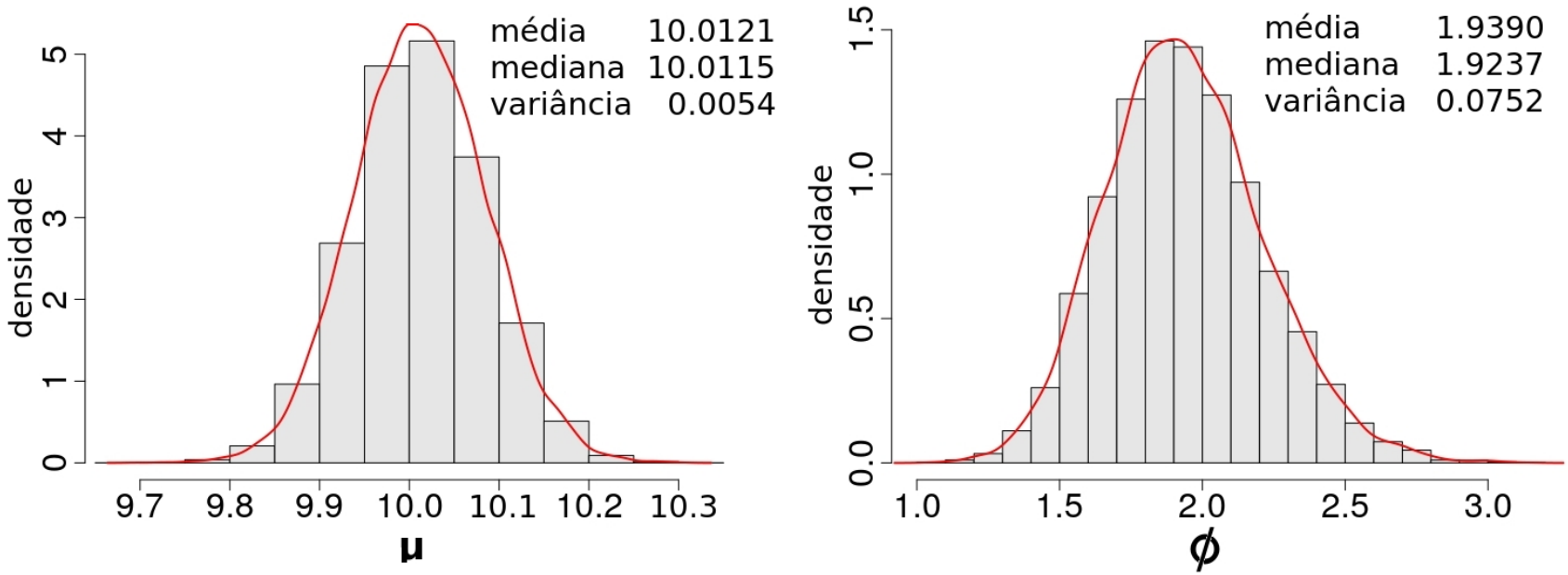
```
for(j in 1:N){

  # concidional completa de mu.
  vs = 1/(n*phi + 1/v)
  ms = vs * (phi*sum(y) + m/v)
  mu = rnorm(1,ms,sqrt(vs))

  # concidional completa de phi.
  as = a + n/2
  bs = b + 0.5*(sum(y^2) -2*mu*sum(y) +n*mu^2)
  phi = rgamma(1,as,rate=bs)

  # salvando cadeia.
  save_mu = c(save_mu, mu)
  save_phi = c(save_phi, phi)

}
```

As cadeias do GS para o exemplo Normal com média μ e precisão ϕ desconhecidos (independência *a priori* entre parâmetros) é mostrada na figura abaixo. A linha horizontal vermelha destaca o valor real do parâmetro (usado para gerar os dados artificiais).



Diferente do Metropolis-Hastings, o GS não possui teste de aceitação/rejeição. Todos os valores amostrados formam a cadeia (não há taxa de aceitação para reportar). Neste exemplo, a convergência foi imediata (ocorreu já na 2^a iteração).

Outro gráfico interessante é o histograma das amostras obtidas após a convergência da cadeia. Assumindo um *burn-in* pequeno (10 primeiras iterações), a figura abaixo exibe a distribuição dos valores amostrados *a posteriori* para μ e ϕ.

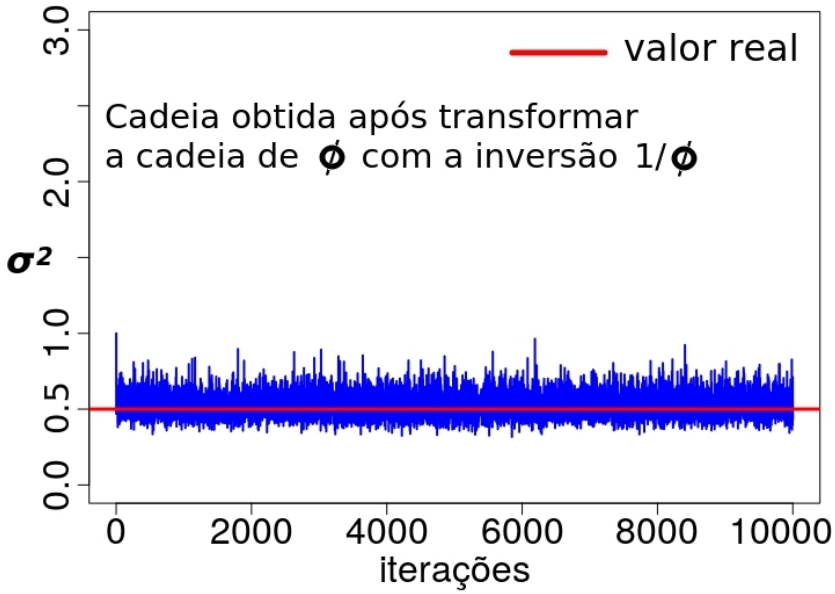


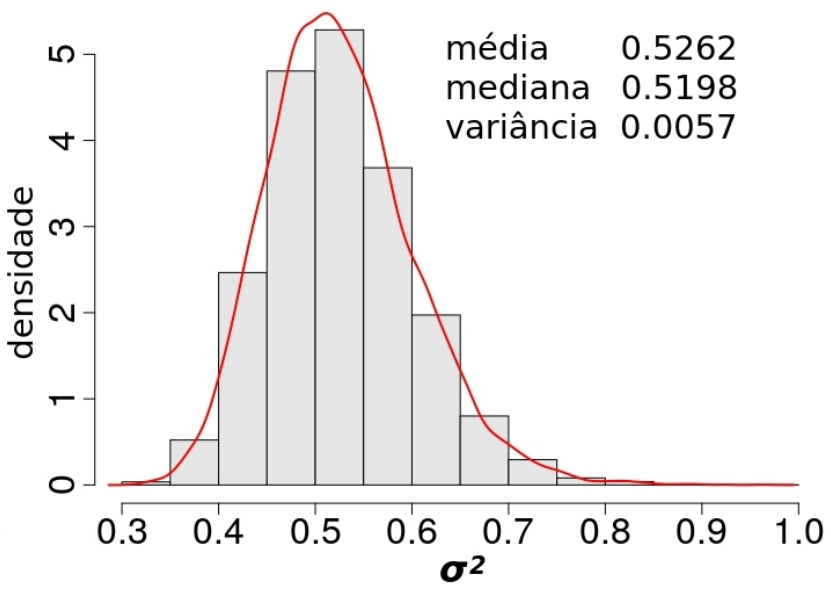
Estas distribuições possuem toda a informação necessária para a realização da inferência com estimação pontual e intervalar a respeito dos parâmetros.

Note que GS para o caso Normal foi implementado para amostrar apenas a média μ e a precisão ϕ. O que fazer se houver interesse pela variância $\sigma^2 = 1/\phi$?

Não precisamos calcular a condicional completa de σ^2 e executar novamente o GS para gerar amostras desse parâmetro.

O analista deve simplesmente aplicar a transformação $\sigma^2 = 1/\phi$ para todos os valores compondo a cadeia de ϕ. O resultado será a cadeia de σ^2. Tal ideia não é exclusiva do GS, vale também para os demais métodos de amostragem indireta.





Um último tópico a ser explorado, nesta aplicação do modelo Normal, é a determinação do intervalo de credibilidade *a posteriori* do tipo HPD (rever slides).

Lembre-se que o intervalo HPD é a estimativa intervalar de menor amplitude. Todos os outros intervalos de credibilidade, de mesma probabilidade, serão maiores.

Para distribuições assimétricas, é necessário aplicar uma rotina computacional para obter o HPD. Uma vez gerada a amostra *a posteriori*, por algum dos métodos estudados, podemos utilizar a função `HPDinterval`, do pacote `R` chamado `coda`, para calcular o intervalo HPD.

Na área de trabalho do `R`, o argumento de entrada da função `HPDinterval` é um objeto do tipo `mcmc`. O usuário do `coda` deve aplicar o comando `as.mcmc(.)` para transformar um vetor (1 cadeia) ou matriz (cadeias nas colunas) em outro do tipo `mcmc` que é reconhecido pelo pacote.

O código R abaixo, mostra o uso do coda para calcular os intervalos HPD de μ e ϕ no exemplo do GS aplicado ao caso Normal.

```
library(coda)

HPDinterval( as.mcmc(save_mu[10:10000]), prob = 0.95 )
HPDinterval( as.mcmc(save_phi[10:10000]), prob = 0.95 )
```

A interpretação do resultado é como segue

- Temos probabilidade 0.95 *a posteriori* de encontrar o verdadeiro μ entre [9.8724 , 10.1580].
- Temos probabilidade 0.95 *a posteriori* de encontrar o verdadeiro ϕ entre [1.4094 , 2.4740].

Exercício: Execute o algorítmio Gibbs Sampling apresentado nestes slides. Gere os dados assumindo os mesmos valores reais para μ e ϕ. Você deverá rodar o método com as seguintes alterações:

- Assuma distribuições *a priori* menos informativas: $\mu \sim N(m = 5, v = 100)$ e $\phi \sim Ga(a = 0.01, b = 0.01)$.
- Adote chutes iniciais para as cadeias mais distantes dos valores reais: $\mu^{(0)} = 50$ e $\phi^{(0)} = 10$.

Calcule os intervalos HPD para μ, ϕ, σ^2 e para $\sqrt{\sigma^2}$. Crie uma tabela contendo nas colunas: o valor real, a média, a moda, a mediana, o desvio padrão e os limites inferior e superior do intervalo HPD para cada um destes parâmetros.

Faça os gráficos das cadeias de μ, ϕ, σ^2 e $\sqrt{\sigma^2}$ conforme apresentado nos slides. Inclua a linha horizontal vermelha demarcando o valor real do parâmetro.

Comentários finais desta seção:

Até o momento estudamos 4 algorítmos que permitem gerar amostras de uma distribuição *a posteriori* desconhecida: *importance sampling*, método da rejeição, Metropolis-Hastings e *Gibbs sampling*.

Os métodos MH e GS são membros da classe de algorítmos MCMC. Eles são mais sofisticados do que o *importance sampling* e o método da rejeição. Uma vantagem do MH e do GS é que podem ser aplicados em modelos com alta dimensão (muitos parâmetros).

Os opções mencionadas acima não são os únicos algorítmos disponíveis na literatura Bayesiana. Na próxima seção deste curso, iremos conhecer os princípios do **Hamiltonian Monte Carlo**. Esta opção também pode ser usada para amostragem indireta e é empregada no *software* `Stan`, que iremos aplicar para desenvolver uma análise Bayesiana para modelos de regressão.