



Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

**MÉTODO EMPÍRICO PARA CÁLCULO
DO VALOR-E COM AUXÍLIO DO
SOFTWARE R**

Matheus Julio Nascimento

Uberlândia-MG

2020

Matheus Julio Nascimento

**MÉTODO EMPÍRICO PARA CÁLCULO
DO VALOR-E COM AUXÍLIO DO
SOFTWARE R**

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Quintiliano Siqueira Schroden Nomelini

**Uberlândia-MG
2020**



**Universidade Federal de Uberlândia
Faculdade de Matemática**

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, _____ de _____ de 20_____

BANCA EXAMINADORA

Prof. Dr. Quintiliano Siqueira Schroden Nomelini

Prof. Dr. Janser Moura Pereira

Profa. Dra. Maria Imaculada de Sousa Silva

**Uberlândia-MG
2020**

AGRADECIMENTOS

Aos meus pais e meu irmão por todo o aprendizado durante toda a minha vida.

À minha namorada, Letícia, por sempre me apoiar em minhas decisões.

Aos meus companheiros de curso, Cássio, Moacir, Rafael, João, Lia, Tiago e Luiz Felipe por todos os trabalhos, estudos e provas juntos.

Aos amigos de trabalho, Juliano, Giba e George, pelo suporte durante o meu período de graduação.

Obrigado também aos meus orientadores, José Waldemar e Quintiliano, pelos ensinamentos e paciência ao longo do desenvolvimento do trabalho em questão.

RESUMO

O presente trabalho foi desenvolvido com a finalidade de apresentar um algoritmo em linguagem R cujo objetivo é, a partir de uma amostra de uma distribuição *a posteriori* passada pelo usuário do código, calcular empiricamente o valor-*e* (Valor de Evidência Bayesiano) baseado em uma hipótese também fornecida previamente. Com o objetivo de testar a qualidade do algoritmo, ele foi validado em amostras aleatórias geradas de distribuições simétricas e assimétricas já conhecidas, nas quais é possível descobrir o valor exato do valor-*e* através de funções existentes. Com o auxílio do pacote *rjags*, também exemplificou-se o trabalho feito testando hipóteses sobre parâmetros de um modelo de regressão linear Bayesiano múltiplo e sobre os dois parâmetros de uma distribuição Normal. Por fim, ressalta-se que o código apresentado aqui tem potencial para virar uma biblioteca da linguagem R em algum momento posterior.

Palavras-chave: Valor-*e*, Linguagem R, Distribuição a Posteriori, *rjags*.

ABSTRACT

The present work was developed with the purpose of presenting an algorithm in R language which objective is, from a sample of a posterior distribution passed by the user of the code, to empirically calculate the e -value (Bayesian Evidence Value) based on a hypothesis also provided previously. In order to test the quality of the algorithm, it was validated in random samples generated from symmetrical and asymmetric distributions already known, in which it is possible to find out the exact value of the e -value through existing functions. With the help of the *rjags* package, the work done was also exemplified by testing hypotheses about parameters of a multiple Bayesian linear regression model and about the two parameters of a Normal distribution. Finally, it is emphasized that the code presented here has the potential to become a library of the R language at some later time.

Keywords: e -value, R Language, Posterior Distribution, *rjags*.

SUMÁRIO

Lista de Figuras	I
Lista de Tabelas	III
1 Introdução	1
2 Desenvolvimento	5
2.1 Utilização do Teorema de Bayes	5
2.2 Estimaco pontual	6
2.3 Estimaco por regies	7
2.4 Testes de Hipteses Bayesianos	7
2.5 Fator de Bayes	8
2.6 Teste de Jeffreys	9
2.7 Teste de Significncia Genuinamente Bayesiano	10
2.8 Clculo do Valor- <i>e</i> Emprico	11
2.9 Propriedades do cdigo	12
2.9.1 Entrada dos dados	12
2.9.2 Condio de funcionamento	13
2.10 Clculo do Valor- <i>e</i> Emprico: verificaes e aplicaes	14
2.10.1 Experimento 1	14
2.10.2 Experimento 2	15
2.10.3 Experimento 3	20
3 Concluses	23
Referncias Bibliogrficas	25
Apndice A Apndice	27
Apndice B Apndice	29
Apndice C Apndice	31
Apndice D Apndice	33
Apndice E Apndice	37

LISTA DE FIGURAS

2.1	Ilustração da área correspondente ao valor- e	10
2.2	Exemplo de arquivo com $n = 3$	13
2.3	Exemplo de entrada de nome de arquivo e hipóteses para $n = 3$	13
2.4	Gráfico das iterações <i>versus</i> valores amostrados para β_0	16
2.5	Gráfico das iterações <i>versus</i> valores amostrados para β_1	16
2.6	Gráfico das iterações <i>versus</i> valores amostrados para β_2	17
2.7	Gráfico das iterações <i>versus</i> valores amostrados para β_3	17
2.8	Gráfico das iterações <i>versus</i> valores amostrados para β_4	17
2.9	Gráfico das iterações <i>versus</i> valores amostrados para β_5	18
2.10	Gráfico de dispersão entre as variáveis do modelo.	19
2.11	Gráfico das iterações <i>versus</i> valores amostrados para μ	20
2.12	Gráfico das iterações <i>versus</i> valores amostrados para τ	20

LISTA DE TABELAS

2.1	Função perda.	8
2.2	Escala de evidências de Jeffreys para o Fator de Bayes.	8
2.3	Resultados para distribuição Normal(5, 4).	14
2.4	Resultados para distribuição Beta(2, 7).	15
2.5	Critério de convergência de Geweke para os parâmetros do modelo.	18
2.6	Valor- e e intervalo HPD para os parâmetros do modelo.	19
2.7	Critério de convergência de Geweke para μ e τ	20
2.8	Valor- e e intervalo HPD para os parâmetros da Normal(15, 4).	21

1. INTRODUÇÃO

De acordo com [15], entender inferência estatística é importante para analisar dados de maneira adequada. Antes de entender conceitos estatísticos utilizados nas mais diversas áreas do conhecimento, como intervalo de confiança, valor- p e modelos de regressão, é importante buscar compreender a ciência que está por trás desses valores e técnicas.

De uma maneira geral, como define [10], inferência estatística é o processo pelo qual decisões a respeito de uma população são feitas baseadas em certas características calculadas a partir de uma amostra de dados pertencente à população em questão. A partir dessa definição, entende-se que existe a necessidade de realizar afirmações não apenas para os objetos de um estudo qualquer, mas sim para uma população maior, visando a possibilidade de realmente estender a conclusão obtida para um grupo mais geral.

Os resultados de métodos estatísticos baseados em inferência podem ser expressos de algumas maneiras diferentes. Em um teste de hipótese, por exemplo, pode-se inferir algo a partir de uma estatística de teste, um valor probabilístico (valor- p) ou algum intervalo de confiança. No caso de modelos de regressão, as estimativas dos parâmetros são o produto final da inferência. É importante destacar que como parte-se de uma amostra para inferir algo sobre uma população, os resultados dessas análises vêm sempre acompanhados de medidas de erro, confiabilidade ou precisão.

A principal justificativa da utilização de técnicas de inferência estatística está associada à dificuldade de realização de experimentos que envolvam toda a população do estudo em questão. São vários os exemplos em que isso ocorre, como por exemplo pesquisas de pretensão eleitoral ou um estudo para comparação de tamanhos de caule em diferentes espécies de orquídeas. Isto é, seria inviável, por questões financeiras e logísticas, entrevistar todos os brasileiros antes de uma eleição, assim como é impossível medir o tamanho do caule de todas as orquídeas do mundo.

Na tentativa de contextualizar historicamente a evolução da inferência estatística, [7] cita três grandes revoluções paramétricas na área: Laplace (1774-1786), Gauss and Laplace (1809-1828) e Fisher (1912-1956). O detalhamento de cada uma dessas revoluções não é o objetivo principal deste trabalho, porém merecem menção honrosa algumas das descobertas realizadas nos períodos citados, como o princípio da probabilidade inversa de Laplace e a utilização da distribuição Normal como uma aproximação da distribuição da média de grandes amostras.

Dentro do panorama contemporâneo, [12] justifica a importância do tema citando sua vasta aplicabilidade em áreas como: estudos médicos, previsões climáticas, testes de qualidade, mer-

cado financeiro de ações, dentre outras.

Segundo [3], quando existe a intenção de saber alguma informação sobre um determinado parâmetro ψ , existem duas abordagens principais de inferência, a clássica e a Bayesiana.

Como explica [13], métodos clássicos e Bayesianos compartilham os mesmos objetivos. Ambos tentam inferir ou aprender algo sobre uma distribuição assumida ou parâmetros de um modelo baseando-se em uma amostra de dados que fornece apenas informações parciais a respeito do objeto de interesse.

A explicação seguinte sobre como cada uma das abordagens resolve problemas a respeito de parâmetros desconhecidos foi dada por [11]. Em inferências clássicas, todas as probabilidades são condicionadas em parâmetros que são assumidamente conhecidos, o que resulta em:

- testes de significância, nos quais o valor do parâmetro de pelo menos um dos parâmetros é fixado pela hipótese;
- intervalos de confiança.

Na teoria de inferência Bayesiana, os parâmetros também são tratados de maneira probabilística, mas resultando em distribuições de probabilidade que expressam o grau de incerteza a respeito do problema nos seguintes momentos:

- antes da observação da amostra (não dependem dos dados): probabilidades *a priori*;
- depois das observação dos dados (são condicionadas aos dados): probabilidades *a posteriori*;
- sobre dados futuros: probabilidades preditivas.

Como afirma [13], os métodos Bayesianos foram nomeados a partir da abordagem filosófica incorporada na obra do século XVII de Thomas Bayes (1702-1761), um ministro presbiteriano e matemático. Seu manuscrito original sobre probabilidade inversa, incluindo o Teorema de Bayes, foi publicado apenas após a sua morte, em 1763.

A ação de inferir atesta o desconhecimento dos parâmetros e, portanto, hipóteses são construídas afim de possibilitar a tomada de decisão. Sob a ótica Bayesiana, o grau de evidência acerca de uma hipótese é determinado a partir da distribuição *a posteriori*. É comum encontrar metodologias de testes de hipóteses quando estas são compostas, tendo como principais exemplos o método baseado na probabilidade *a posteriori* e Fator de Bayes. Estes métodos, em geral, não podem ser aplicados para o caso de hipótese simples, uma vez que dependem do cálculo de uma probabilidade a partir de distribuições contínuas, ou seja, a probabilidade da variável assumir um valor único é sempre igual zero.

Uma solução para testar hipóteses é apresentada em [2]. Os autores definem o valor-*e* como a evidência da hipótese nula simples H_0 em um teste estatístico Bayesiano, também chamado de FSBT (Full Bayesian Significant Test).

Em [2], os autores definem o valor-*e* para situações em que a forma analítica da distribuição *a posteriori* é conhecida, o que nem sempre é o caso em muitos problemas bayesianos. Logo,

há a necessidade de estabelecer uma metodologia para o cálculo do valor- e nesta situação. Os algoritmos baseados no método de Monte Carlo via cadeias de Markov (MCMC) permitem obter iterativamente amostras das distribuições condicionais completas *a posteriori* dos parâmetros e , sob a hipótese de convergência, estas amostras representam as distribuições *a posteriori* [5].

Portanto, pretende-se com este trabalho apresentar um método para cálculo do valor- e a partir de amostras de distribuições condicionais completas *a posteriori* obtidas com o auxílio do pacote *rjags* do software R.

2. DESENVOLVIMENTO

Nesta seção, foi construído o caminho para explicar o cálculo do valor- e empírico, além das justificativas para a sua utilização em testes de hipóteses simples sobre parâmetros de distribuições *a posteriori*.

Primeiramente, na seção 2.1, é apresentado o Teorema de Bayes, que é a base da Inferência Bayesiana, uma vez que ele possibilita a construção da distribuição *a posteriori*, a partir da qual estimam-se e/ou testam-se hipóteses sobre o(s) parâmetro(s) de interesse, seja de maneira pontual (seção 2.2) ou intervalar (seção 2.3).

Nas seções 2.4, 2.5 e 2.6 foram mostradas propostas já existentes para a realização de testes de hipóteses sobre parâmetros de interesse de uma distribuição *a posteriori*. Existem críticas sobre estes métodos que justificam a proposição, pelos autores, do Teste de Significância Genuinamente Bayesiano (FSBT) [2], no qual é definido o valor- e . Este teste foi apresentado na seção 2.7. O FSBT é uma solução para a realização dos testes de hipóteses simples, principal crítica relacionada aos demais testes apresentados neste trabalho. Porém, os autores do FSBT não o propuseram para situações em que a forma analítica da distribuição *a posteriori* é desconhecida.

A partir da utilização de amostras da distribuição ao invés de sua forma analítica, na seção 2.8 foi apresentada a proposta do valor- e empírico para a realização de testes de hipóteses. O algoritmo em linguagem R criado foi então explicado (seção 2.9) e avaliado (seção 2.10).

Ressalta-se que todas as fórmulas e definições das seções 2.1 a 2.7 têm como referência [14], enquanto as seções 2.8 a 2.10 são propostas desenvolvidas neste trabalho.

2.1 UTILIZAÇÃO DO TEOREMA DE BAYES

Haja vista que toda a Inferência Bayesiana é respaldada pelo Teorema de Bayes, justifica-se a apresentação deste, conforme parágrafos seguintes, de forma a elucidar a atualização do conhecimento previamente existente sobre um parâmetro ou parâmetros a partir dos dados experimentais ou amostrais.

Basicamente, a realização de um experimento $\{X = x\}$ visa reduzir a incerteza sobre um parâmetro de interesse θ . Seu modelo probabilístico antes da realização do experimento é chamado de *a priori* de θ , aquele definido após é *a posteriori* de θ e a informação do experimento utilizada para atualizar a incerteza sobre o parâmetro está na função de verossimilhança.

Tal atualização é realizada baseando-se na operação de Bayes:

$$h(\theta|x) = \frac{h(\theta)f(x|\theta)}{\int_{\Theta} h(\theta)f(x|\theta)d\theta} \propto h(\theta) \times L(\theta; x), \quad (2.1)$$

em que:

- Priori: $h(\theta)$;
- Verossimilhança: $L(\theta; x) = f(x|\theta)$;
- Posteriori: $h(\theta|x)$.

Apesar da distribuição *a posteriori* conter toda informação a respeito do parâmetro de interesse θ , geralmente é interessante resumir essa informação em apenas alguns poucos valores numéricos. É nesse cenário que surgem duas opções, a estimação pontual e a intervalar.

2.2 ESTIMAÇÃO PONTUAL

Para explicar a estimação pontual, podem ser utilizados conceitos de Teoria da Decisão. Defina-se:

- a : uma estimativa de θ ;
- $l(a, \theta)$: função que associa uma perda de modo que quanto maior a distância entre o valor real θ e a , maior é o seu valor.

O estimador pontual de Bayes será o valor de a que minimiza a esperança de uma função perda $l(a, \theta)$ definida para *a posteriori* em questão.

$$E[l(a, \theta)|x] = \int l(a, \theta)p(\theta|x)d\theta \quad (2.2)$$

Os estimadores mais populares e suas respectivas funções perdas são:

- Média *a posteriori* (perda quadrática):

$$l(a, \theta) = (a - \theta)^2 \quad (2.3)$$

- Mediana *a posteriori* (perda absoluta):

$$l(a, \theta) = |a - \theta| \quad (2.4)$$

- Moda *a posteriori* (perda 0-1):

$$l(a, \theta) = \begin{cases} 0, & \text{se } |a - \theta| < \varepsilon, \forall \varepsilon \\ 1, & \text{se } |a - \theta| > \varepsilon, \forall \varepsilon \end{cases} \quad (2.5)$$

Apesar da literatura, em geral, classificar como estimadores bayesianos aqueles obtidos a partir das funções perdas, qualquer medida resumo obtida a partir da distribuição *a posteriori* é uma estimação Bayesiana.

2.3 ESTIMAÇÃO POR REGIÕES

Na visão Bayesiana, o termo que equivale à região de confiança é a Região de Credibilidade $R(x)$, a qual é definida abaixo.

$R(x)$ é uma região de credibilidade γ para θ se,

$$P[\theta \in R(x)|x] = \int_{R(x)} h(\theta|x) d\theta \geq \gamma \quad (2.6)$$

Existem infinitas regiões possíveis que podem ser definidas por 2.6. Aqui, o interesse é selecionar a região que contém os valores mais prováveis *a posteriori*, conhecida como Região HPD (*Highest Posteriori Density*).

Assim, $R(x)$ é uma região de credibilidade γ para θ com densidade *a posteriori* máxima se,

$$\sup_{\theta \notin R(x)} h(\theta|x) \leq c_\gamma \leq \inf_{\theta \in R(x)} h(\theta|x), \quad (2.7)$$

para algum c_γ tal que $P[\theta \in R(x)|x] \geq \gamma$.

Quando a abordagem for por parâmetro, de forma individual, tem-se o Intervalo HPD ao invés de Região HPD e, neste caso, o intervalo pode ser utilizado para testar hipóteses. Se o Intervalo HPD contiver o valor do parâmetro sob a hipótese nula, esta não será rejeitada [1]. Este procedimento será utilizado neste trabalho como forma de auxiliar a verificação da validade do método proposto, o método do valor- e empírico.

Nas seções a seguir, serão apresentados os testes de hipóteses já existentes na literatura Bayesiana assim como a proposta de teste deste trabalho de conclusão de curso.

2.4 TESTES DE HIPÓTESES BAYESIANOS

Se Θ é o espaço paramétrico de θ , um problema de testar hipóteses pode ser descrito como:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta - \Theta_0 \end{cases} \quad (2.8)$$

Para testar tais hipóteses então, um caminho é interpretá-las diretamente e calcular as suas respectivas probabilidades *a posteriori* e depois submetê-las a uma função perda definida como:

Tabela 2.1: Função perda.

Decisão	Real	
	H_0 Verdadeira	H_0 Falsa
Aceitar H_0	A	B
Rejeitar H_0	C	D

De acordo com a Tabela 2.1, o risco ou perda esperada de se aceitar a hipótese nula é:

$$R(\text{Aceitar } H_0) = A \times P(H_0) + B \times [1 - P(H_0)] \quad (2.9)$$

Enquanto o risco ou perda esperada de se rejeitar a hipótese nula é:

$$R(\text{Rejeitar } H_0) = C \times P(H_0) + D \times [1 - P(H_0)] \quad (2.10)$$

Como deseja-se minimizar o risco de decisão, toma-se o menor valor de risco entre os dois calculados e conclui-se em favor dele.

As críticas a serem feitas neste método são:

1. Como ele se baseia no cálculo de probabilidade *a posteriori* através de uma integral, a hipótese a ser testada deve ser composta, visto que a área de um único ponto apenas vale 0;
2. Os valores A, B, C e D definidos para a função perda geralmente são arbitrários.

2.5 FATOR DE BAYES

O Fator de Bayes $B_0(X)$ mede a influência dos dados na alteração da credibilidade relativa de H_0 e H_1 . Matematicamente, ele é definido como a razão entre *Odds a posteriori* e *Odds a priori* (*Odds-ratio*):

$$B_0(X) = \frac{O(H_0|X)}{O(H_0)} = \frac{\frac{P(H_0|X)}{1-P(H_0|X)}}{\frac{P(H_0)}{1-P(H_0)}} \quad (2.11)$$

A decisão do teste baseado no Fator de Bayes é dada de acordo com a Escala de evidências de Jeffreys para o Fator de Bayes [9].

Tabela 2.2: Escala de evidências de Jeffreys para o Fator de Bayes.

$B_0(X)$	Evidência em favor de H_0
<1	negativa (evidência em favor de H_1)
1 a 3	insignificante
3 a 10	moderada
10 a 30	forte
30 a 100	muito forte
>100	decisiva

Assim como no Teste de Hipóteses Bayesiano, a necessidade do cálculo de uma integral para calcular as probabilidades que compõem as *Odds a priori* e *a posteriori* impede a formulação de uma hipótese simples. Além disso, a escala apresentada na Tabela 2.2 não nos serve como medida de tomada de decisão clara, permitindo questionamentos do tipo: "Qual a intensidade da diferença entre uma evidência do tipo *forte* e do tipo *muito forte*" ou "Por que a linha de corte entre uma evidência *insignificante* e uma *moderada* é o valor 3?"

Portanto, uma alternativa para realizar um teste de hipóteses do tipo simples é o teste de Jeffreys.

2.6 TESTE DE JEFFREYS

Aqui não existe a crítica baseada em hipótese composta, ou seja, o Teste de Jeffreys surge exatamente com o intuito de permitir testes de hipóteses do tipo simples:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0 \quad (2.12)$$

Devido à impossibilidade de calcular o Fator de Bayes neste caso, já que ter-se-ia o problema da probabilidade em um ponto qualquer ser zero, [9] defende que a própria formulação da hipótese nula simples induz a um juízo de que θ_0 possui, *a priori*, uma ordem de importância diferente da que é atribuída aos demais valores de θ . Dessa maneira, a distribuição *a priori* será mista da seguinte forma:

$$h(\theta) = \begin{cases} p_0, & \text{se } \theta = \theta_0 \\ (1 - p_0)h_1(\theta), & \text{se } \theta \neq \theta_0, \end{cases} \quad (2.13)$$

com $h_1(\theta)$ sendo a distribuição *a priori* dos valores $\theta \neq \theta_0$ e $0 < p_0 < 1$.

Por consequência, a distribuição *a posteriori* de θ será da forma:

$$h(\theta|x) = \begin{cases} \frac{p_0 L(\theta_0; X)}{p(x)}, & \text{se } \theta = \theta_0 \\ \frac{(1-p_0)h_1(\theta)L(\theta; x)}{p(x)}, & \text{se } \theta \neq \theta_0, \end{cases} \quad (2.14)$$

com:

$$p(x) = p_0 L(\theta_0; x) + (1 - p_0) \int_{\theta \neq \theta_0} h_1(\theta) L(\theta; x) d\theta \quad (2.15)$$

Agora, basta calcular o novo Fator de Bayes:

$$B_0(X) = \frac{O(H_0|X)}{O(H_0)} = \frac{\frac{P(H_0|X)}{1-P(H_0|X)}}{\frac{P(H_0)}{1-P(H_0)}} = \frac{\frac{\frac{p_0 L(X|\theta_0)}{p(x)}}{1 - \frac{p_0 L(X|\theta_0)}{p(x)}}}{\frac{p_0}{1-p_0}} = \frac{L(\theta_0; x)}{\int_{\theta \neq \theta_0} h_1(\theta) L(X|\theta) d\theta} \quad (2.16)$$

Apesar de resolver testes baseados em hipóteses simples, a decisão final do teste de Jeffreys ainda se baseia na Tabela 2.2. Dessa maneira, com o objetivo de deixar a tomada de decisão mais objetiva, surge o Teste de Significância Genuinamente Bayesiano.

2.7 TESTE DE SIGNIFICÂNCIA GENUINAMENTE BAYESIANO

Com o intuito de testar hipóteses precisas baseadas no cálculo da probabilidade *a posteriori* da região HPD, surge o conceito do Teste de Significância Genuinamente Bayesiano (FBST - Full Bayesian Significance Test), o qual é um teste baseado somente no valor-*e* [2].

De acordo com [4], o valor-*e* é a área da *posteriori* no conjunto do espaço paramétrico de θ que consiste nos pontos com densidade menor do que no ponto (máximo) da densidade à posteriori sob H_0 .

Formalmente, tem-se:

$$\text{valor} - e = 1 - P(\theta \in T(x)|X), \quad (2.17)$$

com,

$$T(x) = \{\theta \in \Theta : h(\theta|X) > h(\theta^*|X)\}, \quad (2.18)$$

no qual $\theta^* \in \arg \max h(\theta|X)$, com $\theta \in \Theta$ é o valor de θ que maximiza $h(\theta|X)$ sob H_0 .

A Figura 2.1 explica o conceito de maneira gráfica:

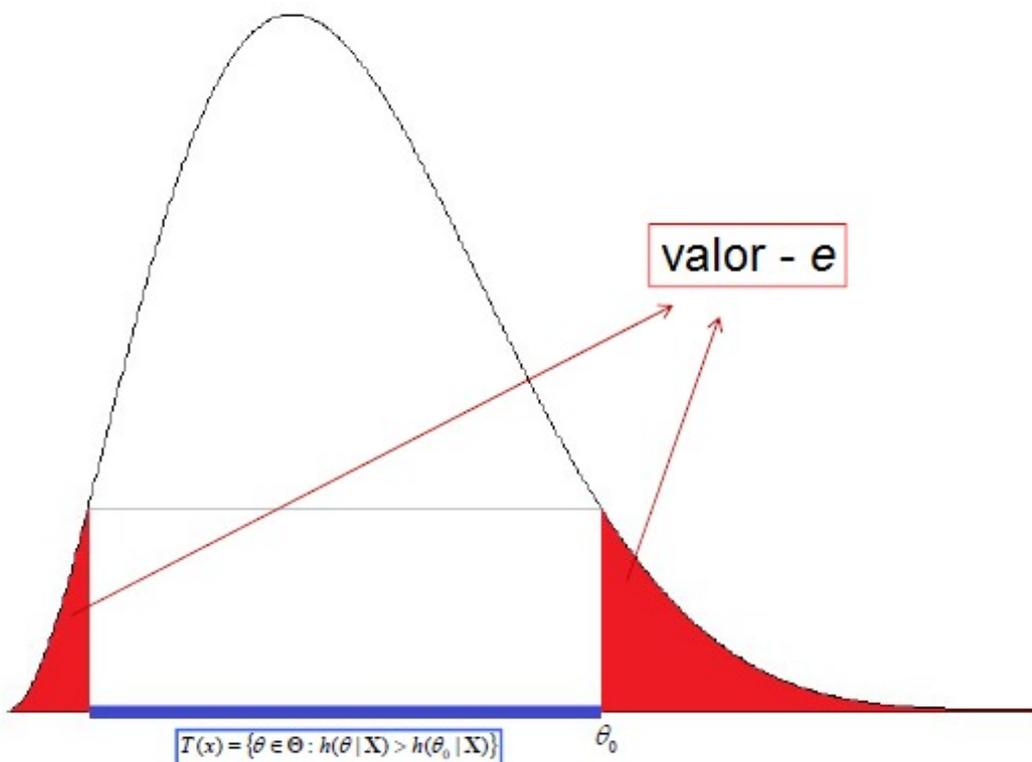


Figura 2.1: Ilustração da área correspondente ao valor-*e*.

Apesar dessa proposta não apresentar problemas como as hipóteses compostas e tomadas de decisão arbitrárias, ela oferece duas dificuldades principais:

1. É necessário o conhecimento da forma analítica da distribuição *a posteriori* em questão, o que envolve a realização de cálculos manuais nem sempre triviais;
2. Não é trivial encontrar o(s) ponto(s) que possui(m) a mesma densidade (altura) que o ponto testado na hipótese nula H_0 . Como nem sempre possui-se uma distribuição simétrica, deve-se traçar uma reta paralela ao eixo X na altura da densidade do ponto da hipótese nula. Depois, deve-se encontrar as raízes da função densidade em relação à reta traçada. Tais raízes são os pontos de interesse e encontrá-las pode exigir o uso de funções auxiliares da linguagem R.

Portanto, o objetivo do trabalho em questão é buscar alternativas a esses dois problemas de maneira a conseguir realizar testes de hipóteses simples/precisas tomando decisões baseadas em um valor de evidência.

2.8 CÁLCULO DO VALOR-*e* EMPÍRICO

A proposta é fornecer uma maneira de calcular o valor-*e* sem a necessidade do conhecimento da forma analítica da distribuição *a posteriori* do parâmetro de interesse. Dada uma amostra da distribuição condicional completa deste parâmetro juntamente com alguma hipótese que se deseja testar sobre ele, o algoritmo proposto gera uma distribuição empírica e calcula o valor-*e*, a partir do qual se toma uma decisão relativa à hipótese nula. Sob a hipótese de convergência do processo amostral, a amostra da distribuição condicional completa representa de forma satisfatória a distribuição de interesse [5]. Critérios de convergência são apresentados em [18], onde o autor apresenta métodos informais e formais. Dentre os métodos informais de verificação de convergência, será utilizado neste trabalho, o gráfico das iterações versus valores amostrados através das distribuições condicionais. Assume-se que um padrão aleatório observado neste gráfico implique em convergência. Por outro lado, como alternativa ao caráter subjetivo dos métodos informais, existem os critérios formais, dentre os quais o critério de Geweke [6], que foi utilizado neste trabalho. A ideia deste teste é comparar uma parte inicial da cadeia com uma parte final utilizando uma estatística de teste que, sob convergência, aproxime-se de uma distribuição normal padrão. Desta forma, quantis da distribuição normal padrão, como por exemplo -1,96 e 1,96 podem ser utilizados como valores críticos para tomada de decisão. Se a estatística de teste calculada for um valor pertencente a este intervalo, a hipótese de convergência não é rejeitada a um nível de significância de 5%.

A regra de tomada de decisão para o valor-*e* é a mesma utilizada para o valor-*p* no caso da inferência clássica, isto é, dado um nível de significância α , rejeita-se H_0 para valor-*e* $< \alpha$ e não rejeita-se caso contrário.

O código (apêndice A) em linguagem R que realiza a proposta acima tem como parte central a seguinte função:

```
(0) e_value<-function(lista, hipotese){  
(1)
```

```

(2)   d<-density(lista)
(3)   dz<-approx(d$x, d$y, xout = lista)$y
(4)   theta<-hipotese
(5)   dtheta<-approx(d$x, d$y, xout = theta)$y
(6)   valor_e<-length(dz[dz<dtheta])/length(dz)
(7)
(8)   return(valor_e)
(9) }

```

A função `e_value` recebe dois parâmetros de entrada, `lista` e `hipotese`. O primeiro contém a amostra de tamanho n do parâmetro de interesse e o segundo o valor do parâmetro sob a hipótese nula. Ambos os parâmetros são inseridos pelo usuário do programa.

Na linha (2), a função `density` é utilizada para calcular a densidade empírica da amostra fornecida. Tal cálculo baseia-se na dispersão dos valores passados em um grid regular de pelo menos 512 pontos (*default*). Utiliza-se então a Transformada Rápida de Fourier para melhorar a aproximação e então transformações lineares para o cálculo da densidade nos pontos especificados [17].

A linha (3) é responsável pela interpolação linear da função densidade gerada no passo anterior [16]. São guardados apenas os valores y que a função `approx` retorna, uma vez que eles são os valores de densidade que serão utilizados no cálculo final do valor- e .

Em (4) e (5), a variável `theta` recebe o valor da hipótese passada como parâmetro para o cálculo da sua densidade de acordo com a função definida em `d` na linha (2).

Finalmente, na linha (6), o valor- e empírico é calculado pela divisão entre a quantidade de valores da amostra que tem densidade inferior à densidade do valor sendo testado e a quantidade de valores total da amostra. O valor- e é então retornado como saída da função em (8).

2.9 PROPRIEDADES DO CÓDIGO

O código criado tem basicamente duas propriedades importantes para seu funcionamento, a maneira de entrada dos dados e a condição de parada por erro.

2.9.1 ENTRADA DOS DADOS

O usuário deve entrar com dois tipos de dados, o nome de um arquivo do tipo `.csv` e os valores a serem testados nas hipóteses.

O arquivo poder conter quantas amostras de condicionais para parâmetros forem desejadas, cada uma em uma coluna, sendo a primeira linha reservada para um título genérico. A Figura 2.2 ilustra uma situação de arquivo para testar hipóteses sobre três parâmetros, por exemplo.

	A	B	C
1	amostra_normal	amostra_gamma	amostra_beta
2	3.81793779483126	1.44549197199353	0.329295954469284
3	5.05318873803233	2.86143374610346	0.0891643439573567
4	1.96689380583627	6.06451713074397	0.439579815390765
5	2.27469330140838	1.23551468980766	0.388921739165195
6	7.35697831206324	1.81727133439068	0.292554882577427
7	3.13169736065327	4.53420390167099	0.197611010671771
8	7.64721129242404	4.07736080701378	0.330991758610594

Figura 2.2: Exemplo de arquivo com $n = 3$.

2.9.2 CONDIÇÃO DE FUNCIONAMENTO

A condição de parada por erro é a parte do código responsável por checar se o número de hipóteses informadas pelo usuário é igual ao número de colunas do arquivo lido. Desta forma, o código não executa caso não exista uma hipótese para cada parâmetro. A parte do algoritmo (apêndice A) responsável por manter esse funcionamento está descrita abaixo.

```
(1) if(length(hipotese) != dim(dados)[2]){
(2)   print('Quantidade de hipóteses incompatível!')
(3) }
```

Basicamente, tem-se um operador condicional, o qual determina que se não existem as condições necessárias para o programa funcionar corretamente, uma mensagem de erro deve ser exibida na tela do usuário.

Um exemplo de entrada está na Figura 2.3. Foram passadas três hipóteses a serem testadas, cada uma referente a uma das três colunas do arquivo *dados.csv*.

```
> nome_arquivo<-readli .... [TRUNCATED]
dados.csv

> hipoteses<-scan()
1: 3
2: 2
3: 1
4:
Read 3 items
```

Figura 2.3: Exemplo de entrada de nome de arquivo e hipóteses para $n = 3$.

Neste caso, as hipóteses sendo testadas seriam, respectivamente:

$$\begin{cases} H_0 : \theta_1 = 3 \\ H_1 : \theta_1 \neq 3 \end{cases} \quad (2.19)$$

$$\begin{cases} H_0 : \theta_2 = 2 \\ H_1 : \theta_2 \neq 2 \end{cases} \quad (2.20)$$

$$\begin{cases} H_0 : \theta_3 = 1 \\ H_1 : \theta_3 \neq 1 \end{cases} \quad (2.21)$$

2.10 CÁLCULO DO VALOR- e EMPÍRICO: VERIFICAÇÕES E APLICAÇÕES

Com o objetivo de exemplificar e verificar a qualidade do algoritmo proposto, foram criadas três situações:

1. Amostras aleatórias das distribuições Normal (simétrica) e Beta (assimétrica). Aqui considera-se que são distribuições *a posteriori*;
2. Amostras de parâmetros de um modelo de regressão linear Bayesiano múltiplo geradas pelo pacote *rjags*;
3. Amostras dos parâmetros μ e σ de uma distribuição Normal geradas pelo pacote *rjags*.

2.10.1 EXPERIMENTO 1

Considera-se dois experimentos quaisquer, nos quais as distribuições *a posteriori* sejam, respectivamente, uma Normal(5, 4) e uma Beta(2, 7). Tal situação foi simulada com o suporte de funções nativas da linguagem R e, assim, amostras de tamanho $n = 50, 2500, 5000$ e 100000 foram criadas (apêndices B e C).

Como em R também existe o suporte nativo para o cálculo da densidade acumulada para algumas distribuições famosas (incluindo Normal e Beta), podemos calcular o valor- e exato para as situações propostas e compará-lo com o resultado do algoritmo do trabalho.

As hipóteses de teste aqui foram definidas de maneira arbitrária.

Para o caso da Normal, seja θ_1 o parâmetro de interesse referente à média da distribuição, testou-se:

$$\begin{cases} H_0 : \theta_1 = 1, 2 \\ H_1 : \theta_1 \neq 1, 2 \end{cases} \quad (2.22)$$

Os resultados estão na Tabela 2.3:

Tabela 2.3: Resultados para distribuição Normal(5, 4).

Valor- e exato	Valores- e empíricos			
	$n = 50$	$n = 2500$	$n = 5000$	$n = 100000$
0,0574	0,0400	0,0660	0,0534	0,0549
HPD _{95%}	[1,727; 8,091]	[0,973; 8,735]	[1,247; 8,931]	[1,169; 8,978]

Portanto, a um nível de significância de 5%, a hipótese nula não seria rejeitada tomando o valor- e exato e aqueles calculados com as amostras de tamanho $n = 2500, 5000$ e 100000 . Analisando os intervalos HPD gerados com 95% de probabilidade, a hipótese nula seria rejeitada para os casos de amostras $n = 50$ e 5000 , isto é, em um dos casos ($n = 5000$), os métodos divergiram quanto a tomada de decisão.

Aqui neste exemplo fica evidente a importância de uma amostra grande para a qualidade do resultado. Outro aspecto a ser notado é que a partir de $n = 5000$ não parece haver ganhos consideráveis na aproximação entre o valor calculado e real.

Já para a Beta gerada, sendo θ_2 o parâmetro de interesse, deseja-se testar sobre a seguinte proporção:

$$\begin{cases} H_0 : \theta_2 = 0,5 \\ H_1 : \theta_2 \neq 0,5 \end{cases} \quad (2.23)$$

Obteve-se como resultados:

Tabela 2.4: Resultados para distribuição Beta(2, 7).

Valor- e exato	Valores- e empíricos			
	$n = 50$	$n = 2500$	$n = 5000$	$n = 100000$
0,0370	0,0000	0,0320	0,0352	0,0366
HPD _{95%}	[0,025; 0,473]	[0,022; 0,447]	[0,008; 0,468]	[0,011; 0,476]

Avaliando a Tabela 2.4, verifica-se que assim como para o valor- e real, para todos os tamanhos de amostra, a um nível de 5% de significância, a hipótese nula seria rejeitada. Todos os intervalos HPD com probabilidade de 95% corroboram com tal decisão.

2.10.2 EXPERIMENTO 2

Para o segundo experimento, utilizou-se o banco de dados *mtcars*, proveniente do pacote *datasets*, o qual é nativo da linguagem R. Todo o código para o desenvolvimento deste experimento encontra-se no apêndice D.

De acordo com [8], os dados foram extraídos da revista *Motor Trend US*, em 1974. Eles dizem respeito ao consumo de combustível e mais dez aspectos de *design* e performance para 32 veículos (modelos 73-74).

O objetivo é ajustar um modelo de regressão linear Bayesiano múltiplo a partir dos dados disponibilizados. A variável dependente do experimento é a quantidade de milhas por galão, a qual será explicada pelas seguintes variáveis independentes:

- x_1 - Deslocamento do motor, que é o volume (em polegadas cúbicas) de todos os pistões dentro dos cilindros do automóvel;
- x_2 - Cavalos de potência, o qual representa a potência do motor;
- x_3 - Relação do eixo traseiro, a qual é uma taxa que está diretamente ligada às rotações de trabalho do motor;
- x_4 - Massa do veículo (em 1000 libras);
- x_5 - Tempo (em segundos) percorrido pelo automóvel em 1/4 de milha.

Deseja-se então estimar parâmetros para cada uma das cinco variáveis descritas mais o intercepto. Isto será feito através do pacote *rjags*, o qual permitirá gerar amostras de tamanho definido pelo usuário (aqui escolheu-se 100.000), utilizando o método do Amostrador de Gibbs [4], para cada parâmetro. Foram descartados os 5000 valores iniciais de cada amostra com o objetivo de diminuir o efeitos dos chutes iniciais, processo conhecido por *burn-in*. Após isso, armazenou-se um valor a cada 10 para compor as amostras a partir das quais foram realizadas as inferências. Estes "pulos" tem como objetivo eliminar uma possível autocorrelação entre os valores, ou seja, espera-se que os valores que compõem a amostra sejam independentes.

Com o objetivo de verificar a convergência das amostras das distribuições condicionais completas, a seguir são apresentados dois métodos: o gráfico das iterações contra os valores amostrados (método informal) e o critério de Geweke (método formal).

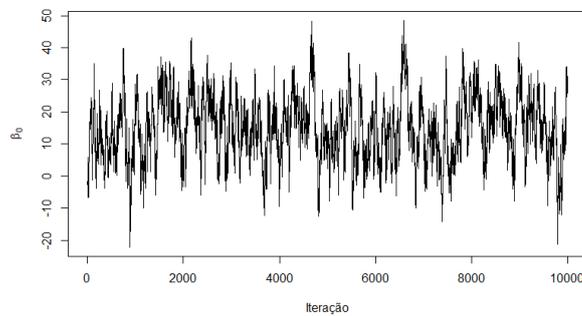


Figura 2.4: Gráfico das iterações *versus* valores amostrados para β_0 .

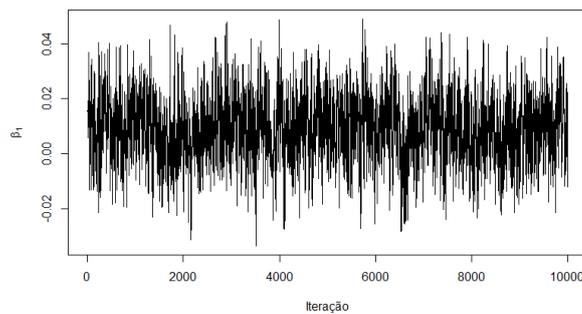


Figura 2.5: Gráfico das iterações *versus* valores amostrados para β_1 .

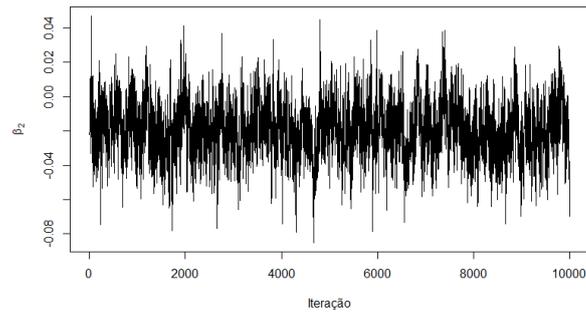


Figura 2.6: Gráfico das iterações *versus* valores amostrados para β_2 .

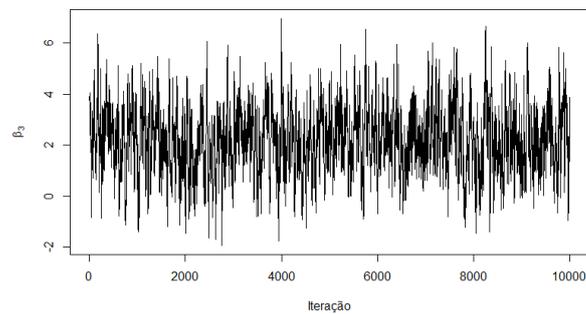


Figura 2.7: Gráfico das iterações *versus* valores amostrados para β_3 .

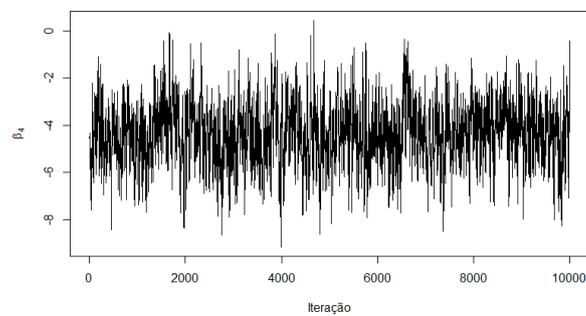


Figura 2.8: Gráfico das iterações *versus* valores amostrados para β_4 .

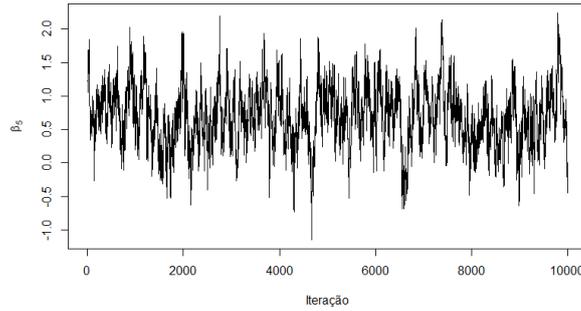


Figura 2.9: Gráfico das iterações *versus* valores amostrados para β_5 .

Em todos os gráficos gerados, nas Figuras de 2.4 a 2.9, é possível identificar que há variação entre os valores gerados, porém a ocorrência deles se dá em uma faixa aparentemente limitada ao longo das iterações. Dessa maneira, conclui-se que, graficamente, a convergência ocorreu para todas as amostras.

A Tabela 2.5 apresenta o critério de Geweke.

Tabela 2.5: Critério de convergência de Geweke para os parâmetros do modelo.

Parâmetro	Estatística do teste (Z)
β_0	0,0930
β_1	0,4372
β_2	-0,4285
β_3	-0,0610
β_4	-0,0565
β_5	-0,0778

A estatística de teste para o critério de Geweke segue uma distribuição Normal padrão, ou seja, a um nível de significância de 5%, não rejeita-se a hipótese nula de convergência caso o valor obtido esteja entre -1,96 e 1,96. Portanto, para todas as amostras geradas, não rejeita-se a condição de convergência.

Com as amostras geradas, foram testadas hipóteses com o objetivo de verificar quais variáveis são ou não significativas para o modelo de regressão, ou seja, para $n = 0, 1, 2, 3, 4, 5$ têm-se:

$$\begin{cases} H_0 : \beta_n = 0 \\ H_1 : \beta_n \neq 0 \end{cases} \quad (2.24)$$

Os resultados do algoritmo para o experimento 2 são apresentados na Tabela 2.6:

Tabela 2.6: Valor- e e intervalo HPD para os parâmetros do modelo.

Parâmetro	Valor- e	HPD _{95%}
β_0	0,20816	[-5,414; 35,750]
β_1	0,40176	[-0,013; 0,031]
β_2	0,23747	[-0,050; 0,009]
β_3	0,10495	[-0,270; 4,874]
β_4	0,00154	[-6,853; -1,939]
β_5	0,11194	[-0,175; 1,559]

Conclui-se que a hipótese nula é rejeitada apenas para o parâmetro referente à variável x_4 , isto é, apenas ele é estatisticamente diferente de 0 a um nível de significância de 5%. O resultado dos intervalos HPD gerados com 95% de probabilidade é o mesmo, apenas x_4 é diferente de 0.

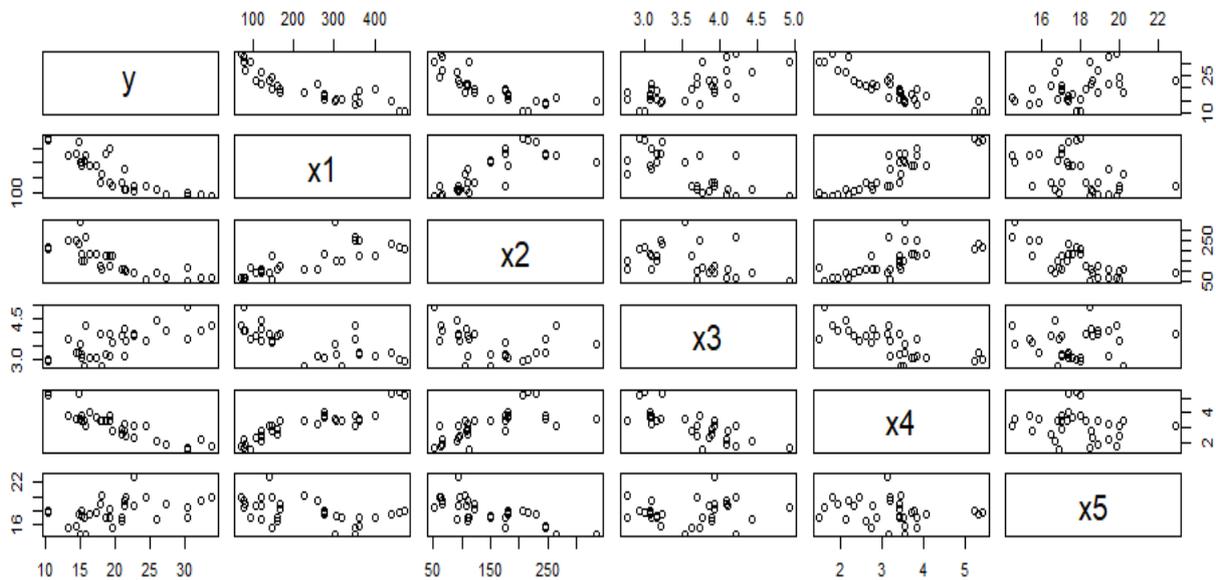


Figura 2.10: Gráfico de dispersão entre as variáveis do modelo.

Os gráficos de dispersão apresentados na Figura 2.10 podem corroborar com os resultados apresentados na Tabela 2.6. Ao avaliar, por exemplo, a dispersão entre a variável resposta e x_4 (única significativamente diferente de zero), observa-se uma tendência linear inversamente proporcional, dando indícios de que um modelo de regressão linear ajustado com tais variáveis possivelmente será bom. Já as variáveis x_3 e x_5 , por exemplo, não aparentam ter correlação linear alta (em módulo) com a variável resposta.

Um algoritmo de seleção de variáveis como o *backward*, por exemplo, poderia ser implementado excluindo a variável em que no teste de hipótese sobre o correspondente parâmetro, obteve-se maior valor- e . O procedimento seria aplicado de tal forma a excluir uma variável por vez e o critério de parada seria quando não houvesse mais parâmetros não significativos. A implementação deste algoritmo não é objeto deste trabalho.

2.10.3 EXPERIMENTO 3

O último experimento (código em apêndice E) também utiliza o pacote *rjags* para gerar amostras de tamanho $n = 10000$ dos parâmetros de uma distribuição Normal(15, 4). Lembrando que as amostras geradas são consideradas distribuições *a posteriori*.

Mais uma vez, realizou-se o *burn-in* de 5000 valores e, posteriormente, "pulos" de 10 valores para compor as amostras a serem utilizadas.

Novamente, a condição de convergência das condicionais completas geradas pelo algoritmo deve ser avaliada.

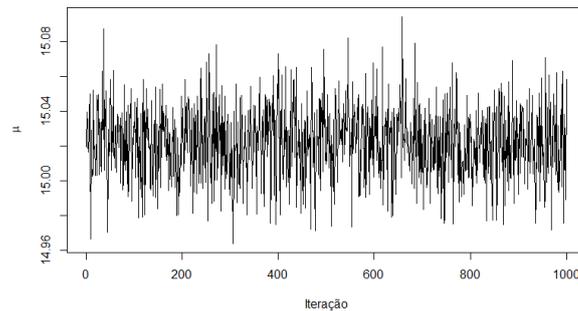


Figura 2.11: Gráfico das iterações *versus* valores amostrados para μ .

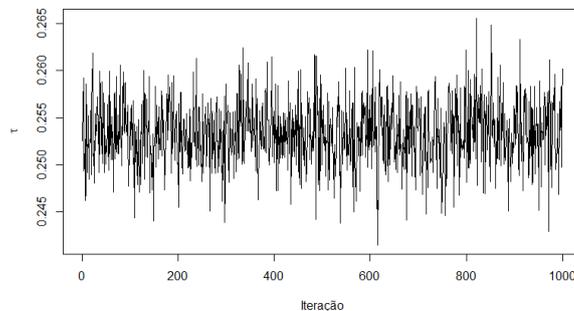


Figura 2.12: Gráfico das iterações *versus* valores amostrados para τ .

Tabela 2.7: Critério de convergência de Geweke para μ e τ .

Parâmetro	Estatística do teste (Z)
μ	0,8266
τ	0,8054

Confirmando os resultados gráficos das Figuras 2.11 e 2.12 (variação de valores dentro de uma faixa semelhante), o critério de Geweke (Tabela 2.7), a um nível de 5% de significância, não rejeita a hipótese nula de convergência para ambas as amostras.

Com a convergência confirmada, para μ , deseja-se testar:

$$\begin{cases} H_0 : \mu = 14,97 \\ H_1 : \mu \neq 14,97 \end{cases} \quad (2.25)$$

Já para τ , com $\tau = 1/\sigma^2$:

$$\begin{cases} H_0 : \tau = 0,25 \\ H_1 : \tau \neq 0,25 \end{cases} \quad (2.26)$$

Conforme observa-se na Tabela 2.8, a um nível de 5% de significância, não rejeita-se a hipótese nula de que μ seja igual a 14,97 e nem rejeita-se que τ vale 0,25. Os intervalos HPD com 95% de probabilidade gerados confirmam a decisão do algoritmo.

Tabela 2.8: Valor- e e intervalo HPD para os parâmetros da Normal(15, 4).

Parâmetro	Valor- e	HPD _{95%}
μ	0,543	[14,941; 15,018]
τ	0,090	[0,249; 0,263]

3. CONCLUSÕES

Neste trabalho, foi apresentada uma maneira de testar hipóteses dentro do cenário Bayesiano. Para isto, sugeriu-se o cálculo do valor- e empírico a partir de amostras de uma distribuição *a posteriori* qualquer. Desta maneira, o método proporciona três vantagens principais em relação a outros já existentes: possibilidade de testar uma hipótese nula simples (ela assume apenas um valor), não existe a necessidade do conhecimento da forma analítica da distribuição *a posteriori* do parâmetro de interesse e a conclusão é feita a partir de um único valor, o valor- e , comparando-o com o nível de significância desejado.

Para a validação do método, amostras de testes foram utilizadas. A partir delas, ilustrou-se e verificou-se a qualidade do algoritmo conforme o aumento do tamanho da amostra. O primeiro experimento apresentado, por exemplo, mostra como o valor- e empírico fica próximo do valor- e real para o maior valor n testado. A necessidade de amostras relativamente grandes não deve ser considerada uma desvantagem, visto que o pacote *rjags* facilita essa ação.

Portanto, conclui-se a validade do algoritmo criado. Destaca-se que o código descrito pode se tornar uma biblioteca da linguagem R em algum momento futuro visando a divulgação de uma ferramenta útil para outros pesquisadores ou usuários do software R.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] Box, G. E. e Tiao, G. C.: *Bayesian Inference in Statistical Analysis*, Addison-Wesley. Reading, MA, 1973.
- [2] Bragança P., C. A. de e Stern, J. M.: *Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses*. 1999. https://www.researchgate.net/publication/26545096_Evidence_and_Credibility_Full_Bayesian_Significance_Test_for_Precise_Hypotheses.
- [3] Cox, D. R.: *Principles of Statistical Inference*. Cambridge University Press, 1ª ed., 2006.
- [4] Ehlers, R. S.: *Inferência Bayesianas*. RICARDO SANDES EHLERS, 5ª ed., 2007.
- [5] Gelfand, A. E. e Smith, A. F.: *Sampling-based approaches to calculating marginal densities*. Journal of the American statistical association, 85(410):398–409, 1990.
- [6] Geweke, J.: *Evaluating the accuracy of sampling-based approaches to calculating posterior moments*. Clarendon, 1992.
- [7] Hald, A.: *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer, 1ª ed., 2004.
- [8] Henderson, H. V. e Velleman, P. F.: *Building multiple regression models interactively*, 1981.
- [9] Jeffreys, H.: *Theory of Probability*. Oxford Univ. Press., 3ª ed., 1961.
- [10] Johnson, L. L., Borkowf, C. B. e Alber, P. S.: *Principles and Practice of Clinical Research*. Academic Press, 2ª ed., 2007.
- [11] Lecoutre, B.: *The Bayesian Approach to Experimental Data Analysis - Essential Statistical Methods for Medical Statistics*. North Holland, 1ª ed., 2010.
- [12] Marsh, J.: *Statistics Role In Real Life*, 2014. https://medium.com/@john_marsh7/statistics-role-in-real-life-a6ba727e0ad8, acessado em 12/12/2019.
- [13] Martz, H. F. e Waller, R. A.: *Methods in Experimental Physics*. World Scientific, 2ª ed., 1994.
- [14] Nakano, E. Y.: *Testes de Hipóteses Bayesianos*. 62ª RBras, 1ª ed., 2017.

- [15] Petter, L. e Fagerland, M.W.: *Research in Medical and Biological Sciences*. Academic Press, 2^a ed., 2015.
- [16] R: *Interpolation Functions*, 2020. <https://www.math.ucla.edu/~anderson/rw1001/library/base/html/approxfun.html>, acessado em 03/01/2020.
- [17] R: *Kernel Density Estimation*, 2020. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/density>, acessado em 03/01/2020.
- [18] Roy, V.: *Convergence Diagnostics for Markov Chain Monte Carlo*. Annual Review of Statistics and Its Application, 7, 2019.

A. APÊNDICE

O código deste apêndice é responsável pela execução do algoritmo criado neste trabalho.

```
#####  
  
# Dados a serem preenchidos pelo usuário  
nome_arquivo<-readline()  
hipoteses<-scan()  
  
#####  
  
# Leitura dos dados  
dados<-read.csv(nome_arquivo, header = TRUE)  
  
# Função para cálculo do valor-e empírico  
e_value<-function(lista, hipotese){  
  d<-density(lista)  
  
  dz<-approx(d$x, d$y, xout = lista)$y  
  
  theta<-hipotese  
  
  dtheta<-approx(d$x, d$y, xout = theta)$y  
  
  valor_e<-length(dz[dz<dtheta])/length(dz)  
  
  return(valor_e)  
}  
  
# 'Loop' principal do código com mensagem de erro  
if(length(hipoteses) != dim(dados)[2]){  
  print('Quantidade de hipóteses incompatível!')  
} else{  
  resultado<-vector()
```

```
for(i in 1:length(dados)){  
  valor<-e_value(dados[, i], hipoteses[i])  
  resultado[i]<-valor  
}  
print(resultado)  
}
```

B. APÊNDICE

Código responsável pela geração da amostra da distribuição Normal utilizada no Experimento 1.

```
# Setando seed para reprodução de resultado
set.seed(11)

# Setando o workspace
setwd('C:/Users/32033_matheus/Downloads')

# Definindo tamanho da amostra
n<-50

# Gerando amostra aleatória da distribuição Normal
amostra_normal<-rnorm(n, mean = 5, sd = 2)

# Gravando a amostra gerada em um arquivo .csv
aux<-data.frame(amostra_normal)
write.csv(aux, file = 'dados.csv', row.names = FALSE)

# Intervalo HPD
library(coda)
HPDinterval(mcmc(aux), prob = 0.95)

# Calculando valor-e real
# Normal
area1<-pnorm(1.2, 5, 2)
area2<-1-pnorm(8.8, 5, 2)
valor_e<-area1+area2
```


C. APÊNDICE

Código responsável pela geração da amostra da distribuição Beta utilizada no Experimento 1.

```
# Setando seed para posterior reproducao de resultado
set.seed(11)

# Setando o workspace
setwd('C:/Users/32033_matheus/Downloads')

# Definindo tamanho da amostra
n<-50

# Gerando amostra aleatória da distribuição Beta
amostra_beta<-rbeta(n, shape1 = 2, shape2 = 7)

# Gravando a amostra gerada em um arquivo .csv
aux<-data.frame(amostra_beta)
write.csv(aux, file = 'dados.csv', row.names = FALSE)

# Intervalo HPD
library(coda)
HPDinterval(mcmc(aux), prob = 0.95)

# Calculando valor-e real
# Beta
hipo<-0.5

g1<-function(x) (x^(2-1))*(exp(-x))
g2<-function(x) (x^(7-1))*(exp(-x))
g3<-function(x) (x^(9-1))*(exp(-x))

integral1<-integrate(g1, 0, Inf)
val1<-integral1$value
```

```
integral2<-integrate(g2, 0, Inf)
val2<-integral2$value

integral3<-integrate(g3, 0, Inf)
val3<-integral3$value

high<-dbeta(hipo, 2, 7)
dens_beta<-function(x) ((val3/(val1+val2))*((x^(2-1))*((1-x)^(7-1)))) - high

library(rootSolve)
raizes<-uniroot.all(dens_beta, c(-1, 1))

area1<-pbeta(raizes[1], 2, 7)
area2<-1-pbeta(raizes[2], 2, 7)
valor_e<-area1+area2
```

D. APÊNDICE

Código utilizado para a realização do Experimento 2

```
# Utilizando o pacote rjags para amostrar
# coeficientes de um modelo de regressão linear
# para teste

# Carregando pacotes que serão utilizados
library(rjags)
library(coda)

# Setando o workspace
setwd('C:/Users/32033_matheus/Downloads')

# Carregando o banco de dados 'mtcars' para o
# nosso ambiente
data(mtcars)

# Escolhendo as variáveis que farão parte da
# nossa análise
y<-mtcars$mpg
x1<-mtcars$disp
x2<-mtcars$hp
x3<-mtcars$drat
x4<-mtcars$wt
x5<-mtcars$qsec

# Escrevendo as variáveis em um arquivo 'csv'
aux<-data.frame(y, x1, x2, x3, x4, x5)
write.csv(aux, file = 'mtcars.csv', row.names = FALSE)

# Carregando o arquivo/variáveis
data<-read.csv('mtcars.csv')
```

```
y<-data[, 1]
x1<-data[, 2]
x2<-data[, 3]
x3<-data[, 4]
x4<-data[, 5]
x5<-data[, 6]

n<-nrow(data)

# Definindo valores iniciais para os parâmetros
# de interesse do modelo

model.inits<-list(tau = 1, beta0 = 1, beta1 = 0, beta2 = 0,
                  beta3 = 0, beta4 = 0, beta5 = 0)

# Definindo número de iterações
iterations<-100000

# Definindo o modelo
modelstring='model{

  for(i in 1:n){
    y[i]~dnorm(mu[i], tau)
    mu[i]<-beta0 + beta1*x1[i] + beta2*x2[i] + beta3*x3[i] + beta4*x4[i]
              + beta5*x5[i]
  }

  # prioris
  tau~dgamma(0.01, 0.01)

  beta0~dnorm(0.0, 0.001)
  beta1~dnorm(0.0, 0.001)
  beta2~dnorm(0.0, 0.001)
  beta3~dnorm(0.0, 0.001)
  beta4~dnorm(0.0, 0.001)
  beta5~dnorm(0.0, 0.001)

}'

# Rodando o modelo jags
```

```
model<-jags.model(textConnection(modelstring),
                  data = list(n = n, x1 = x1, x2 = x2, x3 = x3, x4 = x4,
                              x5 = x5), inits = model.inits)

update(model, n.iter = 5000)

output<-coda.samples(model = model, c('beta0', 'beta1', 'beta2', 'beta3',
'beta4', 'beta5'), n.iter = iterations, thin = 10)

summary(window(output))

m<-as.matrix(mcmc.list(output))

# Gravando as amostras em um arquivo .csv
aux<-data.frame(m[, 2], m[, 3], m[, 4], m[, 5], m[, 6], m[, 7])
write.csv(aux, file = 'modelo-rjags.csv', row.names = FALSE)

# Verificação informal de convergência
plot.ts(m[, 2], xlab="Iteração", ylab=expression(mu))
plot.ts(m[, 3], xlab="Iteração", ylab=expression(mu))
plot.ts(m[, 4], xlab="Iteração", ylab=expression(mu))
plot.ts(m[, 5], xlab="Iteração", ylab=expression(mu))
plot.ts(m[, 6], xlab="Iteração", ylab=expression(mu))
plot.ts(m[, 7], xlab="Iteração", ylab=expression(mu))

# Verificação formal de convergência (critério de Geweke)
geweke.diag(aux[, 1])
geweke.diag(aux[, 2])
geweke.diag(aux[, 3])
geweke.diag(aux[, 4])
geweke.diag(aux[, 5])
geweke.diag(aux[, 6])

# Intervalo HPD
HPDinterval(mcmc(aux), prob = 0.95)
```


E. APÊNDICE

Código utilizado para a realização do Experimento 3

```
# Setando a seed para reprodução de resultados
set.seed(11)

# Carregando pacotes necessários
library(rjags)
library(coda)

# Setando o diretório
setwd('C:/Users/32033_matheus/Downloads')

x<-rnorm(10000, 15, 2)

data<-list(x = x, n = length(x))

hyper<-list(a = 0.01, b = 0.01, c = 10, d = 0.01)

init<-list(mu = 0, tau = 1)

modelstring='
  model {
    for (i in 1:n) {
      x[i]~dnorm(mu,tau)
    }
    mu~dnorm(c,d)
    tau~dgamma(a,b)
  }
'

# Rodando o modelo jags
model<-jags.model(textConnection(modelstring), data = append(data, hyper),
                  inits = init)
```

```
update(model, n.iter = 5000)

output<-coda.samples(model = model,variable.names = c("mu", "tau"),
                     n.iter = 10000, thin = 10)
print(summary(output))

m<-as.matrix(mcmc.list(output))

# Gravando as amostras em um arquivo .csv
aux<-data.frame(m[, 1], m[, 2])
write.csv(aux, file = 'mi-sigma2.csv', row.names = FALSE)

# Verificação informal de convergência
plot.ts(m[, 1], xlab="Iteração", ylab=expression(mu))
plot.ts(m[, 2], xlab="Iteração", ylab=expression(tau))

# Verificação formal de convergência (critério de Geweke)
geweke.diag(aux[, 1])
geweke.diag(aux[, 2])

# Intervalo HPD
HPDinterval(mcmc(aux), prob = 0.95)
```