Universidade Federal do Rio Grande
FURG

Djidénou Hans Amos Montcho

# Método de Laplace em estatística bayesiana: uma aproximação para a distribuição posterior em Estatística Bayesiana

Bacharelado em Matemática Aplicada

Rio Grande, Rio Grande do Sul, Brasil
2016

Djidénou Hans Amos Montcho

# Método de Laplace para aproximar Distribuições de Probabilidade a Posteriori em Estatística Bayesiana

Bacharelado em Matemática Aplicada

Monografia submetida por Djidénou Hans Amos Montcho como requisito para obtenção do grau de Bacharel em Matemática Aplicada pelo curso de Matemática Aplicada junto ao Instituto de Matemática, Estatística e Física da Universidade Federal do Rio Grande, sob orientação dos Dr. Paul Gerhard Kinas e Dr. Vanderlei Manica

BANCA EXAMINADORA

---

Dr Paul Gerhard Kinas
(Orientador)

---

Dr Vanderlei Manica
(Co-Orientador)

---

Dra Raquel da Fontoura Nicolette
(Banca)

---

Dr Matheus Jatkoske Lazo
(Banca)

Rio Grande, 02 de dezembro 2016

# Agradecimentos

A Deus

Aos meus Pais, Rosaline GBEHOUN e Fiacre MONTCHO. Sem vocês, eu não estaria aqui e essas linhas nunca teriam existido.

Aos meus irmãos, Gilles, Méléda, Dietrich e Ulrich pelo apoio emocional.

Aos Governos da República do Benin e Federativa do Brasil pelo apoio financeiro durante esses anos de formação.

Aos recursos humanos da Universidade Federal do Rio Grande- FURG, do Instituto de Matemática, Estatística e Física- IMEF especialmente ao professor Dr Mario Retamoso pelos conselhos durante essa caminhada.

Á família Samaniego pela recepção familiar.

A todo o pessoal do laboratório de Estatística Ambiental, Vanderlei meu coorientador, Raquel, Juliano, Liana, Rayd, Marie, Fernando, Laura, Ana, Marcus, Carlos, Bruno, obrigado pelas discussões inspiradoras.

Á minha namorada Caroline pela paciência nas horas de estudo, pelas releituras, pelos sonhos compartilhados e pelo apoio até hoje.

Por fim, ao meu professor e amigo Dr Paul Gerhard Kinas, a definição de orientador, pelos dois anos de convivência, deixo meus melhores agradecimentos pelo conhecimento, pelo tempo compartilhado, pelas orientações e oportunidades concedidas.

## Resumo

Apesar de sua notável inserção no meio científico ao longo das últimas cinco décadas, a inferência bayesiana ainda permanece um desafio quando trata-se da otimização ou criação de novas metodologias para obtenção da distribuição posterior. Nessa ótica, revisamos sob um enfoque matemático e bayesiano o método de Laplace, poderosa ferramenta e peça fundamental de vários pacotes estatísticos, R-INLA, Laplace's Demon, para aproximar a distribuição posterior de maneira rápida e eficiente.

**Palavras-chave: Inferência bayesiana, Distribuição posterior, Método de Laplace.**

# Lista de Figuras

# Sumário

# 1 Introdução

O conceito de determinismo cientifico, cujas raízes remontam às ideias de Galileu, Kepler e posteriormente ao sucesso das leis de Newton, permeou a mente do cientista francês Pierre-Simon Laplace [1749-1827] quando apresentou o seu artigo "*A Philosophical Essay on Probabilities* " em 1814. Nesse artigo, Laplace apresenta a ideia da possibilidade de um intelecto **prever completamente** estados futuros caso tenha **total conhecimento** do presente. Hoje, sabemos que tal intelecto, conhecido como *Demônio de Laplace*, é impossível de ser concebido segundo os princípios da irreversibilidade e da incerteza [18]. A consequência lógica dessa impossibilidade é a necessidade de teorias aplicáveis à incerteza definida como incompletude da informação: teoria da probabilidade. De posse dessa teoria e de técnicas de inferência, poderemos fazer razoáveis previsões seguindo aqui a abordagem bayesiana.

Na linha bayesiana, a inferência é definida como um conjunto de métodos baseados diretamente no teorema do Rev. Thomas Bayes [1702-1761]. Este último, cujo artigo " *An essay towards solving a problem in the doctrine of chances* " póstumo apresentado em 1763 por Richard Price [1723-1791], mostrava pela primeira vez uma tentativa de determinar probabilidades num "sentido inferencial" [2]. Ele procurava determinar a probabilidade de um evento futuro numa situação na qual não se tem nenhuma informação e com base na atualização ao surgir novas evidências. Porém, a comunidade científica precisou esperar dez anos para que Laplace apresentasse o desenvolvimento formal e matemático da teoria da probabilidade particularmente sob a ótica bayesiana motivado pela necessidade de analisar dados astronômicos sem ter conhecimento do trabalho de Bayes.

Englobando a abordagem *Fisheriana* cuja formalização atribuí-se ao matemático russo Andrei Kolmogorov [1], probabilidade bayesiana segundo Harold Jeffreys[1891-1989] é definida como métrica universal para quantificação de incertezas decorrentes de informação incompleta. Por isso, mesmo parâmetros definidos como constantes em modelos matemático-

estatísticos poderão ter distribuições de probabilidades associadas a eles caracterizando incertezas sobre o seu valor real e à medida que aprimoramos o conhecimento sobre o parâmetro, reduzimos o nosso grau de incerteza que se refletirá sobre sua distribuição de probabilidade contrariamente à lógica *frequentista*. Logo, a inclusão de informações prévias nos modelos é de suma importância, pois serve de alicerce para a reconstrução do conhecimento pela atualização das estruturas existentes entre "**aquilo que se sabia**" ou priori e "**aquilo que se sabe**" ou posteriori à luz de novas evidências sob forma de dados. A construção da posterior dá-se pelo teorema de Bayes a partir da distribuição a priori e de dados sumarizados na função de verossimilhança [12, 16]. Ou seja, toda inferência bayesiana resume-se ao bom uso do teorema de Bayes para a construção da distribuição posterior, âmago de toda a análise de decisão.

Em alguns casos, tais como famílias conjugadas, em que as distribuições a priori e posteriori pertencem à mesma classe de funções, a construção da posterior pode ser feita analiticamente. Porém, em muitos casos reais essa construção analítica é complexa ou até mesmo impossível [3]. É, portanto, indispensável recorrer a aproximações analíticas ou numéricas. Nessa última categoria, podemos citar especialmente os pacotes estatísticos frequentemente usados no software livre $\mathcal{R}$ tais como JAGS, INLA, *Laplace's Demon* que auxiliam na computação numérica.

O JAGS usa o *Markov Chain Monte Carlo*(MCMC), um algoritmo estocástico iterativo eficaz pela flexibilidade em relação à classe de distribuições mas limitado pela lentidão na sua convergência [12]. Um método alternativo ao anterior para obtenção de distribuições à posteriori foi proposto em 1986 por Tierney e Kadane [17]. Eles utilizaram o método de Laplace como aproximação assintótica para obter densidades marginais a posteriori. Rue et al. [10], em 2009, propuseram um novo método determinístico, o método *Integrated Nested Laplace Approximations* ($INLA$), que permite fazer inferência Bayesiana aproximada para os chamados modelos Gaussianos latentes de maneira rápida e precisa, baseados no mesmo método de aproximação de Laplace usado por Tierney e Kadane, além de

não ser necessário a verificação de convergência. Ainda temos o *Laplace's Demon*, um algoritmo misto que usa o MCMC e beneficia-se da aproximação de Laplace para estabelecer ótimos pontos de partida para as cadeias de Markov que aceleram a convergência.

Visamos com esse trabalho apresentar a definição de probabilidade adequada ao contexto bayesiano, exemplificar as limitações das famílias conjugadas no cálculo da distribuição posterior, descrever matematicamente o processo da aproximação de Laplace e seu uso no INLA como ferramenta para obter a posterior e apresentar medidas de discrepância para aproximações.

# 2 Probabilidade e Teorema de Bayes

Gelman et al. [7] definem inferência bayesiana como:

> "o processo de ajuste de um modelo de probabilidade a um conjunto de dados e resumindo o resultado em uma distribuição de probabilidade dos parâmetros do modelo e de quantidades não observadas, tais como previsões para novas observações."

Conjuntamente, Merriam-Webster sugere que o adjetivo "bayesiano"refere-se:

> "àquilo, relacionado ou envolvendo métodos estatísticos que associam probabilidades ou distribuições a eventos(chover amanhã) ou parâmetros (média populacional) baseados na experiência ou melhores sugestões antes do experimento e aplicando o teorema de Bayes para atualizar as probabilidades e distribuições após obter dados experimentais."

O bayesianismo tem duas vertentes: uma **objetiva** e outra **subjetiva**. Essas duas abordagens se diferenciam pelas suas respectivas metodologias e consequências da definição adotada de probabilidade [5]. Os desenvolvimentos a seguir seguem uma lógica objetiva de probabilidade baseada em [5]. Comentários sobre bayesianismo subjetivo são feitos ao longo do texto quando necessário. Nas definições de Gelman et al. e Merriam-Webster, fica claro o lugar ocupado pelas **distribuições de probabilidade** e pelo **teorema de Bayes** no âmbito bayesiano. Por isso, definiremos com precisão o que são probabilidade, teorema de Bayes assim como outros "ingredientes"indispensáveis na inferência bayesiana. .

## 2.1 Probabilidade no contexto bayesiano

Expressões do tipo: "isso é possível", "aquilo é plausível "são exemplos das muitas avaliações que fazemos no nosso cotidiano. Entretanto, a verdadeira tomada de decisão baseia-se não somente nessas avaliações, mas

também no **quão possível ou plausível** elas são, e toma raiz, muitas vezes, na nossa experiência. Definiremos portanto **probabilidade** como sendo *o grau de plausibilidade associado a uma incerteza decorrente de falta de informação por um individuo.* Em outras palavras, ela é uma medida numérica que associamos à plausibilidade de uma asserção incerta. Dessa última frase, saí o nosso primeiro **desiderato** na quantificação dessa métrica, como uma lógica estendida de introduzir probabilidade [5].

**Desiderato I (DI): graus de plausibilidade são representados por números reais.**

Além disso, precisamos garantir que a construção esteja "de acordo com nosso senso comum". Para tal enunciemos o segundo desiderato:

**Desiderato II (DII): exigência de uma correspondência qualitativa com o senso comum.**

Por isso, queremos dizer que:

- a uma plausibilidade maior, estará associada uma probabilidade(número) maior,
- um acréscimo infinitesimal na plausibilidade levará a um acréscimo infinitesimal na sua probabilidade, conhecida como propriedade de continuidade.

Todavia precisamos nortear a medida de plausibilidade dando lhe a propriedade de **consistência** introduzida pelo terceiro desiderato:

**Desiderato III (DIII)**:

- **IIIa: se a tomada de decisão puder ser feita de várias formas, todas devem levar ao mesmo resultado**,
- **IIIb: toda informação relevante deverá ser utilizada impossibilitando toda exclusão arbitrária para tomada de decisão**,

- **IIIc: estados equivalentes de informação tem mesma plausibilidade**.

Uma vez estabelecidas as regras de dedução da probabilidade, consideramos decisões como sendo *proposições lógicas* às quais podemos aplicar todas as operações e leis da álgebra de Boole(1812).

Sejam $A$, $B$, $C$ proposições, e desejemos avaliar a probabilidade conjunta de $A$ e $B$ condicionada a $C$ dada por $AB \mid C$. Ela pode ser feita por dois caminhos que, pelo desiderato IIIa, devem levar ao mesmo caminho:

- Avaliar a plausibilidade de $B \mid C$ e em seguida avaliar a de $A \mid BC$ ou
- Avaliar a plausibilidade de $A \mid C$ e em seguida avaliar a de $B \mid AC$ .

A obtenção da medida de probabilidade com base nos desideratos segundo Jeffreys, por ser extensa, foge do escopo deste trabalho, mas pode ser encontrada em [5] que a apresenta construída com base nos trabalhos de Keynes, Jeffreys, Pólya, R.T Cox, Tribus, de Finetti, Rosenkrantz. Estes deduziram dos três desideratos a probabilidade conjunta de A , B condicionada a C e uma condição necessária para medir grau de plausibilidade dadas respectivamente por:

$$p(AB \mid C) = p(A \mid C)p(B \mid AC) = p(B \mid C)p(A \mid BC) \tag{1}$$

$$p(A \mid C) + p(\overline{A} \mid C) = 1 \tag{2}$$

onde $\overline{A}$ é a negação ou complemento da asserção $A$. Destacamos que dessas duas proposições, é possível deduzir todos os teoremas da teoria da probabilidade. Essa abordagem da probabilidade foi adotada para apresentar as ideias pouco convencionais e geralmente não encontradas em livros de estatística que usam a teoria de medida e integração onde a probabilidade é definida como métrica num caso particular. O interessado em versões clássicas pode consultar [6], [4] e sobretudo [5] que faz uma comparação entre esses dois viés. Cabe salientar que essas deduções foram feitas usando

três proposições e uma extensão dessas ideias para um conjunto de "n" proposições é feita por indução via regra da cadeia.

A avaliação da probabilidade de um parâmetro continuo, $\gamma$ digamos, dada por $p(\gamma \leq b \mid C)$ pode ser reduzida a sentenças mutuamente exclusivas do tipo $p(A \mid y)$ e $p(B \mid C)$ onde $A \equiv (a \leq \gamma)$ e $B \equiv (a < \gamma \leq b)$ e C uma condição ou hipótese restritiva. Assim podemos usar a teoria anterior para construir a probabilidade de parâmetros contínuos.

As notações $p(A \mid C)$ e $p(A \mid BC)$ são denominadas probabilidades condicionais . Na inferência bayesiana, encontramos outros termos como prior e posterior. A prior, $p(\Theta)$, é o reflexo da subjetividade da probabilidade no âmbito bayesiano, pois ela é o grau de plausibilidade que o especialista ou pesquisador atribui inicialmente a um evento baseado na suas experiencia, crenças. Precisamos nos certificar, usando a lógica objetiva, que sujeitos com mesmas informações associem mesmas prioris e para isso notemos a importância dos desideratos IIIa e IIIc, delineando essa construção. A lógica bayesiana objetiva pretende retirar toda decisão subjetiva, crença pessoal e subsidiar meios racionais e lógicos para que a decisão seja tomada. Jaynes(1956) imagina o objetivismo bayesiano como um algoritmo a ser utilizado por um robô imparcial para tomar decisão seguindo algumas regras. Uma das criticas feitas a este modelo é que a priori por si é uma avaliação subjetiva visto que representa a crença pessoal pré-experimental. Questionamentos importantes nesse estágio como: *o que considerar como priori caso eu julgar que todas as possibilidades tem o mesmo nível de credibilidade? caso eu não tenha nenhuma informação anterior à pesquisa ? caso minha intuição seja muito vaga ? etc* levaram aos conhecidos métodos de elicitação de priori que por si só representam ainda uma área de muito interesse dado que uma priori inadequada pode enviesar a pesquisa e conduzir a resultados absurdos.

No primeiro caso, a construção da priori pode ser feita seguindo o método de Bayes-Laplace ou *princípio da razão insuficiente* que afirma que na ausência de razões suficientes para privilegiar um evento em detrimento de outro, deve-se optar por uma distribuição de probabilidade uniforme

para os vários parâmetros. Obviamente isso tem alguns problemas: no caso continuo $\int p(\theta)d\theta = \infty$ e a distribuição uniforme não é invariante via reparametrização. Muitas vezes, isso se resolve por truncamento para obter um intervalo limitado para $\theta$ de interesse. Ainda temos os *métodos de Jeffrey, Box-Tiao, Berger-Bernado* para elicitar prioris cujos detalhes podem ser encontrados em [15].

Outra alternativa interessante conhecida como *método de entropia máxima* desenvolvida por Jaynes e baseada na teoria de informação de Shannon emprega ideias oriundas da mecânica estatística afim de obter uma priori não informativa sujeita a algumas restrições. Entropia é definida como métrica para quantificar a desordem de um sistema, ou seja a dificuldade em prever seu estado futuro. Um sistema de entropia máxima seria portanto equivalente a um estado de ignorância. Portanto diremos que uma distribuição tem entropia máxima se ela representar um estado de ignorância. Observemos então que essa distribuição de probabilidade seria adequada para responder à questão: *que priori considerar caso eu não tenha nenhuma informação anterior à pesquisa?*. Define-se a entropia da distribuição de probabilidade $p(\theta)$ nos casos discreto e continuo respectivamente pelos funcionais:

$$\varepsilon[p(\theta)] = \sum p(\theta) ln[p(\theta)]$$

ou

$$\varepsilon[p(\theta)] = \int p(\theta) ln[p(\theta)] d\theta.$$

**Outras medidas funcionais**

$$-\int p(\theta) ln[p(\theta)] d\theta, -\int p(\theta)^2 d\theta, \int ln[p(\theta)] d\theta, \int \sqrt{p(\theta)} d\theta$$

Estaremos interessados em obter a distribuição $p(\theta)$ que maximiza as equações anteriores com certas restrições por exemplo: $\int p(\theta)d\theta = \sum p(\theta) = 1$. Esses problemas são resolvíveis pela teoria do calculo variacional com multiplicadores de Lagrange de onde seu apreço matemático interessante.

**Exemplo 2.1** *Uniforme e Exponencial* $\int g(\theta)p(\theta)d\theta = M$

***Teorema 2.1*** *(Calculo variacional, Multiplicadores de Lagrange)*

*Seja* $J[y] = \int F(x, y, y^{'})dx$ *um funcional sujeito a restrição* $I[y] = \int G(x, y, y^{'})dx = L$. *Uma condição necessária para que* $y*$ *minimize* $J[y]$ *é:* $\exists \ \lambda$ *tal que* $\Psi(x, y*, y*^{'}) = F(x, y*, y*^{'}) - \lambda[G(x, y*, y*^{'}) - L]$ *satisfaça:*

$$\frac{\partial \Psi}{\partial y} - \frac{d}{dx}\left(\frac{\partial \Psi}{\partial y^{'}}\right) = 0. \tag{3}$$

$$\varepsilon(p) = -\int_a^b p(\theta) ln[p(\theta)] d\theta$$

*s.a*

$$\int p(\theta) d\theta = 1$$

*e*

$$\int \theta p(\theta) d\theta = \mu$$

*então:*

$$\Psi(.) = -pln[p] - \lambda_1[\theta p - 1] - \lambda_2[p - 1]$$

$$\frac{\partial \Psi}{\partial p} = -ln[p] - 1 - \lambda_1\theta - \lambda_2 = 0$$

$$p(\theta) = \frac{1}{e^{\lambda_2 + 1}} e^{-\lambda_1 \theta}$$

- *Resolvendo a integral das restrições para a e b finitos obtemos a distribuição* ***uniforme****:*

$$p(\theta) = k$$

- *Quando* $a = 0$ *e* $b \to \infty$, *obtemos* $e^{\lambda_2 + 1} = \mu$ *e* $\lambda_1 = \frac{1}{\mu}$ *ou seja a distribuição* ***exponencial****:*

$$p(\theta) = \frac{1}{\mu} e^{-\mu\theta}$$

Mais detalhes e discussões sobre a dedução de priori podem ser encontrados em [15] e [5].

A posterior, $p(\theta \mid y)$ é uma atualização da prior depois da análise de novas evidências sob forma de dados $y$. Ela é o fruto do relacionamento, um compromisso entre a priori e a verossimilhança cuja ilustração é feita na figura 1. Os dados quanto a eles são resumidos na função de verossimilhança $L(\theta \mid y)$ que dá uma ideia de quão coerente é $\theta$ baseado nos dados.

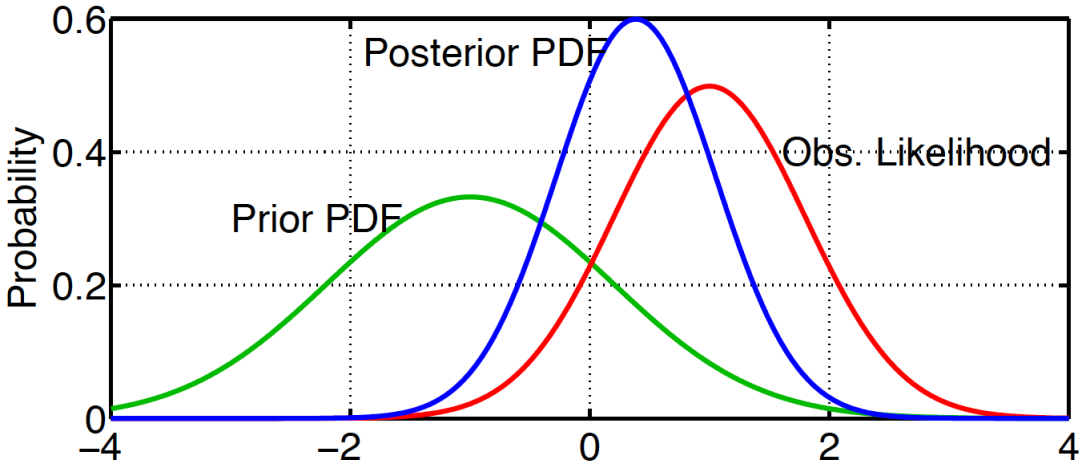


Figura 1: Posterior $\propto$ F(Prior,Verossimilhança)

## 2.2 Teorema de Bayes

O teorema de Bayes é a chave principal da inferência bayesiana. É a expressão da atualização do conhecimento com a experiência. Ele parte do pressuposto que para obter a posterior $p(\theta \mid y)$, iniciamos com a probabilidade conjunta $p(\theta, y)$ e como escrito em (1) temos :

$$p(\theta, y) = p(\theta) * p(y \mid \theta) = p(y) * p(\theta \mid y) \tag{4}$$

Da equação (4) deduzimos o teorema de Bayes que nos dá a posterior:

$$p(\theta \mid y) = \frac{p(\theta) * p(y \mid \theta)}{p(y)} \tag{5}$$

onde: $p(y) = \sum p(\theta) * p(y \mid \theta)$ para um conjunto discreto de parâmetros

ou $p(y) = \int p(\theta) * p(y \mid \theta) d\theta$ no caso continuo. Caso o parâmetro $\theta$ seja multidimensional, a marginalização é o processo pelo qual retiramos uma especifica dimensão de interesse. Nesse caso temos:

$$p(\theta_i \mid y) = \int p(\theta \mid y) d\theta_{-i} \tag{6}$$

com $\theta_{-i} = \theta \setminus \theta_i$. A $p(y)$ distribuição marginal de y para todos os valores de $\theta$, sendo uma constante normalizadora, muitas vezes é preferível usar a forma proporcional do teorema de Bayes para expressar a posterior. O seu uso é importante quando pretendemos comparar diferentes modelos. Portanto a equação (5) tem a forma:

$$p(\theta \mid y) \propto p(\theta) * P(y \mid \theta).$$

**Definição 2.1** *Duas variáveis aleatórias* $\boldsymbol{x}$, $\boldsymbol{y}$ *são ditas condicionalmente independentes dado* $\theta$ *se :* $p(x \mid \theta, y) = p(x \mid \theta)$ *. Tem-se nessas condições que :* $p(x, y \mid \theta) = p(x \mid y, \theta) * p(y \mid \theta) = p(x, y) * p(y \mid \theta)$.

O pesquisador poderia estar interessado em um segundo experimento ou observação de novos dados $x_i$ para aumentar o seu grau de plausibilidade e poderia conduzir seu experimento de tal forma a ter $x_i$ e $y$ condicionalmente independentes dado $\theta$. Nesse caso, a posterior $p(\theta \mid y)$ será a nova priori e aplicando novamente a forma proporcional do teorema de bayes obtemos:

$$\begin{aligned}
p(\theta \mid y) &\propto p(\theta) p(y \mid \theta) \\
p(\theta \mid x_1, y) &\propto p(x_1 \mid \theta) p(\theta \mid y) \\
&\propto p(\theta) p(y \mid \theta) p(x_1 \mid \theta) \\
&\vdots \propto \vdots \\
p(\theta \mid x_n, x_{n-1}, \ldots, x_1, y) &\propto \left[\prod p_i(\theta, x_i)\right] p(\theta \mid y) p(\theta)
\end{aligned} \tag{7}$$

## 2.3 Distribuição posterior

Dado que toda a análise bayesiana de decisão, particularmente a inferência bayesiana reside nos diversos usos da distribuição posterior, focaremos aqui nas mais conhecidas técnicas para sua obtenção através da literatura.

### 2.3.1 Famílias conjugadas

É sempre possível fazer uma dedução analítica da distribuição posterior para alguma classe ou seja obter uma solução exata [12]. Este grupo é conhecido como **famílias conjugadas** e define-se como segue:

**Definição 2.2** *seja $\mathcal{F}$ uma família de distribuições para a verossimilhança ou distribuições amostrais $p(y|\theta)$ e $\mathcal{P}$ uma família de distribuições a priori $p(\theta)$. Dizemos que $\mathcal{F}$ e $\mathcal{P}$ são famílias conjugadas de distribuições quando a posterior $p(\theta|y)$ for também da família $\mathcal{P}$. Matematicamente, $\mathcal{P}$ e $\mathcal{F}$ são conjugadas se*

$$\forall\ p(y|\theta) \in \mathcal{F}\ e\ p(\theta) \in \mathcal{P} \Rightarrow p(\theta \mid y) \in \mathcal{P}.$$

Segue uma tabela de famílias conjugadas muito presentes na literatura [12]

| $\mathcal{P}$ | $\mathcal{F}$ |
|---|---|
| Beta | Binomial |
| Gama | Poisson |
| Normal-Gama | Normal |
| Dirichlet | Multinomial |

Tabela 1: Familias conjugadas

A procura dessas famílias é majoritariamente pela facilidade e exatidão da distribuição posterior. Mostraremos exemplos de famílias conjugadas onde a distribuição posterior pode ser obtida de forma analítica exata:

**Exemplo 2.2** ***P-beta, F-binomial***

*Uma variável aleatória (v.a) Y, lembrando que aleatoriedade é sinonimo de incompletude de informação, tem distribuição de probabilidade*

*binomial Bin* $(n, \theta)$ *para inteiros de 0 a n se sua função de massa de probabilidade(f.m.p) é definida por:*

$$p(y \mid \theta) = \begin{pmatrix} n \\ y \end{pmatrix} .\theta^y.(1-\theta)^{n-y} \propto \theta^y.(1-\theta)^{n-y} \tag{8}$$

*Uma v.a* $\theta \in [0, 1]$ *tem distribuição beta com parâmetros a e b positivos, t.q n=a+ b se sua função densidade de probabilidade (f.d.p) é definida por:*

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a).\Gamma(b)}\theta^{a-1}.(1-\theta)^{b-1} \propto \theta^{a-1}.(1-\theta)^{b-1} \tag{9}$$

*onde*

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \tag{10}$$

*Combinando as equações (8) e (9) pelo teorema de bayes, obtemos a posterior dada por :*

$$\begin{aligned} p(\theta \mid y) &\propto p(\theta) * p(y \mid \theta) && (11) \\ p(\theta \mid y) &\propto \theta^y.(1-\theta)^{n-y}.\theta^{a-1}.(1-\theta)^{b-1} \\ &\propto \theta^{y+a-1}.(1-\theta)^{n+b-y-1} && (12) \end{aligned}$$

*Portanto a posterior tem distribuição beta com parâmetros* $a^* = a + y$ *e* $b^* = b + n - y$*, donde as famílias conjugadas* ***beta*** *e* ***binomial****.*

**Exemplo 2.3** ***P-Gamma, F-Poisson***

*A distribuição Gama*$(\alpha, \beta)$ *é definida para todo valor* $\theta > 0$ *e tem sua f.d.p dada por :*

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}.\theta^{\alpha-1}.e^{-\beta\theta} \propto \theta^{\alpha-1}.e^{-\beta\theta} \tag{13}$$

*Uma v.a. y tem distribuição Poisson* $p(y \mid \theta)$ *com parâmetro* $\theta$ *se sua f.m.p é:*

$$p(y \mid \theta) = \frac{e^{n\theta}(n\theta)^y}{y!} \propto e^{-n\theta}.\theta^y, y \in \mathbb{N} \tag{14}$$

*Combinando as equações (13) e (14) que representam respectivamente as distribuições a priori e amostral, pela forma proporcional do teorema de Bayes, obtemos a posterior:*

$$\begin{aligned} p(\theta \mid y) &\propto \theta^{\alpha-1}.e^{-\beta\theta}.e^{-n\theta}.\theta^y \\ &\propto \theta^{y+\alpha-1}.e^{-(n+\beta)\theta} \end{aligned} \tag{15}$$

*O núcleo dessa posterior sendo o de uma Gama($\alpha^*, \beta^*$), com $\alpha^* = y + \alpha, \beta^* = n + \beta$ deduzimos que as famílias* ***Gamma*** *e* ***Poisson*** *são conjugadas.*

É óbvio que nos casos em que a priori e a verossimilhança pertencem à mesma família, a posterior também pertencerá à mesma dado que o núcleo da distribuição será mantida. Falaremos de um caso geral que engloba distribuições usuais em estatística: família exponencial.

**Exemplo 2.4** ***Família exponencial***

*Uma distribuição de probabilidade $p(y \mid \theta)$ pertence à família exponencial se ela puder ser escrita na forma:*

$$p(y \mid \theta) = f(\theta)g(y)exp\left[\phi(\theta)s(y)\right].$$

*Pode-se mostrar facilmente que as distribuições* ***binomial, beta, normal, gamma, poisson, exponencial, binomial-negativa, weibull, Dirichlet, Wishart, normal-gamma etc*** *pertencem à família exponencial.*

***Teorema 2.2*** *Seja uma amostra $Y = y_1, \ldots, y_n$ seguindo uma distribuição $p(y \mid \theta)$ e sua função de verossimilhança dada por:*

$$L(y \mid \theta) = \prod_{1=1}^{n} p(y \mid \theta) = f(\theta)^n \prod_{1=1}^{n} g(y_i)exp\left[\phi(\theta)\sum s(y_i)\right]$$

*. Então a prior dada por*

$$p(\theta) = f(\theta)^a exp\,[\phi(\theta)b]$$

*é conjugada a* $L(y \mid \theta)$.

*Com efeito, basta ver que:*

$$p(\theta \mid y) \propto f(\theta)^{n+a} \prod_{1=1}^{n} g(y_i) exp\left[\phi(\theta)(\sum sy_i + b)\right]$$

*que também é da família exponencial. Esse teorema nos evidencia um jeito simples de encontrar uma priori conjugada no caso da família exponencial e é de importante uso visto a grande gama de distribuições que envolve.*

Todo cuidado deve ser tomado para evitar o uso inapropriado de uma prior conjugada, cuja utilidade se justifica apenas pela facilidade em obter a posterior.

Na ausência de famílias conjugadas, recorre-se a aproximações analíticas ou numéricas para obter a distribuição posterior pois seu cálculo depende, muitas vezes, de integrais complexas [3]. As aproximações numéricas vão de discretizações das integrais a simulações estocásticas. Na próxima secção, introduziremos o método de Laplace, uma aproximação analítica com uso difundido na literatura e falaremos das abordagens clássicas que existem na literatura.

As funções que usaremos a seguir serão consideradas suficientemente continuas e diferenciáveis, de classe $C^\infty$ sem perda de generalidade, mas na verdade de classe $C^2$ ou $C^3$ na prática. Essa hipótese não é tão restritiva pois na maioria das aplicações estatísticas, as distribuições de probabilidade possuem essa característica. todos os códigos utilizados podem ser encontrados no anexo 1.

# 3 Método de Laplace

Interessado em resolver integrais do tipo $I(\lambda) = \int_a^b q(t)e^{\lambda p(t)}dt$, $\lambda > 0$ Laplace(1774) desenvolveu um método cuja demonstração formal feita por ele foi submetida em [14] (1814).

Note que:

$$I(\lambda) = \int_a^b p(t)e^{\lambda q(t)}dt = \frac{1}{\lambda}\int_a^b \frac{p(t)}{q^{'}(t)}(e^{\lambda q(t)})^{'}dt$$

Integrando por partes temos:

$$I(\lambda) = \left[\frac{1}{\lambda}\frac{p(t)}{q^{'}(t)}e^{\lambda q(t)}\right]_a^b - \frac{1}{\lambda}\int_a^b \left(\frac{p(t)}{q^{'}(t)}\right)^{'} e^{\lambda q(t)}dt$$

**Proposição 3.1** *Se p(t) , q(t) , $q^{'}(t)$ forem continuas, $q^{'}(t) \neq 0 \; \forall \; t \; \in [a,b]$ e $p(a) * p(b) \neq 0$ então: $I(\lambda) \sim \left[\frac{1}{\lambda}\frac{p(t)}{q^{'}(t)}e^{\lambda q(t)}\right]_a^b$, $\lambda \to \infty$*

Ou seja podemos obter uma aproximação assintótica :

$$I(\lambda) \simeq \frac{1}{\lambda}\left[\frac{p(b)}{q^{'}(b)}e^{\lambda q(b)} - \frac{p(a)}{q^{'}(a)}e^{\lambda q(a)}\right]$$

.

Todavia, pode existir um único $c \in [a,b]$ tq $q^{'}(t) = 0$ e $q^{''}(t) \leqslant 0$ a integração por partes falha. Nesse caso, a avaliação do valor de $I(\lambda)$ numa vizinhança, $V_\epsilon(c)$ de raio $\epsilon$ centrada em c, quando $\lambda \to \infty$ representa uma boa aproximação assintótica.

Observamos que $q^{'}(c) = 0$ e $q^{''}(c) \leqslant 0$ implica que c é um máximo global de q(t) em [a,b] ou seja $\forall \; t \; \in [a,b], q(t) \leqslant q(c)$. Tomemos c=0 sem perda de generalidade.

**Teorema 3.1** *Sejam p(t), q(t) funções de classe $C^2$, com $p(t) > 0$ de ordem exponencial tal que $lim_\infty p(t) = 0$ c o máximo de q(t). Então:*

$$I(\lambda) = \int_{-\infty}^{+\infty} p(t)e^{\lambda q(t)}dt \simeq \int_{c-\epsilon}^{c+\epsilon} p(t)e^{\lambda q(t)}dt \quad \textit{quando} \quad \lambda \to \infty.$$

$$I(\lambda) \simeq \int_{-\epsilon}^{+\epsilon} p(t)e^{\lambda q(t)}dt \qquad c = 0 \ \in (a,b) \tag{16}$$

$$I(\lambda) \simeq \int_{0}^{+\epsilon} p(t)e^{\lambda q(t)}dt \qquad c = 0 = a$$

$$I(\lambda) \simeq \int_{-\epsilon}^{0} p(t)e^{\lambda q(t)}dt \qquad c = 0 = b$$

Usando 16 e expandindo p(t) em serie de taylor em torno de 0 até a segunda ordem, e tomando $p(t) \simeq p(0) \neq 0$ temos :

$$\begin{aligned}
I(\lambda) &\simeq \int_{-\epsilon}^{+\epsilon} p(t)e^{\lambda q(t)}dt \\
&\simeq \int_{-\epsilon}^{+\epsilon} p(0)e^{\lambda\left[q(0)+\frac{1}{2}q^{''}(0)(t)^2+O(t^3)\right]}dt \\
&\simeq p(0)e^{\lambda q(0)}\int_{-\epsilon}^{+\epsilon} e^{\lambda\left[\frac{1}{2}q^{''}(c)(t)^2\right]}dt \\
&\simeq p(0)e^{\lambda q(0)}\int_{-\infty}^{+\infty} e^{\lambda\left[\frac{1}{2}q^{''}(c)(t)^2\right]}dt \\
&= p(0)e^{\lambda q(0)}\sqrt{\frac{2\pi}{-\lambda q^{''}(0)}} \\
I(\lambda) &\sim p(0)e^{\lambda q(0)}\sqrt{\frac{2\pi}{-\lambda q^{''}(0)}} \qquad \lambda \to +\infty
\end{aligned} \tag{17}$$

Onde no penúltimo passo foi usado a igualdade $\int_{-\infty}^{+\infty} e^{-s^2}ds = \sqrt{\pi}$ .

Expansões de ordem "n" são importantes caso temos por exemplo $q^{'}(0) = q^{''}(0) = \ldots = q^{n-1}(0) = 0$.

Uma observação importante é no caso da existência de múltiplos máximos para a função q(t). Reduzimos $I(\lambda)$ ao calculo de $I(\lambda)_{\epsilon_i}$ onde $\epsilon_i$ representa o raio da vizinhança em torno de cada máximo $c_i$. Portanto

$$I(\lambda) = \sum_{i=1}^{n} p(c_i)e^{\lambda q(c_i)}\sqrt{\frac{2\pi}{-\lambda q^{''}(c_i)}}$$

No caso multidimensional, sejam $q(T) : \mathbb{R}^n \to \mathbb{R}$, e $p(T) : \mathbb{R}^n \to \mathbb{R}$ e desejemos obter uma aproximação assintótica para:

$$I(\lambda) = \int_{\Omega} q(T)e^{\lambda p(T)}dT,$$

com $\Omega$ o domínio de $p(t) * q(t)$ uma região simplesmente conexa com bordo de classe $C^1$, e $T \in \mathbb{R}^n$.

Usaremos a mesma ideia da dedução anterior admitindo inicialmente que p(T) possui um único máximo em $C = 0 \in \Omega$. Expandimos p(T) em serie de Taylor até a segunda ordem. Para mais precisão e necessidade ( $q^{'}() = q^{''}(0) = \ldots = q^{n-1}(0) = 0$ ), ordens superiores podem ser adotadas. Temos:

$$\begin{aligned} p(T) &= p(0) + \langle T, \bigtriangledown p(0)\rangle + \langle \frac{1}{2}T, H(0)\rangle + O(|T|^3) \\ p(T) &= p(0) + \frac{1}{2}T^{\perp}H(0)T + O(|T|^3) \end{aligned} \tag{18}$$

onde:

$$\bigtriangledown p = \left(\frac{\partial P}{\partial t_1}, \ldots, \frac{\partial P}{\partial t_n}\right)$$

é o gradiente de p e

$$H_{ij} = P_{t_i t_j} = \frac{\partial^2 P}{\partial t_i \partial t_j} = \begin{vmatrix} P_{t_1 t_1} & \ldots & P_{t_1 t_n} \\ \vdots & \vdots & \vdots \\ P_{t_n t_1} & \ldots & P_{t_n t_n} \end{vmatrix}$$

é a matriz Hessiana de P.

**Proposição 3.2** *Dada a hessiana H*

*1 H é simétrica pelo teorema de Clairaut-Schwarz: $H_{ij} = H_{ji}$*

*2 H possui autovalores reais e autovetores ortogonais( que podemos normalizar para obter uma base ortonormal) pelo teorema espectral.*

*3 assumindo H negativa-definida, todos os seus autovalores são negativos.*

De [2], podemos escrever $H = ADA^{\perp}$ onde: A é uma matriz ortonormal formada pelos autovetores de H e D é uma matriz diagonal formada pelos autovalores de H.

$$A = [v_1, \ldots, v_n] \quad |A| = 1$$

$$D_{ij} = -{a_i}^2 \;\; se \;\; i = j \;\; e \;\; D_{ij} = 0 \;\; cc \; .i, j : 1 \ldots n.$$

Fazendo a mudança de variável: $T = AX$

$$\begin{aligned} p(T) &= p(0) + \frac{1}{2}T^{\perp}HT + O(|T|^3) \\ p(AX) &= p(0) + \frac{1}{2}(AX)^{\perp}ADA^{\perp}AX + O(|AX|^3) \\ p(AX) &= p(0) + \frac{1}{2}X^{\perp}A^{\perp}ADA^{\perp}AX + O(|X^3|) \\ p(AX) &= p(0) + \frac{1}{2}X^{\perp}DX + O(|X^3|) \end{aligned} \tag{19}$$

Ora D é diagonal logo:

$$X^{\perp}DX = -\sum_{i=0}^{n} a_i^2 x_i^2$$

Essa equação em (19) nos dá:

$$p(AX) = p(0) - \sum_{i=0}^{n} a_i^2 x_i^2 + O(|X^3|)$$

e

$$\begin{aligned} I(\lambda) &\approx \int_{A^{\perp}X} q(AX) e^{\lambda\left[p(0)-\sum a_i^2 x_i^2 + O(|X^3|)\right]} dX \\ &\approx e^{\lambda p(0)} \int_{A^{\perp}X} q(AX) e^{\lambda\left[-\sum a_i^2 x_i^2\right]} dX \\ &\approx e^{\lambda p(0)} \int_{A^{\perp}X} q(AX) \prod e^{-\lambda a_i^2 x_i^2} dX \end{aligned} \tag{20}$$

Supondo que $A^{\perp}X = X_1 \times X_2 \times \ldots \times X_n$ e $q(AX) \approx q(0)$ temos:

$$\begin{aligned} I(\lambda) &\approx e^{\lambda p(0)} \int_{A^{\perp}X} q(AX) \prod e^{-\lambda a_i^2 x_i^2} dX \\ &\approx e^{\lambda p(0)} \prod \int_{-\epsilon}^{\epsilon} q(0) e^{-\lambda a_i^2 x_i^2} dx \end{aligned} \tag{21}$$

Utilizando o resultado da versão unidimensional, chegamos a:

$$I(\lambda) \approx e^{\lambda p(0)} q(0) \prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda a_i^2}} \qquad \lambda \to \infty \tag{22}$$

Em suma, o método baseia-se nos teoremas de expansão em série de Taylor e de aproximação assintótica e consiste em expandir, sob determinadas condições para garantir convergência da integral, a função $p(t)$ em série de potências em torno do ponto máximo da mesma afim de transformar a integral inicial numa mais fácil. A aproximação $I(\lambda)$ será chamada aproximação de Laplace. De posse dessa ferramenta, podemos obter aproximações para a distribuição posterior.

# 4 Aproximações para a distribuição da Posterior

Gelman et. al em [7] enumerou os seguintes métodos como técnicas mais usadas para obter a posterior:

- Approximate Bayesian Computation (ABC)
- Iterative Quadrature
- Markov Chain Monte Carlo (MCMC)
- Importance sampling
- Laplace approximation
- Gaussian approximation
- Variational Bayes (VB)

Grande parte dessas técnicas encontra-se disponível em pacotes estatísticos dentro do software livre *R* tais como: **R-JAGS, R-INLA, LAPLACE'S DEMON**.

O pacote $\mathcal{R}$-JAGS usa o *Markov Chain Monte Carlo* (MCMC), o LAPLACE'S DEMON usa o método de Laplace conjuntamente com o MCMC para acelerar a convergência se tratando de um algoritmo iterativo [9].

O nosso foco está nas aproximações analíticas especialmente na **aproximação de Laplace** por este método envolver uma formulação matemática interessante e por ele ser o coração do $\mathcal{R}-\mathcal{INLA}$. Antes de apresentar o método de Laplace no cunho bayesiano, falaremos da aproximação gaussiana um procedimento semelhante ao método de Laplace.

# 5 Aproximação Gaussiana

Walker(1969) provou que para um n amostral grande e sob algumas condições, a distribuição posterior multiparamétrica é aproximadamente normal. Em analise bayesiana, esse teorema é o primeiro recurso analítico quando esbarramos com dificuldades para obter a posterior.

Lembremos que:

$p(\theta \mid y) \propto p(\theta)p(y \mid \theta) = exp(ln[p(\theta)] + ln[p(y \mid \theta)])$

Expandindo $ln[p(y \mid \theta)]$ em serie de Taylor em torno do seu estimador de máxima verossimilhança $\theta_n$ temos:

$$ln[p(y \mid \theta)] = ln[p(y \mid \theta_n)] - \frac{1}{2}[\theta - \theta_n]^t H(\theta_n)[\theta - \theta_n] + R_n$$

onde $H(\theta_n)$ é a hessiana de $ln[p(y \mid \theta)]$ em $\theta_n$.

Expandindo também $ln[p(\theta)]$ em torno da sua moda $\theta_o$ tem-se:

$$ln[p(\theta)] = ln[p(\theta_o)] - \frac{1}{2}(\theta - \theta_o)^t H(\theta_o)(\theta - \theta_o) + R_o$$

onde $H(\theta_n)$ é a hessiana de $ln[p(\theta)]$ em $\theta_o$.

Combinando essas duas expressões obtemos:

$$\begin{aligned}
p(\theta \mid y) &\propto exp\left[ln[p(\theta)] + ln[p(y \mid \theta)]\right] \\
&\propto exp[ln[p(y \mid \theta_n)] - \frac{1}{2}(\theta - \theta_n)^t H(\theta_n)(\theta - \theta_n) + ln[p(\theta_o)] - \frac{1}{2}(\theta - \theta_o)^t H(\theta_o \\
&\propto p(\theta_o)p(y \mid \theta_n)exp(-\frac{1}{2}(\theta - \theta_n)^t H(\theta_n)(\theta - \theta_n) - \frac{1}{2}(\theta - \theta_o)^t H(\theta_o)(\theta - \theta_o) + \\
p(\theta \mid y) &\propto p(\theta_o)p(y \mid \theta_n)exp[-\frac{1}{2}(\theta - \mu)^t H(\theta - \mu)]
\end{aligned}$$

com: $H = H(\theta_n) + H(\theta_o)$ e $\quad H\mu = H(\theta_n)\theta_n + H(\theta_o)\theta_o$.

Ou seja a posterior tem uma distribuição multivariada aproximadamente normal com média $\mu$ e matriz de covariância $H^{-1}$. Essa aproximação não é muito boa quando as distribuições envolvidas possuem alguma assimetria. Por isso recorremos à aproximação de Laplace.

# 6 Aproximação de Laplace

Como foi apresentado, a aproximação de Laplace pretende reduzir uma dada integral a uma do tipo gaussiana isto é se a expressão $q(t)e^{\lambda p(t)}$ representasse uma distribuição de probabilidade, estaríamos aproximando-a por uma distribuição normal com parâmetros a determinar. Na estatistifica essa integral pode estar representando a função geradora de momentos de p(t) onde q(t)é a densidade de probabilidade de t ou pode ser a esperança de $e^{\lambda p(t)}$ caso q(t) seja a posterior de onde a importância dessa integral.

A ideia de Laplace era a de que o valor da integral I dependesse unicamente da vizinhança do máximo da função p(t) quando $\lambda$ tendesse para o infinito. A figura abaixo mostra essa dependência para alguns valores de lambda.

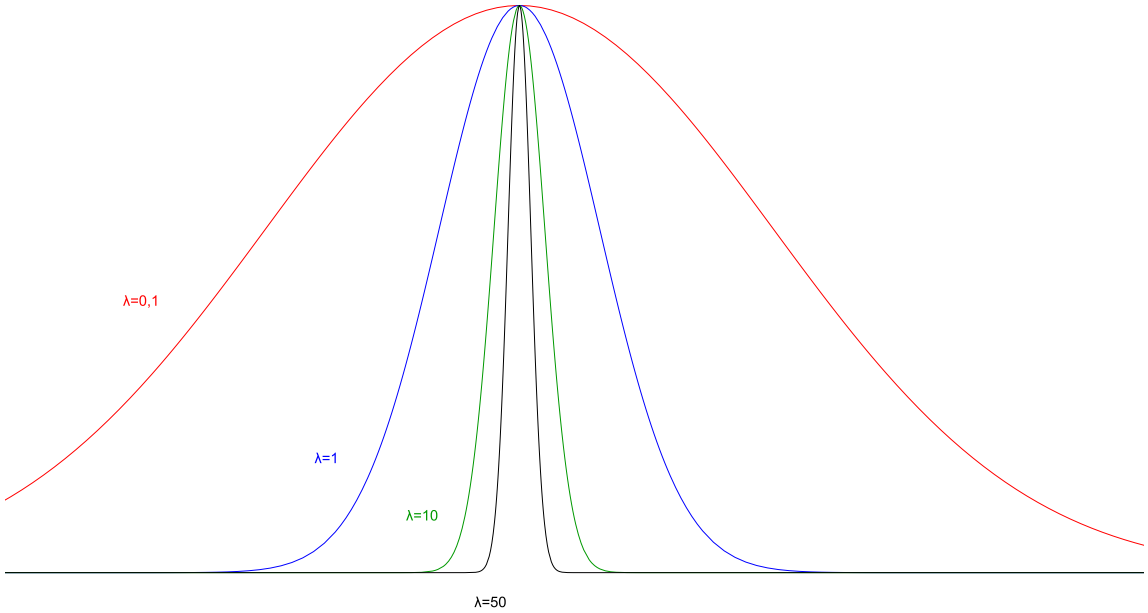


Figura 2: Área = F($\lambda$)

Essa ideia permaneceu "morta"até 1986 quando Tierney e Kadane em [17] fizeram uso dela para obter uma boa aproximação para densidades marginais com um erro de ordem $O(n^{-2})$. Observa-se que este é um método analítico relativamente simples pois exige apenas capacidade de determinar pontos extremos, notadamente moda da distribuição posterior, com alguns pressupostos, e capaz de produzir uma aproximação razoavelmente boa. Relembremos a distribuição posterior obtida pelo teorema de Bayes: $p(\theta \mid y) \propto p(\theta) * p(y \mid \theta)$. Portanto, se desejarmos fazer previsões para alguma função $g(\theta)$, inicialmente tomada positiva, basta computar a distribuição preditiva dada por:

$$\begin{aligned} E[g(\theta)] &= \int g(\theta)p(\theta \mid y)d\theta \\ &= \frac{\int g(\theta)p(\theta)p(y \mid \theta)d\theta}{\int p(\theta)p(y \mid \theta)d\theta} \end{aligned} \tag{23}$$

Tierney e Kadane e aplicaram o método de Laplace tanto ao numerador quanto ao denominador de $E[g(\theta)]$ provando que os respectivos erros de ordem $n^{-1}$ se cancelam dando uma aproximação com erro de ordem $n^{-2}$.

Dado que $p(\theta) > 0$ , $p(y \mid \theta)$, $g(\theta)$ são positivos, tomemos:

$$\begin{aligned} \lambda f(\theta) &= ln[p(\theta)] + ln[p(y \mid \theta)] \quad \text{e} \\ \lambda f^*(\theta) &= ln[g(\theta)] + ln[p(\theta)] + ln[p(y \mid \theta)]. \\ \text{Logo} \qquad E[g(\theta) \mid y] &= \frac{\int e^{[\lambda f^*(\theta)]} d\theta}{\int e^{[\lambda f(\theta)]} d\theta} \end{aligned}$$

Sejam $\theta_o$ e $\theta^*$ os respectivos máximos de $f$ e $f^*$ ou as modas das respectivas distribuições . Usando a eq 3 temos

$$\begin{aligned} E[g(\theta) \mid y] &\approx \frac{e^{\lambda f^*(\theta^*)} \sqrt{\frac{2\pi}{-\lambda f^{*''}(\theta^*)}}}{e^{\lambda f(\theta_o)} \sqrt{\frac{2\pi}{-\lambda f''(\theta_o)}}} \\ E[g(\theta) \mid y] &= (\sigma^*/\sigma_o) e^{\lambda[f(\theta_o) - f^*(\theta^*)]} \end{aligned} \tag{24}$$

onde, $\sigma^* = \sqrt{\frac{2\pi}{-\lambda f^{*''}(\theta^*)}}$ e $\sigma_o = \sqrt{\frac{2\pi}{-\lambda f''(\theta_o)}}$ .

A extensão para o caso mutliparamétrico se faz usando a equação 3 e

obtemos

$$\begin{aligned} E[g(\theta) \mid y] &\approx \frac{e^{\lambda f^*(\theta^*)} \prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda a^{*2}_i}}}{e^{\lambda f(\theta_o)} \prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda a_i^2}}} \\ E[g(\theta) \mid y] &\approx e^{\lambda[f^*(\theta^*) - f(\theta_o)]} \frac{\prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda a^{*2}_i}}}{\prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda a_i^2}}} \\ E[g(\theta) \mid y] &\approx e^{\lambda[f^*(\theta^*) - f(\theta_o)]} (\sigma^*/\sigma_o) \end{aligned} \tag{25}$$

onde, $\sigma_o = \prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda {a_i}^2}}$ e $\sigma^* = \prod_{i=1}^{n} \sqrt{\frac{2\pi}{\lambda a^{*2}_i}}$ .

No caso geral em que $g(\theta)$ assume alguns valores não positivos Tierney et al (1989) sugerem que determinamos primeiro uma aproximação para a função geradora de momentos, $E[exp(sg(\theta))]$. Dado que ela é positiva o método pode ser usado e visto também que:

$$E[g(\theta)] = \frac{d}{ds} log \left[ \int p(\theta \mid y) e^{sg(\theta)} d\theta \right]_{s=0} = \frac{d}{ds} log \left[ E[e^{sg(\theta)}] \right]_{s=0}$$

, basta então obter uma aproximação para a função geradora de momentos e em seguida obter $E[g(\theta)]$ derivando o logaritmo da geradora aproximada calcula em s=0. Mais detalhes sobre isso podem ser vistos em [17].

A seguir, usaremos a aproximação de Laplace para fim de comparação com distribuições oriundas de famílias conjugadas e mostraremos a sua importância no INLA.

**Exemplo 6.1** *: aproximação de Laplace para distribuição beta Seja* $x \in (0,1)$ *uma v.a. seguindo uma distribuição* $\beta(a,b)$ $\lambda = N = a + b$*, onde* $N$ *é o tamanho amostral. Sua f.m.p é dada por:*

$$p(x) = \frac{\Gamma(a+b)}{\Gamma(a).\Gamma(b)} x^{a-1}.(1-x)^{b-1} \propto x^{a-1}.(1-x)^{b-1}.$$

*O caso a=b=1 leva a uma distribuição uniforme que não necessita de aproximações analíticas. Pelo método de Laplace, a aproximação melhora quanto maior for $\lambda$ ou seja quanto mais dados tivermos. Comecemos tomando o logaritmo de p(x) e derivemos para obter a moda $x^*$.*

$$log(p(x))' = \frac{a-1}{x} - \frac{b-1}{1-x}$$

$e\ x^* = \dfrac{a-1}{a+b-2} \quad a > 1, \ \ b > 1.$

$$\sigma^2 = \frac{-1}{log(p(x^*))''} = \frac{(a-1)(b-1)}{(a+b-2)^3}$$

*A aproximação de Laplace para beta é a normal: $N(x^*, \sigma^2)$. Os gráficos abaixo mostram a aproximação para alguns valores de (a,b).*

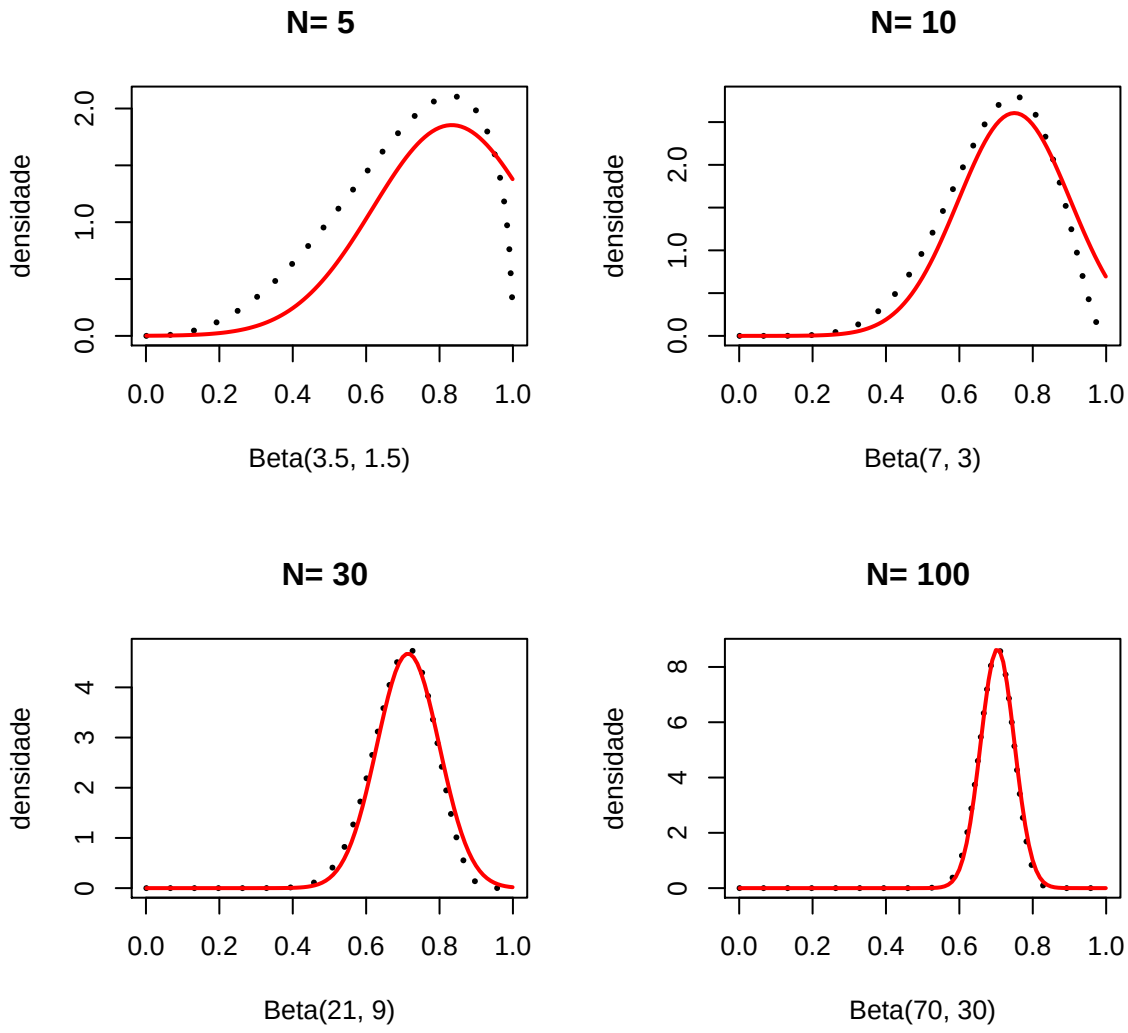


Figura 3: Convergência da aproximação de Laplace( curva cheia) para a distribuição beta (curva pontilhada) com o aumento do tamanho amostral N

**Exemplo 6.2** *: aproximação de Laplace para distribuição Qui-quadrado Sejam $X_i$ v.a normalmente distribuídas. A v.a $Y = \sum_i^N X_i^2$ tem distribuição $\chi^2$ com média N(grau de liberdade) e variância 2N e sua f.m.p é dada por:*

$$p(x) = \frac{1}{2^{N/2}\Gamma(N/2)} x^{N/2-1} exp^{-x/2}$$

*Da mesma forma tomamos o logaritmo de p(x) que derivamos pra obter a moda $x^*$.*

$$log(p(x))' = \left[\frac{N}{2} - 1\right]\frac{1}{x} - \frac{1}{2}$$

*e*

$$x^* = N - 2$$

$$\sigma^2 = \frac{-1}{log(p(x^*))''} = 2(N-2)$$

*A aproximação de Laplace para a Qui-quadrado é a normal:* $N(x^*, \sigma)$. *Os gráficos que seguem mostram a convergência para alguns valores de n.*

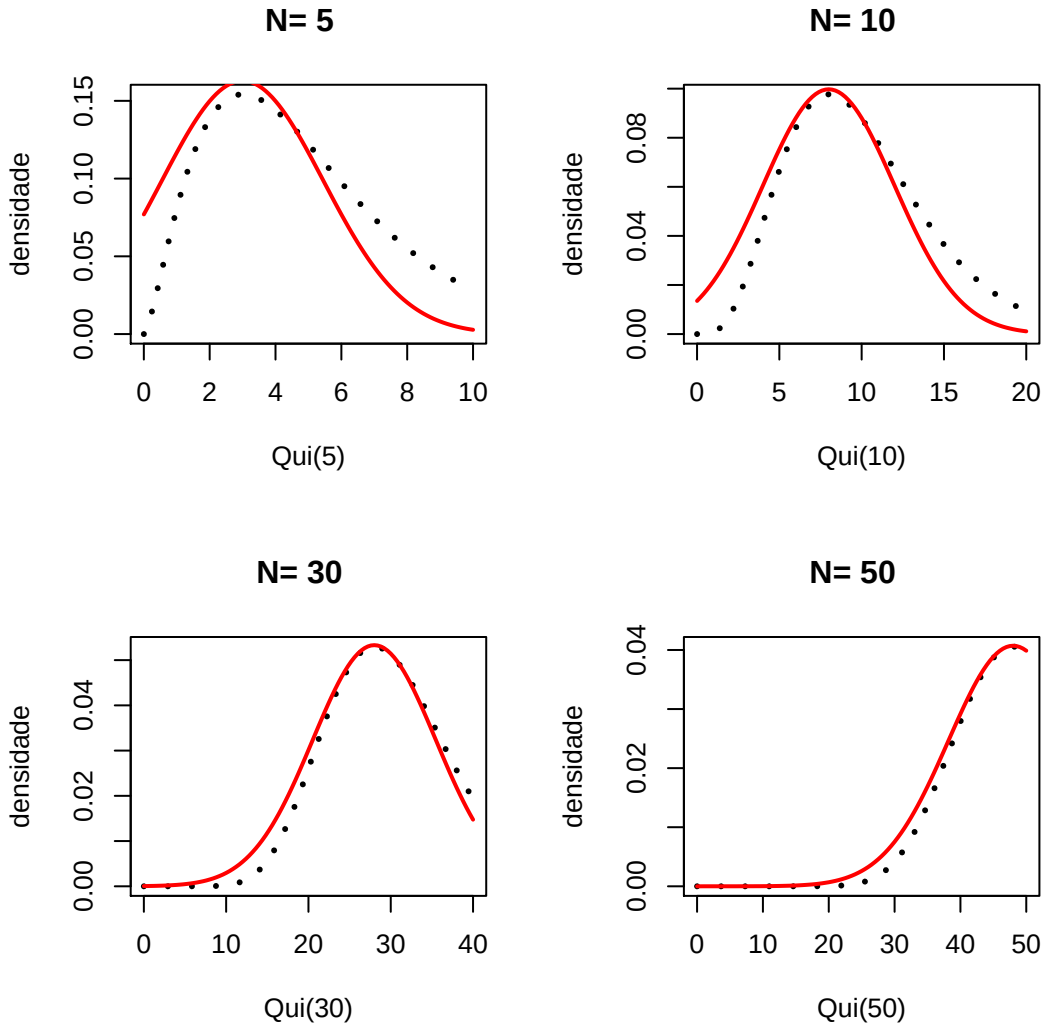


Figura 4: Convergência da aproximação de Laplace(curva pontilhada) para a distribuição qui-quadrado(curva cheia) com o aumento do grau de liberdade N

# 7 Medida de Discrepância

Quando aproximamos uma distribuição p por $\bar{p}$, é necessário avaliar a qualidade da aproximação ou seja o quão próximo ou não está da distribuição exata. Para isso temos várias opções.

- Fazer predições para cada um dos modelos e avaliar a acurácia,
- **usar medidas de discrepâncias( razão, ou diferenças entre as duas distribuições)**
- etc.

Nos debruçaremos sobre o segundo item e usaremos a medida de discrepância de Kullback–Leibler divergence (KLD) e definida em [3].

**Definição 7.1** *A discrepância $\delta[p(\theta) \mid \bar{p}(\theta)]$ entre uma distribuição estritamente positiva $p(\theta)$ e sua aproximação $\bar{p}(\theta)$ é dada por*

$$\delta[p(\theta) \mid \bar{p}(\theta)] = \int p(\theta) log \frac{p(\theta)}{\bar{p}(\theta)} d\theta$$

Vemos que quando as duas distribuições coincidem a discrepância é nula devido ao logaritmo. Com essa medida, o pesquisador pode então de acordo com seus interesses obter aproximações por vários meios( gaussiana, laplace, etc) e decidir por aquela com a menor discrepância.

Seguem dois exemplos de discrepância para aproximações obtidas via método de Laplace nos exemplos 6.1 e 6.2.

**Exemplo 7.1** *Discrepância para a aproximação de Laplace da distribuição beta.*
*Pelos gráficos do exemplo 6.1 vemos que a aproximação se assemelha à distribuição beta com aumento de N. Verificamos esse fato com a discrepância*

*delta. Usando a definição de discrepância dada acima temos para beta:*

$$\delta(. \mid .) \propto \int_0^1 x^{a-1}(1-x)^{b-1} log \left[ \frac{x^{a-1}(1-x)^{b-1}}{exp \dfrac{-(x-x^*)^2}{2\sigma^2}} \right] dx$$

*O gráfico a seguir mostra* $\delta \to 0$ *quando n cresce.*

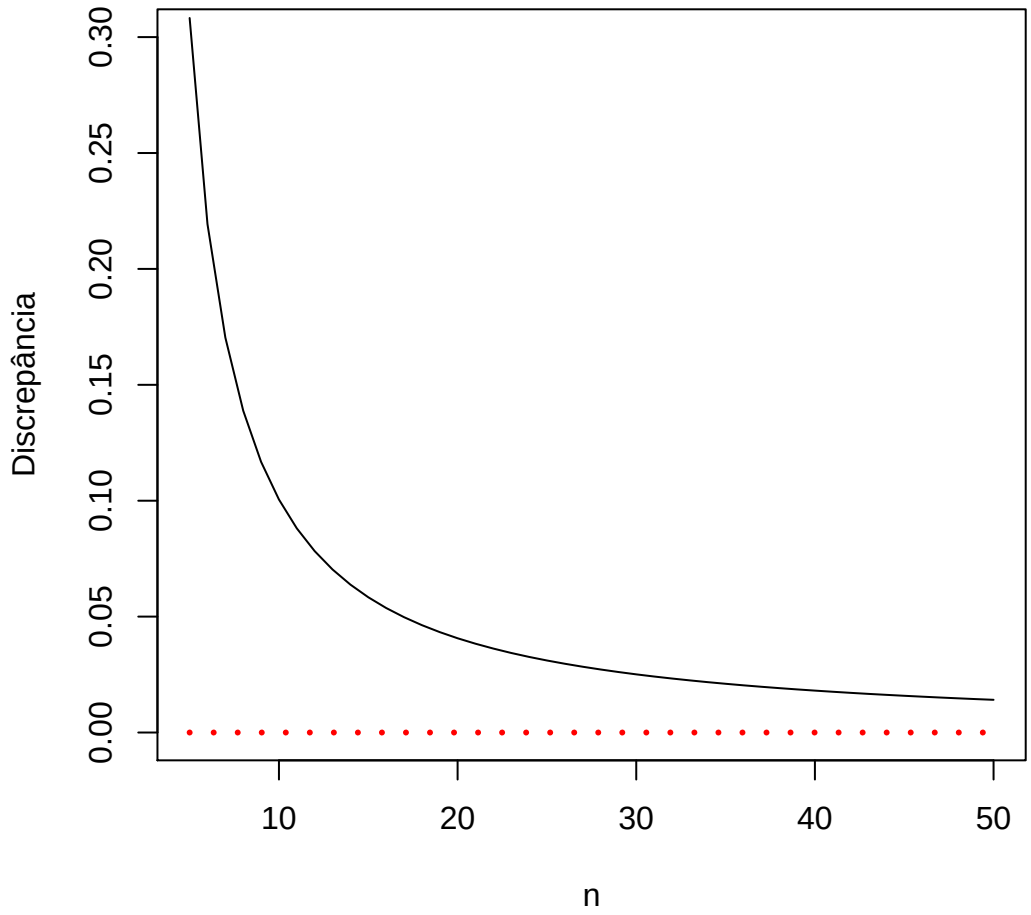


Figura 5: Convergência da discrepância para 0 o aumento do tamanho amostral n

**Exemplo 7.2** *Discrepância para a aproximação de Laplace da distribuição Qui-quadrado*

*No caso da qui-quadrado, utilizamos o método anterior:*

$$\delta(.|.) \propto \int_0^1 x^{n/2-1} exp^{-x/2} log \left[ \frac{x^{n/2-1} exp^{-x/2}}{exp^{\frac{-(x-n+2)^2}{n-2}}} \right] dx$$

*Abaixo segue a convergência de* $\delta \to 0$ *quando n cresce.*

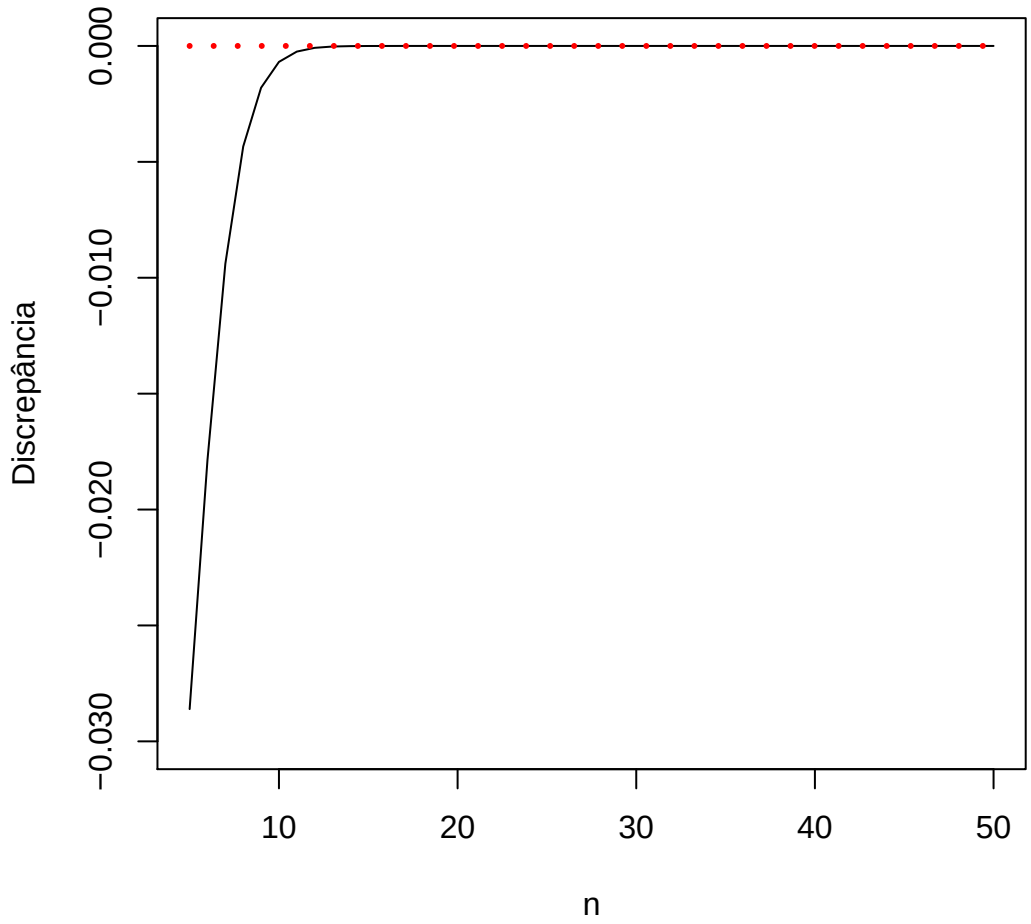


Figura 6: Convergência da discrepância para 0 com o grau de liberdade n

Como tínhamos mencionado, a aproximação de Laplace é uma aproximação assintótica e nos gráficos vemos a diminuição da discrepância com o aumento do tamanho amostral. Esse decréscimo ocorreria mais rapidamente caso a nossa aproximação de Laplace(que assume uma distribuição gaussiana) tivesse os parâmetros ótimos para tal, isto é as melhores média

e variância. Diante desse fato, podemos estar interessados em saber qual a melhor aproximação, ou seja a com menor discrepância, para alguma distribuição dadas algumas condições. Uma pergunta que surge por exemplo é: Qual a melhor aproximação gaussiana(aqui podíamos estipular outras distribuições) $N(\theta \mid \mu, \sigma)$ para uma distribuição uniparamétrica $p(\theta)$ que possui os dois primeiros momentos (média e variância) dados por:

$$\int_{-\infty}^{+\infty} \theta p(\theta) d\theta = m_1$$

$$\int_{-\infty}^{+\infty} (\theta - m_1)^2 p(\theta) d\theta = m_2$$

.

Rescrevendo a expressão da discrepância temos:

$$\begin{aligned} \delta[p(\theta) \mid N(\theta \mid \mu, \sigma)](\mu, \sigma) &= \int_{-\infty}^{+\infty} p(\theta) log \frac{p(\theta)}{N(\theta \mid \mu, \sigma)} d\theta \\ &= \int_{-\infty}^{+\infty} p(\theta) log p(\theta) d\theta - \int_{-\infty}^{+\infty} p(\theta) log N(\theta \mid \mu, \sigma) d\theta \end{aligned} \tag{26}$$

O primeiro termo sendo constante em relação a $\mu$ e $\sigma$, minimizar $\delta(. \mid .)(\mu, \sigma)$ em relação a $\mu$ e $\sigma$ equivale a minimizar apenas o segundo termo.

$N(\theta \mid \mu, \sigma) \propto \frac{1}{\sigma} e^{\frac{-1}{2}\left[\frac{(\theta - \mu)}{\sigma}\right]^2}$ logo minimizaremos:

$$\begin{aligned} \delta(. \mid .)(\mu, \sigma) &\equiv -\int_{-\infty}^{+\infty} p(\theta) log \frac{1}{\sigma} e^{\frac{-1}{2}\left[\frac{(\theta - \mu)}{\sigma}\right]^2} d\theta \\ &\equiv -\int_{-\infty}^{+\infty} p(\theta) log \frac{1}{\sigma} d\theta + \frac{1}{2\sigma^2} \int_{-\infty}^{+\infty} p(\theta)(\theta - \mu)^2 d\theta \\ &\equiv log\sigma + \frac{1}{2\sigma^2} \int_{-\infty}^{+\infty} p(\theta)(\theta - \mu)^2 d\theta \end{aligned} \tag{27}$$

Derivando em relação à $\mu$ e $\sigma$ e igualando a zero obtemos:

$$\frac{d\delta(.\mid .)}{d\mu} = \frac{d}{d\mu}\int_{\infty}^{+\infty} p(\theta)[\theta-\mu]d\theta = -\mu + \int_{\infty}^{+\infty} p(\theta)\theta d\theta = -\mu + m_1 = 0$$

ou seja $\mu = m_1$.

$$\frac{d\delta(.\mid .)}{d\sigma} = \frac{1}{\sigma} - \frac{1}{\sigma^3}\int_{-\infty}^{+\infty} p(\theta)[\theta-\mu]^2 d\theta = 0$$

$$\Updownarrow$$

$$\int_{-\infty}^{+\infty} p(\theta)[\theta-\mu]^2 d\theta = \sigma^2 = \int_{-\infty}^{+\infty} p(\theta)[\theta-m_1]^2 d\theta$$

ou seja $\sigma^2 = m_2$.

Concluímos que a aproximação normal ótima para a distribuição $p(\theta)$ com média $m_1$ e variância $m_2$ é aquela com a mesma média e mesma variância.

# 8 Integrated Nested Laplace Approximation: INLA

Apesar de existir antes da abordagem frequentista, a análise bayesiana foi realmente aceita e difundida devido aos saltos computacionais que tivemos no século XX. É comum escutarmos que as melhores invenções desse século são: *o forno de microondas, a internet, o celular etc.* Todavia, para o analista bayesiano, a melhor invenção é sem dúvida o **amostrador de Gibbs**. O seu sucesso é inegável, devido ao seu uso nas cadeias de Markov para obtenção da distribuição posterior e consequentemente na resolução de muitos problemas teóricos e práticos na estatística.

Porém com o"boom" na quantidade de informações, dados, hoje , é necessário investir em técnicas de otimização de tempo computacional, memória e nessa ótica surge em 2009 o **Integrated Nested Laplace Approximation: INLA**, um software de análise bayesiana desenvolvido em linguagem C por Rue et al [10]. O **R-INLA** é um pacote desenvolvido

para o software R que faz a conexão entre o INLA e o R para efetuar as computações necessárias. Essa ferramenta foi desenvolvida para os chamados **modelos gaussianos latentes**, uma extensão dos modelos lineares generalizados( GLM) dado que muitos problemas de inferência bayesiana se enquadram nessa categoria [10]. Relembremos a estrutura de um GLM para uma variável y da família exponencial:

$$y_i \sim p(y \mid \theta), \quad E(y_i) = \mu_i \eta_i = g(\mu_i) \quad \eta_i = \beta_o + \sum_i^k \beta_k x_{ki}.$$

Os modelos gaussianos latentes são obtidos incluindo uma nova função incorporando efeitos espaciais, temporais, etc e pode ser representado como segue:

$$y_i \sim p(y \mid \theta), \qquad E(y_i) = \mu_i \eta_i = g(\mu_i) \quad \eta_i = \beta_o + \sum_i^k \beta_k x_{ik} + \sum_j^n f_n z_{ij} + \epsilon_i.$$

O modelo é dito gaussiano pois associa-se priores da família gaussiana a: $\beta_k$ , $f_n$, $\epsilon_i$ e latente pois a verdadeira informação y, que não necessariamente é diretamente observável, pode ser obtida via transformação por g e f. Mais detalhes sobre esses modelos podem ser encontrados em [10].

O INLA propõe obter distribuições marginais de maneira rápida, eficiente e seu uso vem sendo difundido desde 2009. Em muitas pesquisas, as marginais são suficientes para responder às perguntas e portanto pode-se lançar mão do INLA. Entretanto, caso o interesse é obter a distribuição conjunta de alguns parâmetros, terá-se que recorrer a outras ferramentas ou de maneira conservadora usar o MCMC, afinal: "Não se mexe no time que está ganhando".

Consideremos o seguinte modelo hierárquico para demonstrar o uso da aproximação de Laplace no INLA:

$$y \mid \theta, \psi = \textstyle\prod_i p(y_i \mid \theta, \psi_2) \;\; modelo \;\; para \;\; os \;\; dados$$
$$\theta \mid \psi \sim p(\theta \mid \psi) \sim Normal(0, \psi_1) \quad prior$$
$$\psi \sim p(\psi_1, \psi_2) \quad \psi_1 \text{ hiperparâmetros (priori sobre priori } \theta) \text{ e } \psi_2 \text{ efeitos}$$

aleatórios sobre y.

Estamos interessados nos parâmetros $\theta$ e $\psi$ dado que baseado neles podemos fazer predições para y.

$$p(\theta_j \mid y) \overset{Marg}{=} \int p(\theta_j, \psi \mid y) d\psi \overset{T.Bayes}{=} \int p(\psi \mid y) p(\theta_j \mid y, \psi) d\psi \tag{28}$$

$$p(\psi_k \mid y) \overset{Marg}{=} \int p(\psi \mid y) d\psi_{-k} \tag{29}$$

Comecemos por determinar os integrandos nas equações 28 e 29:

$$\begin{aligned}
p(\psi \mid y) &\overset{T.B}{=} \frac{p(\theta, \psi \mid y)}{p(\theta \mid \psi, y)} \\
&\overset{T.B}{=} \frac{p(y \mid \theta, \psi) p(\theta, \psi)}{p(y)} \frac{1}{p(\theta \mid \psi, y)} \\
&\overset{yc.id\psi|\theta}{=} \frac{p(y \mid \theta) p(\theta \mid \psi) p(\psi)}{p(y)} \frac{1}{p(\theta \mid \psi, y} \\
&\propto \frac{p(y \mid \theta) p(\theta \mid) p(\psi)}{p(\theta \mid \psi, y)} \\
\bar{p}(\psi \mid y) &\simeq \frac{p(y \mid \theta) p(\theta \mid) p(\psi)}{\bar{p}(\theta \mid \psi, y)})
\end{aligned}$$

Onde T.B $\equiv$ teorema de Bayes, $yc.id\psi \mid \theta \equiv$ y condicionalmente independente de $\psi$ dado $\theta$ e $\bar{p}(\theta \mid \psi, y)$ é a aproximação de Laplace para $p(\theta \mid \psi, y)$ calculada na moda $\theta^*(\psi)$.

Para o outro termo do integrando em 28 temos:

$$
\begin{aligned}
p(\theta_j \mid \psi, y) &\overset{T.Bayes}{=} \frac{p([\theta_j, \theta_{-j}] \mid \psi, y)}{p(\theta_{-j} \mid \theta_j, \psi, y)} \\
&\overset{T.B}{=} \frac{p([\theta_j, \theta_{-j}], \psi \mid y)}{p(\psi \mid y)} \frac{1}{p(\theta_{-j} \mid \theta_j, \psi, y)} \\
&\overset{T.B, yc.id\psi|\theta}{\propto} \frac{p(\psi)p(\theta \mid \psi)p(y \mid \theta)}{p(\theta_{-j} \mid \theta_j, \psi, y)} \\
\bar{p}(\theta_j \mid \psi, y) &\simeq \frac{p(\psi)p(\theta \mid \psi)p(y \mid \theta)}{\bar{p}(\theta_{-j} \mid \theta_j, \psi, y)}
\end{aligned}
\tag{30}
$$

Onde: $\bar{p}(\theta_{-j} \mid \theta_j, \psi, y)$ é a aproximação de Laplace para cada distribuição marginal calculada na moda $\theta^*_{-j}(\theta_j, \psi)$. Fica evidente que no caso de um modelo com um único parâmetro $\theta$ não precisamos do INLA, pois o MCMC é rápido e suficiente.

Resumindo usamos a aproximação de Laplace para obter os integrandos em 28 e 29 e portanto podemos obter as distribuições marginais posteriores para cada $\theta_j$.

# 9 Considerações finais

Fazer predição, em geral pode ser uma tarefa difícil. Todavia temos a estatística para auxiliar nisso. Analisando uma das etapas mais importantes da abordagem bayesiana, obtenção da distribuição posterior, pudemos conhecer uma definição e construção de probabilidade outra que as convencionais ensinadas nos cursos clássicos de estatística sob o prisma da teoria de medida e integração. Nos deparamos inicialmente com o problema de elicitação de prioris uma das origens da falta de unificação da teoria bayesiana, uma área com muitas lacunas a serem preenchidas ainda. Em seguida, adotamos a análise bayesiana objetiva por ela ser axiomatizada e envolver uma matemática atraente que vai dos teoremas da análise real aos problemas isoperimétricos do cálculo variacional. Além disso, identificamos as dificuldades mais encontradas pelos pesquisadores, isto é inexistência de

solução analítica, o que nos incentivou a lançar mão de aproximações em serie de Taylor e assintóticas notadamente o método de Laplace para obter a posterior.

Este estudo foi motivado também pelo sucesso do INLA na obtenção da distribuição posterior marginal assim como pelos artifícios matemáticos usados por ele para tal. Introduzimos também uma medida de discrepância para aproximações cuja interpretação servirá de meio para o pesquisador validar ou não uma aproximação. Cumprimos todas as metas e este trabalho foi um grande incentivo ao futuro estudo das teorias assintóticas matemáticas para aproximação de distribuições, para desenvolvimento de critério, isto é algum adimensional para validar essas aproximações e a possível criação de um pacote estatístico para contribuir ao avanço dessa área que é a teoria bayesiana.

# 10 Anexos

### Códigos para os exemplos 6.1 e 6.2

1* Beta(a,b)

```
# Aproximacao Laplace para Beta(a,b)
rm(list=ls())
y <- seq(0.001,0.999,by=0.001)
par(mfrow=c(2,2))
#1. Beta(a,b)  com a+b = n =  e a/(a+b) = p = 0.7
n <- 5
p <- 0.7
a <- n*p
b <- n - a
c(a,b)
E.x <- a/(a+b)
V.x <- (a*b)/(((a+b)^2)*(a+b+1))
x1<-(a-1)/(a+b-2)
x2<-(a-1)*(b-1)/(a+b-2)^3
curve(dbeta(x,a,b),from=0.001,to=0.999,lwd=2,
      main= "N= 5  ", xlab = "Beta(3.5, 1.5)")
curve(dnorm(x,x1,sqrt(x2)),from=0.001,to=0.999,lwd=2,col="red",ad
```

2* Qui-quadrado(n)

```
# Aproximacao Laplace para Qui-quadrado(n)
rm(list=ls())
y <- seq(0.001,0.999,by=0.001)

par(mfrow=c(2,2))
#1.  n=5
```

```
n <- 5
x1<-n-2
x2<-2*(n-2)
curve(dchisq(x,n),from=0.001,to=10,lty=3,lwd=3,
      main= "N= 5  ", xlab = "Qui(5)",ylab="densidade")
curve(dnorm(x,x1,sqrt(x2)),from=0.001,to=10,lwd=2,col="red",add=T
```

### Códigos para a medida de discrepância

1* Beta(a,b)

```
#Discrepância para beta.
rm(list=ls())
 dx<-rep(0,46)
for(i in 5:50) {
 p <- 0.7
 a <- i*p
 b <- i - a
 x1<-(a-1)/(a+b-2)
 x2<-(a-1)*(b-1)/(a+b-2)^3

z1<-function(x){dbeta(x,a,b)*log( dbeta(x,a,b)/dnorm(x,x1,sqrt(x2
dx[i-4]<-integrate(z1,0.001,1,subdivisions=20)
}

nn<-seq(5,50)



plot(nn,dx,type='l',xlab="n", ylab="Discrepância")
```

2* Qui-quadrado(n)

```
#Discrepância para Qui.
rm(list=ls())
```

```
dx <- rep(0,46)
for (i in 5:50) {

  x1 <- i-2
  x2 <- 2*(i-2)
  z1 <- function(x){dchisq(x,i)*log( dchisq(x,i)/dnorm(x,x1,sqrt(
  dx[i-4] <-integrate(z1,0.001,1,subdivisions=20)
}

nn<-seq(5,50)

plot(nn,dx,type='l',xlab="n", ylab="Discrepância")
```

## Referências


[1] Kolmogorov A., Foundations of the theory of probability, *Chelsea Publishing Company*, New York, USA, 1956.

[2] Bayes T. , An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London* 53, 1763, 370-418.

[3] Bernardo J. M, Smith A. F. M., Bayesian Theory,*John Wiley and Sons, Chichester* , West Sussex, Inglaterra, 2000.

[4] Durrett R. , Probability : Theory and Examples, *Cambridge university press 4th ed.*,Cambridge, Inglaterra, 2013.

[5] E.T. Jaynes, Probability theory: the logic of science,*Cambridge university press*, Cambridge, Inglaterra, 2003.

[6] Fernandez P.J., Introdução à teoria das probabilidades, *Associação Instituto De Matemática Pura e Aplicada- IMPA*, Rio de Janeiro, Brasil, 2005.

[7] Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin B. D., Bayesian Data Analysis ,*Chapman and* Hall/CRC 3*st* Ed, 1995.

[8] Hall B., LaplacesDemon: Software for Bayesian Inference. R package version 11.06.13, URL http://cran.r-project.org/web/packages/LaplacesDemon/index. 2011 html.

[9] Hall B., LaplacesDemon: Software for Bayesian Inference. R package version 12.07.02, URL http://cran.r-project.org/web/packages/LaplacesDemon/index. html. 2012

[10] Havard R., Martino S., Chopin N., Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations, *J. R. Statist. Soc. B* **71**, 2009, 319-392.

[11] Havard R., Held L., Gaussian Markov Random Fields: Theory and applications, *Chapman and* Hall/CRC, 2005.

[12] Kinas P. A., Andrade H. A., Introdução à Análise Bayesiana (com R), *maisQnada*, Porto Alegre, Brasil, 2010.

[13] Laplace P. ,Mémoire sur la Probabilité des Causes par les Evenements , *l'Académie royale des sciences* ,6 , 1774, 621-656.

[14] Laplace P. ,Essai Philosofique sur les Probabilités , 1814.

[15] Paulino C. D., Turkman M. A., Murteira B., Estatística Bayesiana , *Fundação Calouste Gulbenkian* ,Lisboa, Portugal, 2003.

[16] Ribeiro P.J.,Bonat W.H., Krainski E.T., Zeviani W.M., Métodos Computacionais em Inferência Estatística, *20ª SINAPE*, João Pessoa, Brasil, 2012.

[17] Tierney L., Kadane J.B., Accurate Aproximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association* **81**, 1986, 82-86.

[18] Wheeler, J. A. and Zurek, W. H: The physical contents of quantum kinematics and mechanics, *Quantum Theory and Measurement , Princeton University Press* ,Princeton, New Jersey, 1983, pp. 62–84