**Relação entre níveis de significância
Bayesiano e freqüentista:
e-value e *p-value* em tabelas de contingência**

Cátia Petri

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO DE MESTRE
EM
CIÊNCIAS

Área de concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

São Paulo, fevereiro de 2007.

**Relação entre níveis de significância
Bayesiano e freqüentista:
e-value e *p-value* em tabelas de contingência**

**Este exemplar corresponde à redação final da
dissertação devidamente corrigida e defendida
por Cátia Petri e aprovada pela Comissão Julgadora.**

São Paulo, 20 de Abril de 2005.

Banca Examinadora:

Prof. Dr. Carlos Alberto de Bragança Pereira (orientador) - IME/USP

Prof. Dr. José Afonso Mazzon - FEA/USP

Prof. Dr. Sergio Wechsler - IME/USP

*À memória de Elizabeth,
minha mãe querida*

Agradecimentos

Em primeiro lugar, agradeço ao grande mestre e amigo, o Professor Carlinhos, pela orientação nesta dissertação e por todos os ensinamentos que valerão para a vida inteira.

Agradeço ao meu pai, João, por todos os conselhos perfeitos que me guiaram até aqui. Aos meus irmãos, João e Maciel, e meu sobrinho Marquinhos, pelo carinho e apoio sempre.

Ao Danillo Nakano e aos amigos do IME, que não me deixaram desistir dos estudos diante das dificuldades da vida.

Ao grande amigo Paulo Oliveira, da Poli, que com seu conhecimento e dedicação tornou possível a otimização dos programas aqui utilizados.

À Universidade de São Paulo e ao Instituto de Matemática e Estatística pela oportunidade concedida de aperfeiçoar meus estudos.

Resumo

O FBST (*Full Bayesian Significance Test*) é um procedimento para testar hipóteses precisas, apresentado por Pereira e Stern (1999), e baseado no cálculo da probabilidade posterior do conjunto tangente ao conjunto que define a hipótese nula. Este procedimento é uma alternativa *Bayesiana* aos testes de significância usuais. Neste trabalho, estudamos a relação entre os resultados do FBST e de um teste freqüentista, o TRVG (Teste da Razão de Verossimilhanças Generalizado), através de alguns problemas clássicos de testes de hipóteses. Apresentamos, também, todos os procedimentos computacionais utilizados para a resolução automática dos dois testes para grandes amostras, necessária ao estudo da relação entre os testes.

Abstract

FBST (Full Bayesian Significance Test) is a procedure to test precise hypotheses, presented by Pereira and Stern (1999), which is based on the calculus of the posterior probability of the set tangent to the set that defines the null hypothesis. This procedure is a Bayesian alternative to the usual significance tests. In the present work we study the relation between the FBST's results and those of a frequentist test, GLRT (Generalised Likelihood Ratio Test) through some classical problems in hypothesis testing. We also present all computer procedures that compose the automatic solutions for applying FBST and GLRT on big samples what was necessary for studying the relation between both tests.

Sumário

1. Introdução	9
2. Distribuição de Dirichlet.....	11
2.1 Definição.....	11
2.2 Uma Propriedade.....	11
2.3 Uma Conjectura	12
2.4 Família Conjugada de Distribuições	13
3. Testes Bayesiano e Clássico (Freqüentista)	16
3.1 O FBST (Full Bayesian Significance Test).....	16
3.2 O TRVG (Teste da Razão de Verossimilhanças Generalizado)	18
4. Aplicações	20
4.1 Teste para Proporção	21
4.1.1 Hipótese Nula.....	22
4.1.2 FBST	22
4.1.3 TRVG	23
4.1.4 Resultados e Comparação.....	24
4.2 Teste para Homogeneidade de Proporções	28
4.2.1 Hipótese Nula.....	29
4.2.2 FBST	29
4.2.3 TRVG	31
4.2.4 Resultados e Comparação.....	32
4.3 Teste de Homogeneidade de Marginais (O Problema de McNemar)	38
4.3.1 Hipótese Nula.....	39
4.3.2 FBST	39
4.3.3 TRVG	40
4.3.4 Resultados e Comparação.....	41
4.4 Teste do Equilíbrio Populacional de Hardy-Weinberg	46
4.4.1 Hipótese Nula.....	47
4.4.2 FBST	47
4.4.3 TRVG	48
4.4.4 Resultados e Comparação.....	50
4.5 Teste de Independência	55
4.5.1 Hipótese Nula.....	57

4.5.2 FBST	57
4.5.3 TRVG	58
4.5.4 Resultados e Comparação	60
5. Considerações Finais	65
6. Referências Bibliográficas	67
A. Anexo - Programação no MatLab	68
A.0 O Ajuste da função Beta Acumulada	68
A.1 Teste para Proporção	70
A.1.1 FBST	70
A.1.2 TRVG	71
A.1.3 O programa para calcular as duas estatísticas para grandes amostras ...	71
A.2 Teste para Homogeneidade de Proporções	74
A.2.1 FBST	74
A.2.2 TRVG	75
A.2.3 O programa para calcular as duas estatísticas para grandes amostras ...	76
A.3 Teste de Homogeneidade de Marginais (O Problema de McNemar)	79
A.3.1 FBST	79
A.3.2 TRVG	80
A.3.3 O programa para calcular as duas estatísticas para grandes amostras ...	81
A.4 Teste do Equilíbrio Populacional de Hardy-Weinberg	83
A.4.1 FBST	83
A.4.2 TRVG	85
A.4.3 O programa para calcular as duas estatísticas para grandes amostras ...	86
A.5 Teste de Independência	88
A.5.1 FBST	88
A.5.2 TRVG	90
A.5.3 O programa para calcular as duas estatísticas para grandes amostras ...	91

1. Introdução

A literatura estatística está repleta de procedimentos que visam testar hipóteses estatísticas. Este trabalho se restringe aos testes de significância. Por teste de significância entenda-se um procedimento criado para medir a consistência dos dados com a hipótese sendo testada, denominada hipótese nula. Na literatura estatística freqüentista, o cálculo do valor p (o *p-value*) é o exemplo mais conhecido de tais procedimentos. Diversos métodos estão disponíveis para calcular o valor p . Recentemente, uma alternativa Bayesiana foi criada, e o valor e (*e-value*) passa a ser a alternativa Bayesiana do valor p .

Esta dissertação irá focar a resolução de 5 problemas estatísticos amplamente divulgados na literatura através da utilização de dois testes estatísticos de significância:

- i) Um teste Bayesiano: o *Full Bayesian Significance Test* (FBST) ou Teste de Significância Genuinamente Bayesiano, baseado na distribuição a posteriori, através do qual será calculado o *e-value*;
- ii) Um teste Clássico: o Teste da Razão de Verossimilhanças Generalizado (TRVG), baseado na razão entre os máximos - geral e sob a hipótese nula - através da qual será calculado o *p-value*.

Os problemas aqui estudados serão baseados em hipóteses precisas. Entende-se por hipótese precisa aquela definida em um subespaço do espaço paramétrico cuja dimensão é menor do que a dimensão do espaço paramétrico original.

Os problemas estudados são:

- 1) Teste para a Proporção;
- 2) Teste para Homogeneidade de Proporções;
- 3) Teste de Homogeneidade de Marginais (o problema de McNemar);
- 4) Teste do Equilíbrio Populacional de Hardy-Weinberg; e
- 5) Teste de Independência.

O objetivo desta dissertação é apresentar ambos os cálculos, *e-value* e *p-value*, para os diferentes problemas, apresentar os programas utilizados para a aplicação dos testes para amostras consideradas grandes e, finalmente, determinar o tipo de relação existente entre *p-value* e *e-value* para cada problema. Na verdade

mostraremos que algumas funções Beta acumuladas realizam bem o papel de relacionar o *e-value* ao *p-value*.

O Capítulo 2 apresenta a distribuição de Dirichlet e os principais resultados necessários para a aplicação do FBST. O Capítulo 3 apresenta de forma sucinta o FBST e o TRVG. O Capítulo 4 apresenta os problemas com as definições das hipóteses de interesse, a resolução pelos dois métodos, a comparação entre os resultados e o melhor ajuste entre eles. Toda a programação foi feita no *software* MatLab, versão 6.5.0.180913a *release* 13, e encontra-se comentada no Anexo e disponível em CD. Também estão disponíveis no CD todos os dados obtidos nos exercícios aqui resolvidos e mostrados nos gráficos.

2. Distribuição de Dirichlet

2.1 Definição

Um vetor aleatório $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, com $\theta_i > 0$ e $\mathbf{1}'\boldsymbol{\theta} = 1$, tem distribuição de Dirichlet de ordem k com parâmetros $\mathbf{a} = (a_1, a_2, \dots, a_k)$, $a_i > 0$, se a densidade de $\boldsymbol{\theta}$ é a função

$$g(\boldsymbol{\theta}) = \Gamma\left(\sum_{i=1}^k a_i\right) \prod_{i=1}^k \frac{\theta_i^{a_i-1}}{\Gamma(a_i)},$$

onde $\Gamma(a_i)$ é a função Gama avaliada no ponto a_i . Em símbolos, escreve-se:

$$\boldsymbol{\theta}|\mathbf{a} \sim D_k(\mathbf{a}).$$

2.2 Uma Propriedade

Considerando as componentes do vetor aleatório $\mathbf{z} = (z_1, z_2, \dots, z_k)$ como variáveis possuindo distribuição Gama, mutuamente independentes, com parâmetros $(\mathbf{a}, b)$, com $\mathbf{a} = (a_1, a_2, \dots, a_k)$, um vetor de componentes reais positivas, e mesmo parâmetro escala $b > 0$, ou seja:

$$z_i | (\mathbf{a}, b) \sim G(a_i, b)$$

e

$$f(z_i | \mathbf{a}, b) = \frac{b^{a_i} z_i^{a_i-1}}{e^{bz_i} \Gamma(a_i)},$$

se $t = z_1 + z_2 + \dots + z_k$, então são válidas as seguintes propriedades:

- i) $t | (\mathbf{a}, b) \sim G(\mathbf{1}'\mathbf{a}, b)$;
- ii) se $\boldsymbol{\theta} = \frac{\mathbf{z}}{t}$, então $\boldsymbol{\theta}|\mathbf{a} \sim D_k(\mathbf{a})$;

- iii) $\frac{\mathbf{z}}{t}$ e t são independentes, fixado $\mathbf{a}$;
- iv) a média e a matriz de covariâncias para $\boldsymbol{\theta}$ são, respectivamente:

$$\boldsymbol{\mu} = E(\boldsymbol{\theta} | \mathbf{a}) = \frac{\mathbf{a}}{t} \text{ e } \Sigma = \frac{1}{t+1} \begin{bmatrix} \mu_1 - \boldsymbol{\mu}'\boldsymbol{\mu} & -\boldsymbol{\mu}'\boldsymbol{\mu} & \dots & -\boldsymbol{\mu}'\boldsymbol{\mu} \\ -\boldsymbol{\mu}'\boldsymbol{\mu} & \mu_2 - \boldsymbol{\mu}'\boldsymbol{\mu} & \dots & -\boldsymbol{\mu}'\boldsymbol{\mu} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{\mu}'\boldsymbol{\mu} & -\boldsymbol{\mu}'\boldsymbol{\mu} & \dots & \mu_k - \boldsymbol{\mu}'\boldsymbol{\mu} \end{bmatrix}$$

onde cada μ_i é uma componente do vetor $\boldsymbol{\mu}$.

A demonstração das propriedades acima pode ser verificada em Pereira & Basu (1982).

Diversos autores já mencionaram que o logaritmo de uma variável Gama é bem aproximado por uma variável Normal, ou seja, se um vetor $\mathbf{z}$ tem distribuição Gama, então existe uma variável y com distribuição Log-normal equivalente a $\mathbf{z}$. Para maiores detalhes, veja Aitchison & Shen (1980).

No trabalho de Rodrigues (2005) é feita uma longa discussão a respeito deste resultado, inclusive graficamente mostrando a qualidade da aproximação entre as distribuições Gama($\mathbf{a}, b$) e Log-normal de acordo com as possíveis variações de $\mathbf{a}$ e b . Ainda no mesmo trabalho estão disponíveis outros resultados importantes, como sobre a partição do vetor $\boldsymbol{\theta}$ que possui distribuição Dirichlet resultar em vetores independentes também com distribuição de Dirichlet ou ainda a definição da Dirichlet de segundo tipo obtida através de uma reparametrização do vetor $\boldsymbol{\theta}$.

Com todos os resultados já apresentados, pode-se trabalhar com a aproximação normal para a reparametrização do vetor $\boldsymbol{\theta}$ com distribuição Dirichlet, conforme segue:

2.3 Uma Conjectura

Seja $\boldsymbol{\theta} | \mathbf{a} \sim D_k(\mathbf{a})$, ao aplicar no vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ a reparametrização

$$\mathbf{w} = (w_1, w_2, \dots, w_{k-1}) = \ln \left(\frac{1}{\theta_k} (\theta_1, \theta_2, \dots, \theta_{k-1}) \right),$$

pode-se dizer que $\mathbf{w}$ tem distribuição aproximadamente normal k-dimensional com vetor de médias dado por:

$$\mu_w = E(\mathbf{w}) = \begin{bmatrix} \Psi(a_1) - \Psi(a_k) \\ \Psi(a_2) - \Psi(a_k) \\ \vdots \\ \Psi(a_{k-1}) - \Psi(a_k) \end{bmatrix},$$

e matriz de covariâncias dada por:

$$\Sigma_w = \begin{bmatrix} \Psi'(a_1) + \Psi'(a_k) & \Psi'(a_k) & \cdots & \Psi'(a_k) \\ \Psi'(a_k) & \Psi'(a_2) + \Psi'(a_k) & \cdots & \Psi'(a_k) \\ \vdots & \vdots & \ddots & \vdots \\ \Psi'(a_k) & \Psi'(a_k) & \cdots & \Psi'(a_{k-1}) + \Psi'(a_k) \end{bmatrix},$$

onde $\Psi(a_k)$ e $\Psi'(a_k)$ são respectivamente as funções digama e trigama avaliadas no ponto a_k e definidas como:

$$\Psi(a_k) = \frac{\partial}{\partial a_k} \ln \Gamma(a_k) = \frac{\Gamma'(a_k)}{\Gamma(a_k)} \text{ e } \Psi'(a_k) = \frac{\partial}{\partial a_k} \Psi(a_k).$$

2.4 Família Conjugada de Distribuições

Segundo Berger e Casella (2001), na metodologia Bayesiana, o parâmetro θ (ou vetor de parâmetros $\boldsymbol{\theta}$) é tido como uma quantidade desconhecida, porém fixa, cuja variação pode ser descrita por uma distribuição de probabilidade, chamada de distribuição a priori. Esta distribuição é subjetiva, baseada apenas no conhecimento do pesquisador e é definida antes que os dados sejam observados.

Após se retirar uma amostra da população, a distribuição a priori pode ser atualizada com a informação observada de modo a se obter a distribuição a posteriori. Esta atualização é feita utilizando-se o fator de Bayes.

Ao denotar a distribuição a priori por $\pi(\theta)$ e a distribuição amostral por $f(\mathbf{x} | \theta)$, a distribuição a posteriori é dada por

$$\pi(\theta | \mathbf{x}) = \frac{f(\mathbf{x} | \theta)\pi(\theta)}{\int f(\mathbf{x} | \theta)\pi(\theta)d\theta},$$

note que a distribuição a posteriori é uma distribuição condicional aos dados observados na amostra. Esta distribuição agora é utilizada para se fazer inferências a respeito do parâmetro θ .

Os problemas que serão estudados neste trabalho estão restritos aos vetores de dados observados $\mathbf{x} = (x_1, x_2, x_3, x_4)$, com $\mathbf{1}'\mathbf{x} = n$, associados ao vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$, com $\mathbf{1}'\boldsymbol{\theta} = 1$. A distribuição condicional de $\mathbf{x}$ dado $\boldsymbol{\theta}$ chama-se Distribuição Multinomial e é dada por

$$P(\mathbf{X} = \mathbf{x} | \boldsymbol{\theta}) = n! \prod_{i=1}^4 \frac{\theta_i^{x_i}}{x_i!},$$

em símbolos, escreve-se $\mathbf{x}|\boldsymbol{\theta} \sim M_4(n; \boldsymbol{\theta})$. Para os problemas em que $\mathbf{x}$ e $\boldsymbol{\theta}$ possuem dimensão 2, a distribuição de $\mathbf{x}|\boldsymbol{\theta}$ chama-se trinomial e, no caso de dimensão 1, a distribuição de $\mathbf{x}|\boldsymbol{\theta}$ coincide com a distribuição binomial.

Pereira e Viana (1982) demonstram que a distribuição Dirichlet está naturalmente conjugada com a distribuição Multinomial, por isso, a distribuição Dirichlet torna-se uma escolha natural como distribuição a priori para os parâmetros aqui estudados.

Desta maneira, se $\boldsymbol{\theta}|\mathbf{a} \sim D_4(\mathbf{a})$, conforme definido em 2.1 e se $\mathbf{x}|\boldsymbol{\theta} \sim M_4(n; \boldsymbol{\theta})$, conforme definido acima, então

$$\boldsymbol{\theta}|\mathbf{x} \sim D_4(\mathbf{a}+\mathbf{x}).$$

Demonstração: pelo fator de Bayes, tem-se

$$f(\boldsymbol{\theta} | \mathbf{x}) = \frac{P(\mathbf{x} | \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int P(\mathbf{x} | \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\left(n! \prod_{i=1}^4 \frac{\theta_i^{x_i}}{x_i!} \right) \left(\Gamma\left(\sum_{i=1}^4 a_i \right) \prod_{i=1}^4 \frac{\theta_i^{a_i-1}}{\Gamma(a_i)} \right)}{\int \left(n! \prod_{i=1}^4 \frac{\theta_i^{x_i}}{x_i!} \right) \left(\Gamma\left(\sum_{i=1}^4 a_i \right) \prod_{i=1}^4 \frac{\theta_i^{a_i-1}}{\Gamma(a_i)} \right) d\boldsymbol{\theta}}$$

$$\begin{aligned}
& \frac{n! \Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^4 x_i! \Gamma(a_i)} \left(\prod_{i=1}^4 \theta_i^{x_i} \theta_i^{a_i-1} \right) \\
&= \frac{\frac{n! \Gamma\left(\sum_{i=1}^k a_i\right)}{\prod_{i=1}^4 x_i! \Gamma(a_i)} \int \left(\prod_{i=1}^4 \theta_i^{x_i} \theta_i^{a_i-1} \right) d\boldsymbol{\theta}}{\int \prod_{i=1}^4 \theta_i^{x_i + a_i - 1} d\boldsymbol{\theta}} = \frac{\prod_{i=1}^4 \theta_i^{x_i + a_i - 1}}{\int \prod_{i=1}^4 \theta_i^{x_i + a_i - 1} d\boldsymbol{\theta}}.
\end{aligned}$$

Pereira e Viana (1982) demonstram que

$$\int \prod_{i=1}^4 \theta_i^{a_i-1} d\boldsymbol{\theta} = \frac{\prod_{i=1}^k \Gamma(a_i)}{\Gamma\left(\sum_{i=1}^k a_i\right)}$$

e, desta forma, segue que $\boldsymbol{\theta}|\mathbf{x} \sim D_4(\mathbf{a}+\mathbf{x})$.

As prioris utilizadas neste trabalho serão de dimensão 4 ou menor com vetor de parâmetros $\mathbf{a} = \mathbf{1}$.

3. Testes Bayesiano e Clássico (Frequêntista)

3.1 O FBST (Full Bayesian Significance Test)

O FBST (*Full Bayesian Significance Test* ou Teste de Significância Genuinamente Bayesiano) foi primeiro apresentado por Pereira & Stern (1999) como um teste Bayesiano coerente e intuitivo. Trata-se de um teste de significância estatístico baseado apenas na distribuição a posteriori, com o objetivo de determinar a *evidência* que os dados carregam a favor de uma hipótese precisa. Este teste pode ser implementado utilizando métodos de otimização numérica e técnicas de integração. Como dito anteriormente, por hipótese precisa entende-se $H: \theta \in \Theta_H$, $\Theta_H \subset \Theta$ e $\dim(\Theta_H) < \dim(\Theta)$.

Neste trabalho será utilizada uma versão generalizada do FBST, descrita por Madruga *et al* (2003) que utiliza uma densidade referência no espaço paramétrico.

A densidade referência é escolhida no espaço das densidades sobre o espaço paramétrico original, onde a densidade a priori é definida. Em geral, a escolha da referência recai sobre a densidade que descreve a menor informação sobre θ . Para este trabalho, a escolha da classe de distribuições a priori está restrita à classe de distribuições de Dirichlet de ordem k . Uma escolha intuitiva e natural para a densidade referência é a própria Dirichlet com vetor de parâmetros formado por 1 em todas as posições, que é, na realidade, a própria densidade Uniforme.

Sejam uma hipótese precisa $H: \theta \in \Theta_H$, $f(\theta)$ e $r(\theta)$ as densidades a posteriori e referência para θ , respectivamente, define-se:

$$\theta^* = \arg \max_{\theta \in \Theta_H} \left\{ \frac{f(\theta)}{r(\theta)} \right\} \quad \text{e} \quad s^* = \max_{\theta \in \Theta_H} \left\{ \frac{f(\theta)}{r(\theta)} \right\} = \left\{ \frac{f(\theta^*)}{r(\theta^*)} \right\}.$$

A função $s(\theta) = \frac{f(\theta)}{r(\theta)}$ é chamada “surpresa relativa”. Define-se também no espaço paramétrico Θ o conjunto de maior surpresa relativa (em inglês, highest

relative surprise set - HRSS), Θ^* de pontos $\theta \in \Theta$ com surpresa relativa $s(\theta)$ maior do que em qualquer ponto de Θ_H , ou seja:

$$\Theta^* = \left\{ \theta \in \Theta \left| \frac{f(\theta)}{r(\theta)} \geq s^* \right. \right\}.$$

Note que o conjunto Θ^* é tangente ao conjunto Θ_H em Θ^* . A evidência contra H, de acordo com os dados amostrais $\mathbf{x}$ é dada pela probabilidade a posteriori do conjunto tangente Θ^* :

$$\overline{ev} = \int_{\Theta^*} f(\theta) d\theta.$$

O valor da evidência a favor de H é o complementar de $\overline{ev}$, ou seja, $e\text{-value} = 1 - \overline{ev}$. O FBST rejeita a hipótese nula quando o $e\text{-value}$ resultar em um valor pequeno.

O cálculo do FBST é feito em duas etapas:

- i) **Otimização numérica:** consiste em encontrar o argumento que maximiza a surpresa relativa sob a hipótese H:

$$\theta^* = \arg \max_{\theta \in \Theta_H} \left\{ \frac{f(\theta)}{r(\theta)} \right\};$$

- ii) **Integração numérica:** consiste em integrar a densidade a posteriori sobre a região tangente:

$$\overline{ev} = \int_{\Theta^*} f(\theta) d\theta.$$

Esta definição da evidência contra H é invariante quanto a uma possível reparametrização de θ .

Voltando aos problemas que serão analisados neste trabalho, ao aplicar em $\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ a reparametrização vista na conjectura 2.3, obtém-se:

$$\mathbf{w} = (w_1, w_2, w_3) = \ln \left[\frac{1}{\theta_{22}} (\theta_{11}, \theta_{12}, \theta_{21}) \right],$$

e, de acordo com a conjectura, o vetor $\mathbf{w}$ tem distribuição aproximadamente normal com matriz de médias $\boldsymbol{\mu}_w$ e de covariâncias $\boldsymbol{\Sigma}_w$, respectivamente, ou seja, $g(\mathbf{w}) \sim N(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$.

Da mesma forma, a densidade referência também será aproximada por uma normal com matriz de médias $\boldsymbol{\mu}_r$ e de covariâncias $\boldsymbol{\Sigma}_r$, respectivamente, ou seja, $q(\mathbf{w}) \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$.

Deste modo, a função “surpresa relativa” passa a ser:

$$\begin{aligned} s(\mathbf{w}) &= \frac{f(\mathbf{w})}{r(\mathbf{w})} = \frac{\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_w|}} \exp \left[-\frac{1}{2} \left[\frac{(\mathbf{w} - \boldsymbol{\mu}_w)^2}{\boldsymbol{\Sigma}_w} \right] \right]}{\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_r|}} \exp \left[-\frac{1}{2} \left[\frac{(\mathbf{w} - \boldsymbol{\mu}_r)^2}{\boldsymbol{\Sigma}_r} \right] \right]} \\ &= |\boldsymbol{\Sigma}_w|^{-1/2} |\boldsymbol{\Sigma}_r|^{1/2} \exp \left\{ \frac{1}{2} [(\mathbf{w} - \boldsymbol{\mu}_r)' \boldsymbol{\Sigma}_r (\mathbf{w} - \boldsymbol{\mu}_r) - (\mathbf{w} - \boldsymbol{\mu}_w)' \boldsymbol{\Sigma}_w (\mathbf{w} - \boldsymbol{\mu}_w)] \right\}. \end{aligned}$$

As particularidades para cada tipo de problema e hipóteses serão apresentadas no próximo capítulo junto da resolução dos testes.

3.2 O TRVG (Teste da Razão de Verossimilhanças Generalizado)

Definição 3.2.1 - Seja $x_1, x_2, \dots, x_n$ uma amostra aleatória da variável aleatória X , com função de densidade ou probabilidade $f(\mathbf{x}|\boldsymbol{\theta})$, com $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. A função de verossimilhança de $\boldsymbol{\theta}$ associada a esta amostra é dada por:

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^n f(x_i | \boldsymbol{\theta}).$$

Definição 3.2.1 - O estimador de máxima verossimilhança de θ , chamado de $\hat{\theta}$, é o valor de θ que maximiza a função de verossimilhança definida acima.

Sejam as hipóteses:

$$H: \theta \in \Theta_H$$

$$A: \theta \in \Theta_A$$

onde $\Theta = \Theta_H \cup \Theta_A$, $\Theta_H \cap \Theta_A = \emptyset$, $\Theta_H \neq \emptyset$, $\Theta_A \neq \emptyset$.

O Teste da Razão de Verossimilhanças Generalizado pode ser definido como o teste que utiliza como estatística a razão $\lambda(\mathbf{x})$ de duas maximizações:

- i) o máximo restrito ao subespaço definido por H;
- ii) o máximo da verossimilhança.

A razão $\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_H} L(\theta; \mathbf{x})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x})}$, ou seja, a razão das verossimilhanças calculadas

em seus máximos, deve variar entre 0 e 1. É intuitivo notar que quando H é verdadeira, espera-se que $\lambda(\mathbf{x})$ esteja “próximo” de 1 e, quando H for falsa, espera-se que $\lambda(\mathbf{x})$ esteja “próximo” de 0.

Wilks (1935, 1938) mostrou que, quando $n \rightarrow \infty$, a distribuição nula (distribuição sob H) de $-2\lambda(\mathbf{x})$ é aproximadamente Qui-Quadrado, com número de graus de liberdade determinado pela diferença entre as dimensões do espaço paramétrico original e do subespaço definido por H.

4. Aplicações

Para o Cálculo do FBST, a etapa de otimização numérica pode ser realizada de duas maneiras:

- i) utilizando o estimador de máxima verossimilhança $\hat{\theta}$ de θ como o máximo da função surpresa relativa;
- ii) embora o uso do EMV de θ seja correto, pode-se calcular o ponto de máximo da função utilizando algum método de otimização próprio do *software* utilizado.

Apesar de os dois cálculos apresentarem resultados idênticos, ambos serão apresentados nos programas do anexo. Quando o ponto de máximo $\hat{\theta}$ for de fácil cálculo, basta substituí-lo na função surpresa relativa para encontrar seu máximo, dessa forma economizando tempo de processamento. Quando o ponto de máximo não for conhecido, o máximo da função surpresa relativa pode ser encontrado com técnicas de otimização próprias do *software*. A obtenção do ponto de máximo pelos dois métodos é apresentada no anexo. Quando o número de pontos amostrais cujas funções devem ser otimizadas é muito grande, como em alguns exemplos mostrados a seguir, o fato de conhecer o ponto de máximo através de $\hat{\theta}$ é muito vantajoso pois ajudará a reduzir sensivelmente o tempo de processamento computacional.

Ainda para o FBST, a etapa de integração numérica é feita através do cálculo da função surpresa relativa em 10 mil amostras apresentando distribuição normal com média e variância definida em cada problema, estes pontos amostrais são aleatorizados pelo *software*. A proporção de pontos que apresentarem valor da função surpresa relativa inferior ao valor obtido no ponto de máximo é o próprio *e-value*.

Todos os cálculos do TRVG produzindo os *p-values* também foram programados e serão apresentados no anexo.

Para elucidar a relação entre *p-value* e *e-value*, em todos os exemplos, os pontos (*p-value*, *e-value*) foram apresentados em forma de gráfico. Será possível notar que existe uma forte correlação entre os mesmos. Um dos objetivos deste trabalho é explicitar esta relação através de funções Beta acumuladas, isto é,

mostrar que a relação entre *p-value* e *e-value* é aproximada por uma função de distribuição Beta:

$$f(x | a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt,$$

com $B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$.

Para a obtenção dos parâmetros *a* e *b* da função Beta acumulada, o primeiro passo foi considerar, para os pontos do gráfico dos *e-values* em função dos *p-values*, uma discretização no eixo dos *p-values*, calculando-se para cada intervalo o valor médio dos *e-values*. Após este cálculo, gerou-se uma *spline* de ordem cúbica para interpolar os valores obtidos na discretização.

Esta *spline* é utilizada como base para obter-se os coeficientes da função Beta acumulada. Para isso foram calculados os valores de referência para a *spline* através de uma nova discretização, maior que a utilizada para sua construção.

Os parâmetros da função Beta acumulada foram variados em um dado intervalo e, com um passo conhecido, foi utilizado o método dos mínimos quadrados para minimizar o erro entre as duas curvas testadas (o erro minimizado é a soma dos quadrados das diferenças entre os valores de referência calculados para as duas curvas consideradas), sendo os valores da função Beta acumulada obtidos na mesma discretização utilizada para o cálculo dos valores de referência da *spline*.

A precisão dos parâmetros calculados para as Betas chegou à ordem de 1.0×10^{-10} .

4.1 Teste para Proporção

O teste para a Proporção é um exemplo padrão que tem por objetivo determinar se a taxa de ocorrência de uma determinada característica em uma população *X* pode ser representada por um valor conhecido *p*. O espaço paramétrico é o intervalo unitário $\Theta = \{0 \leq \theta \leq 1\}$.

Uma amostra de *n* indivíduos é retirada da população. Seja *x_i* o número de indivíduos na amostra que apresentam a característica em estudo, se a proporção

de indivíduos com a característica na população for representada por θ_1 , ($0 \leq \theta_1 \leq 1$), então $X \sim \text{Bin}(n, \theta_1)$.

Como exemplos de aplicação deste teste, podemos citar as pesquisas feitas antes e após as propagandas eleitorais para verificar se a preferência por um determinado candidato aumentou; se o tempo de cura para determinada doença diminuiu após a utilização de um certo medicamento ou se as vendas de um produto de consumo aumentaram após a veiculação de uma propaganda na televisão.

4.1.1 Hipótese Nula

Para este teste, as hipóteses de interesse são:

$$H: \theta_1 = p, 0 \leq p \leq 1$$

$$A: \theta_1 \in P, \text{ onde } P \text{ é um conjunto próprio de } [0, 1]$$

4.1.2 FBST

Considere o vetor de parâmetros $\theta = (\theta_1, \theta_2)$ que, para este teste, pode ser reescrito na forma $\theta = (\theta_1, 1 - \theta_1)$, associado ao vetor de dados observados $\mathbf{x} = (x_1, x_2)$. A priori adotada para θ será $D_2(\mathbf{1})$, $\mathbf{x}|\theta$, como já foi mencionado, possui distribuição Binomial com parâmetros n e θ e, portanto, de acordo com a discussão feita em 2.4, $\theta|\mathbf{x} \sim D_2(\mathbf{1}+\mathbf{x})$. Ao aplicar em θ a reparametrização $w = (w_1) = \ln\left(\frac{\theta_1}{\theta_2}\right)$,

testar H fica equivalente a testar:

$$H: w = \ln\left(\frac{\theta_1}{1-\theta_1}\right) = v$$

Com esta reparametrização e de acordo com a conjectura 2.3, a densidade a posteriori $f(w)$ e a densidade de referência $r(w)$ podem ser aproximadas pela normal com médias μ_w e μ_r , e variâncias σ_w^2 e σ_r^2 , respectivamente, dadas por:

$$\mu_w = E(w) = \Psi(1+x_1) - \Psi(1+x_2) \text{ e } \mu_r = E(r) = 0$$

e

$$\sigma_w^2 = \Psi'(1+x_1) + \Psi'(1+x_2) \text{ e } \sigma_r^2 = 2\Psi'(1).$$

Com as densidades a posteriori e referência definidas, pode-se calcular a surpresa relativa:

$$s(w) = \frac{f(w)}{r(w)} = \frac{\sigma_r}{\sigma_w} \exp \left\{ \frac{1}{2} \left[\left(\frac{w^2}{\sigma_r^2} \right) - \left(\frac{(w - \mu_w)^2}{\sigma_w^2} \right) \right] \right\}.$$

O teste é aplicado conforme descrito em 3.1.

4.1.3 TRVG

Seja $\mathbf{x} = (x_1, x_2)$ o vetor de dados observados na amostra. Sob H , $\theta_1 = p$. Sob A , a estimativa para θ_1 é dada pelo estimador de máxima verossimilhança $\theta_1 = \frac{x_1}{n}$.

Deste modo, as funções de verossimilhança sob H e sob $H \cup A$ são:

$$L(\theta_H; \mathbf{x}) = \binom{n}{x_1} p^{x_1} (1-p)^{n-x_1} \text{ e } L(\theta_{H \cup A}; \mathbf{x}) = \binom{n}{x_1} \left(\frac{x_1}{n} \right)^{x_1} \left(1 - \frac{x_1}{n} \right)^{n-x_1}.$$

De modo que a estatística qui-quadrado da razão de verossimilhanças é dada por:

$$Q^2 = -2 * \ln \lambda(\mathbf{x}) = -2 * \ln \left(\frac{\binom{n}{x_1} p^{x_1} (1-p)^{n-x_1}}{\binom{n}{x_1} \left(\frac{x_1}{n} \right)^{x_1} \left(1 - \frac{x_1}{n} \right)^{n-x_1}} \right).$$

Simplificando $\lambda(\mathbf{x})$, obtém-se:

$$Q^2 = -2 * \left(\left(\frac{np}{x_1} \right)^{x_1} \left(\frac{n(1-p)}{n-x_1} \right)^{n-x_1} \right),$$

e, aplicando o $\ln$, chega-se a:

$$Q^2 = -2 * (x_1 \ln p + x_2 \ln (1 - p) + n \ln n - x_1 \ln x_1 - x_2 \ln x_2).$$

Para o caso geral (sob $H \cup A$), o espaço paramétrico é determinado pela proporção θ_1 sujeita à restrição ($0 \leq \theta_1 \leq 1$), portanto a dimensão é 1. Sob H , θ_1 é fixo, portanto a dimensão é 0. A diferença entre as duas dimensões é $1 - 0 = 1$.

Portanto, para amostras grandes, $Q^2 \sim \chi_1^2$ e $p\text{-value} = P(\chi_1^2 > Q^2)$.

4.1.4 Resultados e Comparação

As tabelas a seguir apresentam alguns resultados do *e-value* e do *p-value* para diferentes valores do vetor $\mathbf{x} = (x_1, x_2)$ com diferentes tamanhos de amostra e valores da proporção p :

Tabela 4.1.4.1 - Aplicação do teste da Proporção em amostras de tamanho $n = 100$ com diferentes valores de p

p = 0,3				p = 0,5				p = 0,9			
x₁	x₂	e-value	p-value	x₁	x₂	e-value	p-value	x₁	x₂	e-value	p-value
15	85	0,002	0,000	16	84	0,000	0,000	35	65	0,000	0,000
18	82	0,008	0,006	65	35	0,003	0,003	79	21	0,000	0,001
40	60	0,032	0,034	37	63	0,009	0,009	83	17	0,026	0,032
23	77	0,124	0,117	61	39	0,029	0,027	84	16	0,052	0,063
37	63	0,124	0,134	41	59	0,068	0,071	95	5	0,076	0,068
35	65	0,276	0,282	58	42	0,111	0,109	85	15	0,106	0,118
26	74	0,374	0,376	56	44	0,232	0,230	86	14	0,201	0,206
33	67	0,514	0,516	48	52	0,684	0,689	87	13	0,344	0,337
28	72	0,650	0,660	49	51	0,833	0,841	89	11	0,796	0,742
30	70	0,991	1,000	50	50	1,000	1,000	90	10	0,933	1,000

Tabela 4.1.4.2 - Aplicação do teste da Proporção em amostras de tamanho $n = 1.000$ com diferentes valores de p

p = 0,3				p = 0,5				p = 0,9			
x ₁	x ₂	e-value	p-value	x ₁	x ₂	e-value	p-value	x ₁	x ₂	e-value	p-value
142	858	0,000	0,000	386	614	0,000	0,000	807	193	0,000	0,000
337	663	0,012	0,012	461	539	0,015	0,014	987	13	0,000	0,000
334	666	0,022	0,020	462	538	0,017	0,016	879	121	0,027	0,032
316	684	0,278	0,272	470	530	0,055	0,058	880	120	0,033	0,040
309	691	0,548	0,536	476	524	0,124	0,129	890	110	0,300	0,299
295	705	0,733	0,730	478	522	0,163	0,164	906	94	0,499	0,523
296	704	0,782	0,782	516	484	0,306	0,312	903	97	0,727	0,751
304	696	0,783	0,783	510	490	0,536	0,527	897	103	0,778	0,753
297	703	0,835	0,836	496	504	0,803	0,800	902	98	0,809	0,833
300	700	0,997	1,000	500	500	1,000	1,000	900	100	0,978	1,000

Tabela 4.1.4.3 - Aplicação do teste da Proporção em amostras de tamanho n = 10.000 com diferentes valores de p

p = 0,3				p = 0,5				p = 0,9			
x ₁	x ₂	e-value	p-value	x ₁	x ₂	e-value	p-value	x ₁	x ₂	e-value	p-value
2.905	7.095	0,038	0,038	4.904	5.096	0,053	0,055	8.906	1.094	0,002	0,002
3.066	6.934	0,157	0,151	5.075	4.925	0,134	0,134	8.959	1.041	0,175	0,174
2.951	7.049	0,287	0,284	5.060	4.940	0,235	0,230	8.964	1.036	0,234	0,233
2.959	7.041	0,370	0,370	5.046	4.954	0,362	0,358	8.967	1.033	0,272	0,274
3.037	6.963	0,421	0,420	4.960	5.040	0,420	0,424	8.970	1.030	0,318	0,319
3.028	6.972	0,540	0,542	5.033	4.967	0,506	0,509	8.975	1.025	0,410	0,406
2.980	7.020	0,660	0,662	5.021	4.979	0,685	0,674	9.017	983	0,564	0,570
3.013	6.987	0,786	0,777	4.983	5.017	0,729	0,734	9.010	990	0,742	0,739
2.991	7.009	0,842	0,844	4.992	5.008	0,869	0,873	9.006	994	0,837	0,841
2.998	7.002	0,966	0,965	4.997	5.003	0,950	0,952	9.002	998	0,942	0,947

Com o intuito de verificar se a relação entre *p-value* e *e-value* não se modifica de acordo com o tamanho da amostra observada e com o valor de p testado, foram realizadas diversas simulações com diferentes tamanhos de amostras n e valores de p, varrendo todo o espaço amostral, ou seja, utilizando todas as combinações possíveis de elementos nas duas posições do vetor $\mathbf{x} = (x_1, x_2)$ de modo a se obter soma $x_1 + x_2 = n$ e, também, de forma que nenhum $x_i < 5$.

Os resultados podem ser observados nos gráficos do *e-value* em função do *p-value* disponibilizados a seguir:

Figura 4.1.4.1 - Relação entre *e-value* e *p-value* para $n = 100$ com diferentes valores de p

Figura 4.1.4.2 - Relação entre *e-value* e *p-value* para $n = 1.000$ com diferentes valores de p

Figura 4.1.4.3 - Relação entre *e-value* e *p-value* para $n = 10.000$ com diferentes valores de p

Em todos os gráficos, pode-se verificar que, independentemente dos valores de n e p , a curva que melhor representa os pontos é sempre igual. A linha vermelha representa a curva da função Beta acumulada com parâmetros $a = 0,9957$ e $b = 0,9956$. Estes valores foram ajustados com base nos pontos amostrais obtidos para $n = 10.000$ e $p = 0,5$, conforme descrito no início do capítulo.

Para a obtenção desta curva, o primeiro passo é a discretização dos pontos (p -value, e -value) para obtenção dos pontos médios necessários para garantir a unicidade no mapeamento dos pares (p -value, e -value) e, dessa forma possibilitar o ajuste da curva *spline*:

Figura 4.1.4.4 Discretização dos pontos (p -value, e -value)

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste e os pontos verdes com rótulo “medias” representam as médias dos e -values no intervalo discretizado dos p -values.

Com base nos pontos médios, a *spline* é ajustada e, após mais alguns passos, a curva da Beta acumulada é obtida:

Figura 4.2.4.4 - Ajuste da Beta acumulada pela *spline*

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos ($p\text{-value}$, $e\text{-value}$) obtidos com os dois métodos de resolução do teste, a linha azul com rótulo “*spline*” representa a curva *spline* ajustada após a discretização dos dados e, finalmente, a linha vermelha com rótulo “beta” representa a curva da Beta acumulada melhor ajustada aos pontos deste teste.

4.2 Teste para Homogeneidade de Proporções

O teste de Homogeneidade tem por objetivo determinar se a taxa de ocorrência de uma determinada característica é a mesma para duas populações distintas (X_1 e X_2). Uma amostra de n indivíduos é retirada da primeira população e uma amostra de m indivíduos é retirada da segunda.

Suponha que x_{11} e x_{21} sejam o número de indivíduos de cada uma das amostras que apresentam a característica em estudo, se a proporção de indivíduos com a característica na população X_1 for representada por θ_{11} , ($0 \leq \theta_{11} \leq 1$) e a proporção na população X_2 for representada por θ_{21} , ($0 \leq \theta_{21} \leq 1$), então pode-se afirmar que $X_1 \sim \text{Bin}(n, \theta_{11})$ e $X_2 \sim \text{Bin}(m, \theta_{21})$, com X_1 e X_2 independentes.

A tabela de freqüências observadas para as duas amostras encontra-se a seguir:

Tabela 4.2.1 - Frequências observadas

	Ocorrência		
	Sim	Não	Total
População X	n_{11}	n_{12}	n
População Y	n_{21}	n_{22}	m
Total	$n_{11} + n_{21}$	$n + m - n_{11} - n_{21}$	$n + m$

Para este caso, vale a relação $n_{12} = n - n_{11}$ e $n_{22} = m - n_{21}$.

Dados os vetores de parâmetros $\theta_{(1)} = (\theta_{11}, \theta_{12})$ e $\theta_{(2)} = (\theta_{21}, \theta_{22})$, sujeitos às restrições $\theta_{12} = 1 - \theta_{11}$ e $\theta_{22} = 1 - \theta_{21}$, a função de verossimilhança para x_{11} e x_{21} é dada pelo produto das Binomiais:

$$L(\theta | \mathbf{x}_1, \mathbf{x}_2) = \binom{n}{x_{11}} \theta_{11}^{x_{11}} (1 - \theta_{11})^{n - x_{11}} \binom{m}{x_{21}} \theta_{21}^{x_{21}} (1 - \theta_{21})^{m - x_{21}}.$$

O espaço paramétrico para este caso é dado por $\Theta = \{0 \leq \theta_{ij} \leq 1 \mid \theta_{11} + \theta_{12} = 1 \wedge \theta_{21} + \theta_{22} = 1\}$.

Como exemplos de aplicação deste teste, podemos citar a comparação de duas populações com relação à incidência de uma determinada doença, comportamento de consumo ou preferência eleitoral.

4.2.1 Hipótese Nula

Para este teste, as hipóteses de interesse são:

H: $\theta_{11} = \theta_{21}$ (as probabilidades de ocorrência da característica são iguais para as duas populações)

A: $\theta_{11} \neq \theta_{21}$ (as probabilidades de ocorrência da característica são diferentes)

4.2.2 FBST

Considere os vetores de parâmetros $\theta_{(1)} = (\theta_{11}, \theta_{12})$ e $\theta_{(2)} = (\theta_{21}, \theta_{22})$ que, para este teste, podem ser reescritos na forma $\theta_{(1)} = (\theta_{11}, 1 - \theta_{11})$ e $\theta_{(2)} = (\theta_{21}, 1 - \theta_{21})$,

associados aos vetores de dados observados $\mathbf{x}_1 = (x_{11}, x_{12})$ e $\mathbf{x}_2 = (x_{21}, x_{22})$. As priors adotadas para $\theta_{(1)}$ e $\theta_{(2)}$ serão $D_2(\mathbf{1})$, $\mathbf{x}_1|\theta$ e $\mathbf{x}_2|\theta$, como já foi mencionado, possuem distribuição Binomial com parâmetros $(n, \theta_{(1)})$ e $(m, \theta_{(2)})$ respectivamente e, portanto, de acordo com a discussão feita em 2.4, $\theta_{(1)}|\mathbf{x}_1 \sim D_2(\mathbf{1}+\mathbf{x}_1)$ e $\theta_{(2)}|\mathbf{x}_2 \sim D_2(\mathbf{1}+\mathbf{x}_2)$.

Ao aplicar em θ a reparametrização

$$\mathbf{w} = (w_1, w_2) = \ln\left(\frac{\theta_{11}}{\theta_{12}}, \frac{\theta_{21}}{\theta_{22}}\right) = \ln\left(\frac{\theta_{11}}{1-\theta_{11}}, \frac{\theta_{21}}{1-\theta_{21}}\right),$$

testar H fica equivalente a testar:

$$H: w_1 = w_2 \rightarrow \ln\left(\frac{\theta_{11}}{1-\theta_{11}}\right) = \ln\left(\frac{\theta_{21}}{1-\theta_{21}}\right)$$

Com esta reparametrização e de acordo com a conjectura 2.3, as densidades a posteriori $f(w_1)$ e $f(w_2)$ e a densidade de referência $r(w)$ podem ser aproximadas pela normal, com médias μ_{w_1} , μ_{w_2} e μ_r , e variâncias $\sigma_{w_1}^2$, $\sigma_{w_2}^2$ e σ_r^2 , respectivamente, dadas por:

$$\mu_{w_1} = \Psi(1+x_{11}) - \Psi(1+x_{12}), \mu_{w_2} = \Psi(1+x_{21}) - \Psi(1+x_{22}), \mu_r = 0$$

e

$$\sigma_{w_1}^2 = \Psi'(1+x_{11}) + \Psi'(1+x_{12}), \sigma_{w_2}^2 = \Psi'(1+x_{21}) + \Psi'(1+x_{22}), \sigma_r^2 = 2\Psi'(1).$$

Com as densidades a posteriori e referência definidas, pode-se calcular a surpresa relativa:

$$\begin{aligned} s(w) &= \frac{f_1(w)}{r(w)} \cdot \frac{f_2(w)}{r(w)} \\ &= \frac{\sigma_r}{\sigma_{w_1}} \exp\left\{\frac{1}{2}\left[\left(\frac{w_1^2}{\sigma_r^2}\right) - \left(\frac{(w_1 - \mu_{w_1})^2}{\sigma_{w_1}^2}\right)\right]\right\} \cdot \frac{\sigma_r}{\sigma_{w_2}} \exp\left\{\frac{1}{2}\left[\left(\frac{w_2^2}{\sigma_r^2}\right) - \left(\frac{(w_2 - \mu_{w_2})^2}{\sigma_{w_2}^2}\right)\right]\right\} \end{aligned}$$

$$= \frac{\sigma_r^2}{\sigma_{w_1} \sigma_{w_2}} \exp \left\{ \frac{1}{2} \left[\left(\frac{w_1^2}{\sigma_r^2} \right) - \left(\frac{(w_1 - \mu_{w_1})^2}{\sigma_{w_1}^2} \right) + \left(\frac{w_2^2}{\sigma_r^2} \right) - \left(\frac{(w_2 - \mu_{w_2})^2}{\sigma_{w_2}^2} \right) \right] \right\}.$$

O teste é aplicado conforme descrito em 3.1.

4.2.3 TRVG

Utilizando a função de verossimilhança definida em 4.2, o estimador de máxima verossimilhança para θ sob H : $\theta_{11} = \theta_{21}$ é dado por $\hat{\theta} = \frac{x_{11} + x_{21}}{n + m}$. Sob A , a estimativa para θ_{11} e θ_{21} é dada pelos respectivos estimadores de máxima verossimilhança $\hat{\theta}_{11} = \frac{x_{11}}{n}$ e $\hat{\theta}_{21} = \frac{x_{21}}{m}$. Deste modo, as funções de verossimilhança sob H e sob $H \cup A$ são:

$$L(\theta_H | \mathbf{x}, \mathbf{y}) = \binom{n}{x_{11}} \left(\frac{x_{11} + x_{21}}{n + m} \right)^{x_{11}} \left(1 - \frac{x_{11} + x_{21}}{n + m} \right)^{n - x_{11}} \binom{m}{x_{21}} \left(\frac{x_{11} + x_{21}}{n + m} \right)^{x_{21}} \left(1 - \frac{x_{11} + x_{21}}{n + m} \right)^{m - x_{21}}$$

e

$$L(\theta_{H \cup A} | \mathbf{x}, \mathbf{y}) = \binom{n}{x_{11}} \left(\frac{x_{11}}{n} \right)^{x_{11}} \left(1 - \frac{x_{11}}{n} \right)^{n - x_{11}} \binom{m}{x_{21}} \left(\frac{x_{21}}{m} \right)^{x_{21}} \left(1 - \frac{x_{21}}{m} \right)^{m - x_{21}}.$$

De modo que a estatística qui-quadrado da razão de verossimilhanças é dada por:

$$Q^2 = -2 * \ln \lambda(x)$$

$$= -2 * \ln \left\{ \frac{\binom{n}{x_{11}} \left(\frac{x_{11} + x_{21}}{n + m} \right)^{x_{11}} \left(1 - \frac{x_{11} + x_{21}}{n + m} \right)^{n - x_{11}} \binom{m}{x_{21}} \left(\frac{x_{11} + x_{21}}{n + m} \right)^{x_{21}} \left(1 - \frac{x_{11} + x_{21}}{n + m} \right)^{m - x_{21}}}{\binom{n}{x_{11}} \left(\frac{x_{11}}{n} \right)^{x_{11}} \left(1 - \frac{x_{11}}{n} \right)^{n - x_{11}} \binom{m}{x_{21}} \left(\frac{x_{21}}{m} \right)^{x_{21}} \left(1 - \frac{x_{21}}{m} \right)^{m - x_{21}}} \right\}.$$

Simplificando $\lambda(x)$, obtém-se:

$$Q^2 = -2 * \ln \left\{ \left[\left(\frac{x_{11} + x_{21}}{n + m} \right) \left(\frac{n}{x_{11}} \right) \right]^{x_{11}} \left[\left(1 - \frac{x_{11} + x_{21}}{n + m} \right) \left(\frac{n}{n - x_{11}} \right) \right]^{n - x_{11}} \left[\left(\frac{x_{11} + x_{21}}{n + m} \right) \left(\frac{m}{x_{21}} \right) \right]^{x_{21}} \right. \\ \left. \left[\left(1 - \frac{x_{11} + x_{21}}{n + m} \right) \left(\frac{m}{m - x_{21}} \right) \right]^{m - x_{21}} \right\},$$

e, aplicando o ln, chega-se a:

$$Q^2 = -2 * \{ (x_{11} + x_{21}) \ln(x_{11} + x_{21}) + (x_{12} + x_{22}) \ln(x_{12} + x_{22}) - (n + m) \ln(n + m) \\ + n \ln n + m \ln m - x_{11} \ln x_{11} - x_{21} \ln x_{21} - x_{12} \ln x_{12} - x_{22} \ln x_{22} \}.$$

Para o caso geral (sob $H \cup A$), o espaço paramétrico é determinado pelas proporções θ_{ij} sujeitas às restrições lineares $\sum_{j=1}^2 \theta_{ij} = 1$, para $i = 1, 2$, portanto a dimensão é $1 + 1 = 2$. Sob H , $\theta_{11} = \theta_{21}$, portanto a dimensão é 1. A diferença entre as duas dimensões é $2 - 1 = 1$.

Portanto, para amostras grandes, $Q^2 \sim \chi_1^2$ e $p\text{-value} = P(\chi_1^2 > Q^2)$.

4.2.4 Resultados e Comparação

As tabelas que seguem apresentam alguns resultados do *e-value* e do *p-value* para diferentes valores do vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ com diferentes tamanhos de amostras n e m :

Tabela 4.2.4.1 - Aplicação do teste de Homogeneidade no caso $n = m$

n = m = 30						n = m = 50					
x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>	x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>
20	10	6	24	0,002	0,000	22	28	9	41	0,018	0,004
12	18	25	5	0,003	0,000	30	20	16	34	0,019	0,005
25	5	15	15	0,023	0,005	18	32	30	20	0,055	0,016
10	20	16	14	0,276	0,117	38	12	45	5	0,157	0,059
19	11	13	17	0,292	0,119	23	27	14	36	0,170	0,061
9	21	13	17	0,551	0,283	9	41	14	36	0,478	0,233
8	22	6	24	0,818	0,541	21	29	26	24	0,609	0,316
10	20	9	21	0,958	0,781	22	28	18	32	0,716	0,414
12	18	13	17	0,965	0,793	15	35	17	33	0,906	0,668
5	25	5	25	0,994	1,000	5	45	5	45	0,988	1,000

n = m = 100					
x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>
54	46	78	22	0,002	0,000
41	59	26	74	0,081	0,024
78	22	65	35	0,121	0,041
51	49	40	60	0,289	0,118
14	86	22	78	0,329	0,140
79	21	71	29	0,422	0,191
80	20	76	24	0,785	0,495
28	72	30	70	0,952	0,755
46	54	44	56	0,958	0,776
5	95	5	95	0,967	1,000

Tabela 4.2.4.2 - Aplicação do teste de Homogeneidade no caso $n < m$

n = 30, m = 60						n = 50, m = 100					
x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>	x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>
15	15	13	47	0,022	0,007	26	24	76	24	0,014	0,003
8	22	6	54	0,117	0,046	16	34	50	50	0,111	0,035
24	6	38	22	0,248	0,099	21	29	58	42	0,179	0,064
14	16	36	24	0,475	0,231	10	40	33	67	0,237	0,090
20	10	46	14	0,585	0,317	39	11	88	12	0,269	0,117
13	17	20	40	0,652	0,356	19	31	49	51	0,438	0,200
17	13	29	31	0,750	0,455	41	9	89	11	0,488	0,244
7	23	11	49	0,857	0,580	21	29	52	48	0,503	0,247
11	19	25	35	0,902	0,647	44	6	86	14	0,923	0,732
23	7	46	14	0,999	1,000	23	27	45	55	0,993	0,908

Tabela 4.2.4.3 - Aplicação do teste de Homogeneidade no caso $n > m$

n = 50, m = 25						n = 80, m = 40					
x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>	x_{11}	x_{12}	x_{21}	x_{22}	<i>e-value</i>	<i>p-value</i>
25	25	20	5	0,041	0,010	48	32	32	8	0,085	0,025
17	33	14	11	0,180	0,069	41	39	27	13	0,229	0,088
23	27	17	8	0,194	0,069	44	36	16	24	0,292	0,120
8	42	8	17	0,270	0,118	56	24	33	7	0,310	0,131
14	36	11	14	0,366	0,170	56	24	32	8	0,487	0,235
18	32	13	12	0,403	0,186	40	40	16	24	0,576	0,299
35	15	14	11	0,475	0,233	57	23	25	15	0,617	0,335
7	43	6	19	0,540	0,290	53	27	24	16	0,798	0,502
13	37	7	18	0,984	0,854	30	50	14	26	0,965	0,788
17	33	8	17	0,984	0,862	35	45	18	22	0,993	0,897

Com o intuito de verificar se a relação entre *p-value* e *e-value* não se modifica de acordo com o tamanho da amostra observada, foram realizadas simulações com diferentes tamanhos de amostras n e m , varrendo todo o espaço amostral, ou seja, utilizando todas as combinações possíveis de elementos nas quatro posições do vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ de modo a se obter soma $x_{11} + x_{12} = n$ e $x_{21} + x_{22} = m$ e, também, de forma que nenhum $x_{ij} < 5$.

Os gráficos do *e-value* em função do *p-value* para diferentes valores de n e m , com $n = m$ encontram-se a seguir:

Figura 4.2.4.1 - Relação entre e -value e p -value para $n = m$

Os gráficos do e -value em função do p -value para diferentes valores de n e m , com $n \neq m$ encontram-se a seguir:

Figura 4.2.4.2 - Relação entre *e-value* e *p-value* para $n \neq m$

Em todos os gráficos, pode-se verificar que, independentemente dos valores de n e m e da relação entre eles (iguais ou diferentes), a curva que melhor representa os pontos é sempre igual. A linha vermelha representa a curva da função Beta acumulada com parâmetros $a = 0,8299$ e $b = 1,9586$. Estes valores foram ajustados com base nos pontos amostrais obtidos para $n = m = 100$, conforme descrito no início do capítulo.

Para a obtenção desta curva, o primeiro passo é a discretização dos pontos (*p-value*, *e-value*) para obtenção dos pontos médios necessários para garantir a unicidade no mapeamento dos pares (*p-value*, *e-value*) e, dessa forma possibilitar o ajuste da curva *spline*:

Obtenção dos pontos medios com discretizacao 0.01 e a beta experimental

Figura 4.2.4.1 Discretização dos pontos (p -value, e -value)

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste e os pontos verdes com rótulo “medias” representam as médias dos e -values no intervalo discretizado dos p -values.

Com base nos pontos médios, a *spline* é ajustada e, após mais alguns passos, a curva da Beta acumulada é obtida:

Obtenção das spline cubica com discretizacao 0.01 e a beta experimental

Figura 4.2.4.2 - Ajuste da Beta acumulada pela *spline*

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (*p-value*, *e-value*) obtidos com os dois métodos de resolução do teste, a linha azul com rótulo “*spline*” representa a curva *spline* ajustada após a discretização dos dados e, finalmente, a linha vermelha com rótulo “beta” representa a curva da Beta acumulada melhor ajustada aos pontos deste teste.

4.3 Teste de Homogeneidade de Marginais (O Problema de McNemar)

Dados dois eventos A e B, cada um com 2 categorias, ao classificar n indivíduos de uma população segundo cada uma das categorias de A e B, obtém-se a tabela de contingência 2X2:

Tabela 4.3.1 - Freqüências observadas

Evento A	Evento B		Total
	Categoria 1	Categoria 2	
Categoria 1	n_{11}	n_{12}	$n_{1.}$
Categoria 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

onde

$$n_{i.} = \sum_{j=1}^2 n_{ij} \text{ e } n_{.j} = \sum_{i=1}^2 n_{ij}.$$

Cada indivíduo é classificado em apenas uma combinação de categorias de A e B, em outras palavras, as combinações são exaustivas e mutuamente exclusivas.

Pode-se dizer que esta população apresenta homogeneidade marginal quanto à distribuição dos dois eventos quando:

$$n_{i.} = n_{.j}, \text{ para } i = j.$$

Considerando o vetor de parâmetros $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ associado a cada uma das caselas da tabela acima, a função de verossimilhança para os dados $\mathbf{x}$ é dada pelo modelo Multinomial com parâmetro $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta} | \mathbf{x}) = \left(\frac{n!}{x_{11}! x_{12}! x_{21}! x_{22}!} \right) \theta_{11}^{x_{11}} \theta_{12}^{x_{12}} \theta_{21}^{x_{21}} \theta_{22}^{x_{22}}$$

O espaço paramétrico para este caso é dado por $\Theta = \{0 \leq \theta_{ij} \leq 1 \mid \theta_{11} + \theta_{12} + \theta_{21} + \theta_{22} = 1\}$.

Este teste pode ser aplicado, por exemplo, para verificar se dois professores de uma mesma matéria são igualmente exigentes na avaliação da mesma turma de alunos.

4.3.1 Hipótese Nula

Para este teste, as hipóteses de interesse são:

$$H: \theta_{21} = \theta_{12}$$

$$A: \theta_{21} \neq \theta_{12}$$

4.3.2 FBST

Considere o vetor de parâmetros $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ associado ao vetor de dados observados $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$. A priori adotada para $\boldsymbol{\theta}$ será $D_4(\mathbf{1})$, $\mathbf{x}|\boldsymbol{\theta}$, como já foi mencionado, possui distribuição Multinomial com parâmetros n e $\boldsymbol{\theta}$ e, portanto, de acordo com a discussão feita em 2.4, $\boldsymbol{\theta}|\mathbf{x} \sim D_4(\mathbf{1}+\mathbf{x})$. Ao aplicar em $\boldsymbol{\theta}$ a reparametrização $\mathbf{w} = (w_1, w_2, w_3) = \ln\left(\frac{1}{\theta_{22}}(\theta_{11}, \theta_{12}, \theta_{21})\right)$, testar H fica equivalente a testar:

$$H: w_2 = w_3 \rightarrow \ln\left(\frac{\theta_{12}}{\theta_{22}}\right) = \ln\left(\frac{\theta_{21}}{\theta_{22}}\right)$$

Com esta reparametrização e de acordo com a conjectura 2.3, a densidade a posteriori $f(\mathbf{w})$ e a densidade de referência $r(\mathbf{w})$ podem ser aproximadas pela normal com matrizes de médias $\boldsymbol{\mu}_w$ e $\boldsymbol{\mu}_r$, e de covariâncias $\boldsymbol{\Sigma}_w$ e $\boldsymbol{\Sigma}_r$, respectivamente, dadas por:

$$\boldsymbol{\mu}_w = E(\mathbf{w}) = \begin{bmatrix} \Psi(1+x_{11}) - \Psi(1+x_{22}) \\ \Psi(1+x_{12}) - \Psi(1+x_{22}) \\ \Psi(1+x_{21}) - \Psi(1+x_{22}) \end{bmatrix} \text{ e } \boldsymbol{\mu}_r = E(\mathbf{r}) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

e

$$\boldsymbol{\Sigma}_w = \begin{bmatrix} \Psi'(1+x_{11}) + \Psi'(1+x_{22}) & \Psi'(1+x_{22}) & \Psi'(1+x_{22}) \\ \Psi'(1+x_{22}) & \Psi'(1+x_{12}) + \Psi'(1+x_{22}) & \Psi'(1+x_{22}) \\ \Psi'(1+x_{22}) & \Psi'(1+x_{22}) & \Psi'(1+x_{21}) + \Psi'(1+x_{22}) \end{bmatrix} \text{ e}$$

$$\boldsymbol{\Sigma}_r = \begin{bmatrix} 2\Psi'(1) & \Psi'(1) & \Psi'(1) \\ \Psi'(1) & 2\Psi'(1) & \Psi'(1) \\ \Psi'(1) & \Psi'(1) & 2\Psi'(1) \end{bmatrix}.$$

Com as densidades a posteriori e referência definidas, pode-se calcular a surpresa relativa:

$$s(\mathbf{w}) = \frac{f(\mathbf{w})}{r(\mathbf{w})} = |\boldsymbol{\Sigma}_w|^{-1/2} |\boldsymbol{\Sigma}_r|^{1/2} \exp \left\{ \frac{1}{2} [\mathbf{w}' \boldsymbol{\Sigma}_r^{-1} \mathbf{w} - (\mathbf{w} - \boldsymbol{\mu}_w)' \boldsymbol{\Sigma}_w^{-1} (\mathbf{w} - \boldsymbol{\mu}_w)] \right\}.$$

O teste é aplicado conforme descrito em 3.1.

4.3.3 TRVG

Utilizando a função de verossimilhança definida em 4.3, o estimador de máxima verossimilhança para θ sob $H: \theta_{12} = \theta_{21}$ é dado por $\hat{\theta} = \frac{x_{12} + x_{21}}{2n}$. Sob A , a estimativa para o vetor de parâmetros $\boldsymbol{\theta}$ é dada pelo estimador de máxima verossimilhança $\hat{\theta}_i = \frac{x_i}{n}$. Deste modo, as funções de verossimilhança sob H e sob $H \cup A$ são:

$$L(\theta_H; \mathbf{x}) = \left(\frac{n!}{x_{11}! x_{12}! x_{21}! x_{22}!} \right) \left(\frac{x_{11}}{n} \right)^{x_{11}} \left(\frac{x_{12} + x_{21}}{2n} \right)^{x_{12} + x_{21}} \left(\frac{x_{22}}{n} \right)^{x_{22}}$$

e

$$L(\theta_{H \cup A}; \mathbf{x}) = \left(\frac{n!}{x_{11}! x_{12}! x_{21}! x_{22}!} \right) \left(\frac{x_{11}}{n} \right)^{x_{11}} \left(\frac{x_{12}}{n} \right)^{x_{12}} \left(\frac{x_{21}}{n} \right)^{x_{21}} \left(\frac{x_{22}}{n} \right)^{x_{22}}.$$

De modo que a estatística qui-quadrado da razão de verossimilhanças é dada por:

$$Q^2 = -2 * \ln \lambda(x)$$

$$= -2 * \ln \left\{ \frac{\left(\frac{n!}{x_{11}! x_{12}! x_{21}! x_{22}!} \right) \left(\frac{x_{11}}{n} \right)^{x_{11}} \left(\frac{x_{12} + x_{21}}{2n} \right)^{x_{12} + x_{21}} \left(\frac{x_{22}}{n} \right)^{x_{22}}}{\left(\frac{n!}{x_{11}! x_{12}! x_{21}! x_{22}!} \right) \left(\frac{x_{11}}{n} \right)^{x_{11}} \left(\frac{x_{12}}{n} \right)^{x_{12}} \left(\frac{x_{21}}{n} \right)^{x_{21}} \left(\frac{x_{22}}{n} \right)^{x_{22}}} \right\}.$$

Simplificando $\lambda(x)$, obtém-se:

$$Q^2 = -2 * \ln \left[\left(\frac{x_{12} + x_{21}}{2} \right)^{x_{12} + x_{21}} \left(\frac{1}{x_{12}^{x_{12}} x_{21}^{x_{21}}} \right) \right]$$

e, aplicando o ln, chega-se a:

$$Q^2 = -2 * [(x_{12} + x_{21}) \ln(x_{12} + x_{21}) - (x_{12} + x_{21}) \ln 2 - x_{12} \ln x_{12} - x_{21} \ln x_{21}].$$

Para o caso geral (sob $H \cup A$), o espaço paramétrico é determinado pelas proporções p_{ij} sujeitas à restrição linear $\sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1$, portanto a dimensão é $2 \times 2 - 1 = 3$. Sob H : $\theta_{12} = \theta_{21}$, a dimensão do espaço é 2. A diferença entre as duas dimensões é $3 - 2 = 1$.

Portanto, para amostras grandes, $Q^2 \sim \chi_1^2$ e $p\text{-value} = P(\chi_1^2 > Q^2)$.

4.3.4 Resultados e Comparação

A tabela a seguir apresenta alguns resultados do *e-value* e do *p-value* para diferentes valores do vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ com diferentes tamanhos de amostra:

Tabela 4.3.4.1 - Aplicação do teste de Homogeneidade de Marginais em amostras de tamanhos diferentes

x_{11}	x_{12}	x_{21}	x_{22}	n	<i>e-value</i>	<i>p-value</i>
6	5	12	7	30	0,381	0,085
6	6	12	6	30	0,554	0,153
8	6	10	6	30	0,779	0,315
10	7	8	5	30	0,994	0,796
9	5	15	21	50	0,158	0,022
6	5	14	25	50	0,208	0,035
6	11	18	15	50	0,627	0,191
5	5	7	33	50	0,942	0,563
10	10	30	50	100	0,019	0,001
8	27	35	30	100	0,792	0,309
10	17	22	51	100	0,882	0,423
18	35	38	9	100	0,988	0,725

Com o intuito de verificar se a relação entre *p-value* e *e-value* não se modifica de acordo com o tamanho da amostra observada, foram realizadas simulações com diferentes tamanhos de amostra n , varrendo todo o espaço amostral, ou seja, utilizando todas as combinações possíveis de elementos nas quatro posições do vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ de modo a se obter soma $x_{11} + x_{12} + x_{21} + x_{22} = n$ e, também, de forma que nenhum $x_{ij} < 5$.

Os resultados podem ser observados nos gráficos do *e-value* em função do *p-value* disponibilizados a seguir:

Figura 4.3.4.1 - Relação entre *e-value* e *p-value* para $n = 30$

Figura 4.3.4.2 - Relação entre *e-value* e *p-value* para $n = 50$

Figura 4.3.4.3 - Relação entre *e-value* e *p-value* para $n = 100$

Em todos os gráficos, pode-se verificar que, independentemente dos valores de n , a curva que melhor representa os pontos é sempre igual. A linha vermelha representa a curva da função Beta acumulada com parâmetros $a = 0,6871$ e $b = 3,0189$. Estes valores foram ajustados com base nos pontos amostrais obtidos para $n = 100$, conforme descrito no início do capítulo.

Para a obtenção desta curva, o primeiro passo é a discretização dos pontos (*p-value*, *e-value*) para obtenção dos pontos médios necessários para garantir a unicidade no mapeamento dos pares (*p-value*, *e-value*) e, dessa forma possibilitar o ajuste da curva *spline*:

Figura 4.3.4.4 Discretização dos pontos (p -value, e -value)

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste e os pontos verdes com rótulo “medias” representam as médias dos e -values no intervalo discretizado dos p -values.

Com base nos pontos médios, a *spline* é ajustada e, após mais alguns passos, a curva da Beta acumulada é obtida:

Figura 4.3.4.5 - Ajuste da Beta acumulada pela *spline*

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste, a linha azul com rótulo “*spline*” representa a curva *spline* ajustada após a discretização dos dados e, finalmente, a linha vermelha com rótulo “beta” representa a curva da Beta acumulada melhor ajustada aos pontos deste teste.

4.4 Teste do Equilíbrio Populacional de Hardy-Weinberg

Considere que em uma população uma característica atribuída a um determinado par de genes apresente 3 genótipos: AA, Aa ou aa. As proporções de cada um dos genótipos na população são representadas no vetor $\theta = (\theta_1, \theta_2, \theta_3)$, com $\theta_i > 0$ e sujeitas à restrição linear $\sum_{i=1}^3 \theta_i = 1$. A Lei do Equilíbrio de Hardy-Weinberg estabelece que esta população está em equilíbrio gênico se a proporção de cada um dos genótipos puder ser escrita sob a forma

$$\theta_1 = \theta^2, \theta_2 = 2\theta(1 - \theta) \text{ e } \theta_3 = (1 - \theta)^2, \text{ para algum } 1 \leq \theta \leq 0.$$

A fim de se testar a hipótese de equilíbrio, observa-se uma amostra de n indivíduos desta população. O vetor $\mathbf{x} = (x_1, x_2, x_3)$ representa as frequências observadas de indivíduos classificados sob cada um dos genótipos, e o vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ representa a probabilidade de ocorrência de cada genótipo. A função de verossimilhança para os dados $\mathbf{x}$ é dada pelo modelo Trinomial com parâmetro $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta} | \mathbf{x}) = \left(\frac{n!}{x_1! x_2! x_3!} \right) \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3}.$$

O espaço paramétrico para este caso é dado por $\Theta = \{0 \leq \theta_i \leq 1 \mid \theta_1 + \theta_2 + \theta_3 = 1\}$.

Como exemplos de aplicação deste teste, podemos citar a comparação de duas populações com relação à incidência de uma determinada doença, comportamento de consumo ou preferência eleitoral.

4.4.1 Hipótese Nula

Para este teste, as hipóteses de interesse são:

H: $\exists p \in [0, 1] \mid p_1 = p^2, p_2 = 2p(1-p) \text{ e } p_3 = (1-p)^2$ (a população está em equilíbrio gênico)

A: as 3 proporções acima não se aplicam simultaneamente (a população não está em equilíbrio gênico)

4.4.2 FBST

Considere o vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ que, para este teste, pode ser reescrito na forma $\boldsymbol{\theta} = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)$, associado ao vetor de dados observados $\mathbf{x} = (x_1, x_2, x_3)$. A priori adotada para $\boldsymbol{\theta}$ será $D_3(\mathbf{1})$, $\mathbf{x} | \boldsymbol{\theta}$, como já foi mencionado, possui distribuição Trinomial com parâmetros n e $\boldsymbol{\theta}$ e, portanto, de

acordo com a discussão feita em 2.4, $\theta|\mathbf{x} \sim D_3(\mathbf{1}+\mathbf{x})$. Ao aplicar em θ a reparametrização $\mathbf{w} = (w_1, w_2) = \ln\left(\frac{1}{\theta_3}(\theta_1, \theta_2)\right)$, testar H fica equivalente a testar:

$$H: \mathbf{w} = (w_1, w_2) = \ln\left(\frac{\theta^2}{(1-\theta)^2}, \frac{2\theta(1-\theta)}{(1-\theta)^2}\right)$$

Com esta reparametrização e de acordo com a conjectura 2.3, a densidade a posteriori $f(\mathbf{w})$ e a densidade de referência $r(\mathbf{w})$ podem ser aproximadas pela normal com matrizes de médias $\boldsymbol{\mu}_w$ e $\boldsymbol{\mu}_r$, e de covariâncias $\boldsymbol{\Sigma}_w$ e $\boldsymbol{\Sigma}_r$, respectivamente, dadas por:

$$\boldsymbol{\mu}_w = E(\mathbf{w}) = \begin{bmatrix} \Psi(1+x_1) - \Psi(1+x_3) \\ \Psi(1+x_2) - \Psi(1+x_3) \end{bmatrix} \text{ e } \boldsymbol{\mu}_r = E(\mathbf{r}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

e

$$\boldsymbol{\Sigma}_w = \begin{bmatrix} \Psi'(1+x_1) + \Psi'(1+x_3) & \Psi'(1+x_3) \\ \Psi'(1+x_3) & \Psi'(1+x_2) + \Psi'(1+x_3) \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_r = \begin{bmatrix} 2\Psi'(1) & \Psi'(1) \\ \Psi'(1) & 2\Psi'(1) \end{bmatrix}.$$

Com as densidades a posteriori e referência definidas, pode-se calcular a surpresa relativa:

$$s(\mathbf{w}) = \frac{f(\mathbf{w})}{r(\mathbf{w})} = |\boldsymbol{\Sigma}_w|^{-1/2} |\boldsymbol{\Sigma}_r|^{1/2} \exp\left\{\frac{1}{2} [\mathbf{w}' \boldsymbol{\Sigma}_r^{-1} \mathbf{w} - (\mathbf{w} - \boldsymbol{\mu}_w)' \boldsymbol{\Sigma}_w^{-1} (\mathbf{w} - \boldsymbol{\mu}_w)]\right\}.$$

O teste é aplicado conforme descrito em 3.1.

4.4.3 TRVG

Utilizando a função de verossimilhança definida em 4.4, o estimador de máxima verossimilhança para θ sob H é dado por $\hat{\theta} = \frac{2x_1 + x_2}{2n}$. Sob A, a estimativa

para o vetor de parâmetros θ é dada pelo estimador de máxima verossimilhança

$\hat{\theta}_i = \frac{x_i}{n}$. Deste modo, as funções de verossimilhança sob H e sob $H \cup A$ são:

$$L(\theta_H; \mathbf{x}) = \left(\frac{n!}{x_1! x_2! x_3!} \right) \left(\frac{2x_1 + x_2}{2n} \right)^{2x_1} \left[2 \left(\frac{2x_1 + x_2}{2n} \right) \left(1 - \frac{2x_1 + x_2}{2n} \right) \right]^{x_2} \left(\frac{2x_1 + x_2}{2n} \right)^{2x_3}$$

e

$$L(\theta_{H \cup A}; \mathbf{x}) = \left(\frac{n!}{x_1! x_2! x_3!} \right) \left(\frac{x_1}{n} \right)^{x_1} \left(\frac{x_2}{n} \right)^{x_2} \left(\frac{x_3}{n} \right)^{x_3}.$$

De modo que a estatística qui-quadrado da razão de verossimilhanças é dada por:

$$Q^2 = -2 * \ln \lambda(x) = -2 * \ln \left\{ \frac{\left(\frac{n!}{x_1! x_2! x_3!} \right) \left(\frac{2x_1 + x_2}{2n} \right)^{2x_1} \left[2 \left(\frac{2x_1 + x_2}{2n} \right) \left(1 - \frac{2x_1 + x_2}{2n} \right) \right]^{x_2} \left(\frac{2x_1 + x_2}{2n} \right)^{2x_3}}{\left(\frac{n!}{x_1! x_2! x_3!} \right) \left(\frac{x_1}{n} \right)^{x_1} \left(\frac{x_2}{n} \right)^{x_2} \left(\frac{x_3}{n} \right)^{x_3}} \right\}.$$

Simplificando $\lambda(x)$, obtém-se:

$$Q^2 = -2 * \ln \left\{ 2^{-2x_1 - 2x_3 - x_2} \left(\frac{2x_1 + x_2}{n} \right)^{2x_1} \left(\frac{2n - 2x_1 - x_2}{n} \right)^{2x_3} \left(\frac{(2x_1 + x_2)(2n - 2x_1 - x_2)}{n^2} \right)^{x_2} \left(\frac{x_1}{n} \right)^{-x_1} \left(\frac{x_2}{n} \right)^{-x_2} \left(\frac{x_3}{n} \right)^{-x_3} \right\}$$

e, aplicando o ln, chega-se a:

$$Q^2 = -2 * \{ (-2n + x_2) \ln 2 + (2x_1 + x_2) \ln(2x_1 + x_2) + (x_2 + 2x_3) \ln(x_2 + 2x_3) - n \ln n - x_1 \ln x_1 - x_2 \ln x_2 - x_3 \ln x_3 \}.$$

Para o caso geral (sob $H \cup A$), o espaço paramétrico é determinado pelas proporções θ_i sujeitas à restrição linear $\sum_{i=1}^3 \theta_i = 1$, portanto a dimensão é $3 - 1 = 2$.

Sob H , é determinado por θ e $1 - \theta$, portanto a dimensão é $2 - 1 = 1$. A diferença entre as duas dimensões é $2 - 1 = 1$.

Portanto, para amostras grandes, $Q^2 \sim \chi_1^2$ e $p\text{-value} = P(\chi_1^2 > Q^2)$.

4.4.4 Resultados e Comparação

A tabela a seguir apresenta alguns resultados do *e-value* e do *p-value* para diferentes valores do vetor $\mathbf{x} = (x_1, x_2, x_3)$ com diferentes tamanhos de amostra:

Tabela 4.4.4.1 - Aplicação do teste do Equilíbrio de Hardy-Weinberg em amostras de tamanhos diferentes

n = 30					n = 50					n = 100				
x_1	x_2	x_3	<i>e-value</i>	<i>p-value</i>	x_1	x_2	x_3	<i>e-value</i>	<i>p-value</i>	x_1	x_2	x_3	<i>e-value</i>	<i>p-value</i>
6	6	18	0,021	0,005	13	10	27	0,000	0,000	28	25	47	0,000	0,000
15	9	6	0,182	0,064	14	15	21	0,023	0,006	5	57	38	0,015	0,004
5	19	6	0,314	0,139	5	32	13	0,077	0,024	17	33	50	0,038	0,010
7	11	12	0,401	0,178	12	32	6	0,098	0,033	11	57	32	0,148	0,052
5	11	14	0,584	0,291	6	15	29	0,263	0,100	20	59	21	0,192	0,071
7	12	11	0,584	0,309	16	28	6	0,467	0,235	5	46	49	0,306	0,147
8	17	5	0,704	0,426	18	21	11	0,600	0,311	34	54	12	0,383	0,174
11	13	6	0,839	0,552	6	24	20	0,938	0,768	33	44	23	0,532	0,266
12	13	5	0,907	0,648	9	25	16	0,988	0,888	38	50	12	0,738	0,467
5	15	10	0,978	0,876	5	21	24	0,997	0,898	29	48	23	0,936	0,715

Com o intuito de verificar se a relação entre *p-value* e *e-value* não se modifica de acordo com o tamanho da amostra observada, foram realizadas simulações com diferentes tamanhos de amostra n , varrendo todo o espaço amostral, ou seja, utilizando todas as combinações possíveis de elementos nas três posições do vetor $\mathbf{x} = (x_1, x_2, x_3)$ de modo a se obter soma $x_1 + x_2 + x_3 = n$ e, também, de forma que nenhum $x_i < 5$.

Os resultados podem ser observados nos gráficos do *e-value* em função do *p-value* disponibilizados a seguir:

Figura 4.4.4.1 - Relação entre *e-value* e *p-value* para $n = 30$

Figura 4.4.4.2 - Relação entre *e-value* e *p-value* para $n = 50$

Figura 4.4.4.3 - Relação entre *e-value* e *p-value* para $n = 100$

Em todos os gráficos, pode-se verificar que, independentemente do valor de n , a curva que melhor representa os pontos é sempre igual. A linha vermelha representa a curva da função Beta acumulada com parâmetros $a = 0,8278$ e $b = 1,9751$. Estes valores foram ajustados com base nos pontos amostrais obtidos para $n = 100$, conforme descrito no início do capítulo.

Para a obtenção desta curva, o primeiro passo é a discretização dos pontos (*p-value*, *e-value*) para obtenção dos pontos médios necessários para garantir a unicidade no mapeamento dos pares (*p-value*, *e-value*) e, dessa forma possibilitar o ajuste da curva *spline*:

Figura 4.4.4.4 Discretização dos pontos (p -value, e -value)

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste e os pontos verdes com rótulo “medias” representam as médias dos e -values no intervalo discretizado dos p -values.

Com base nos pontos médios, a *spline* é ajustada e, após mais alguns passos, a curva da Beta acumulada é obtida:

Figura 4.4.4.5 - Ajuste da Beta acumulada pela *spline*

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (*p-value*, *e-value*) obtidos com os dois métodos de resolução do teste, a linha azul com rótulo “*spline*” representa a curva *spline* ajustada após a discretização dos dados e, finalmente, a linha vermelha com rótulo “beta” representa a curva da Beta acumulada melhor ajustada aos pontos deste teste.

A tabela a seguir mostra os parâmetros *a* e *b* obtidos através do modelo de ajuste da Beta acumulada para os testes de Hardy-Weinberg e de Homogeneidade:

Tabela 4.4.4.1 – Comparação entre os parâmetros *a* e *b*

Teste	<i>a</i>	<i>b</i>
Hardy-Weinberg	0,8278	1,9751
Homogeneidade	0,8299	1,9586

Estes resultados mostram que as duas curvas são extremamente próximas. A diferença na curva causada por estas variações entre os parâmetros é muito sutil e, também, pode-se atribuir esta diferença ao fato de, no caso do FBST, ser feita uma aleatorização para se calcular o *e-value* mas, como já foi dito, esta diferença é muito sutil e tende a convergir caso os testes sejam repetidos um número grande de

vezes. Para estes dois testes, a dimensão do espaço paramétrico original é dois, enquanto que sob a hipótese H a dimensão é um.

4.5 Teste de Independência

Dependência e associação são dois conceitos intimamente ligados. Dizer que dois eventos são associados significa que um influencia a ocorrência do outro, ou seja, a ocorrência de um deles pode aumentar ou diminuir a chance do outro ocorrer e, assim, a associação (ou dependência) pode ser chamada de positiva ou negativa.

Dois eventos são independentes quando, ao saber que um deles ocorreu, a probabilidade de o outro ocorrer não se altera.

Dados dois eventos A e B , cada um com 2 categorias, ao classificar n indivíduos de uma população segundo cada uma das categorias de A e B , obtém-se a tabela de contingência 2×2 :

Tabela 4.5.1 - Freqüências observadas

		Evento B	
Evento A	Categoria 1	Categoria 2	Total
Categoria 1	n_{11}	n_{12}	$n_{1.}$
Categoria 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

onde

$$n_{i.} = \sum_{j=1}^2 n_{ij} \text{ e } n_{.j} = \sum_{i=1}^2 n_{ij}.$$

Cada indivíduo é classificado em apenas uma combinação de categorias de A e B , em outras palavras, as combinações são exaustivas e mutuamente exclusivas.

Se a e b são não associadas ou independentes, então:

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}}, i = 1, 2$$

ou, ainda,

$$\frac{n_{ij}}{n_j} = \frac{n_{i.}}{n}, i = 1, 2$$

de onde se deduz que:

$$n_{ij} = \frac{n_{i.} n_{.j}}{n}.$$

Ao dividir as freqüências observadas em cada combinação das categorias de a e b pelo tamanho da amostra, obtém-se a matriz de proporções observadas:

Tabela 4.5.2 - Proporções observadas

		Evento B	
Evento A	Categoria 1	Categoria 2	Total
Categoria 1	θ_{11}	θ_{12}	$\theta_{1.}$
Categoria 2	θ_{21}	θ_{22}	$\theta_{2.}$
Total	$\theta_{.1}$	$\theta_{.2}$	1

onde

$$\theta_{i.} = \sum_{j=1}^2 \theta_{ij} \text{ e } \theta_{.j} = \sum_{i=1}^2 \theta_{ij}.$$

e, para a e b independentes:

$$\theta_{ij} = \theta_{i.} \theta_{.j}$$

A fim de se testar a hipótese de independência, observa-se uma amostra de n indivíduos da população. O vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ representa a freqüência observada de indivíduos classificados na i -ésima categoria do evento A e na j -ésima categoria do evento B, e o vetor de parâmetros $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ representa a probabilidade de ocorrência de cada uma das caselas. A função de verossimilhança para os dados $\mathbf{x}$ é dada pelo modelo Multinomial com parâmetro $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta} | \mathbf{x}) = \left(\frac{n!}{x_{11}! x_{12}! x_{21}! x_{22}!} \right) \theta_{11}^{x_{11}} \theta_{12}^{x_{12}} \theta_{21}^{x_{21}} \theta_{22}^{x_{22}}$$

O espaço paramétrico para este caso é dado por $\Theta = \{0 \leq \theta_{ij} \leq 1 \mid \theta_{11} + \theta_{12} + \theta_{21} + \theta_{22} = 1\}$.

Como exemplos de aplicação deste teste, podemos citar pesquisas para verificar se o hábito de fumar influencia ou não a ocorrência de determinadas doenças ou se a durabilidade de uma peça automotiva depende do tipo de material utilizado ou mesmo do fabricante.

4.5.1 Hipótese Nula

Para este teste, as hipóteses de interesse são:

$$H: \theta_{ij} = \frac{n_{i.} n_{.j}}{n^2} \text{ (os eventos } a \text{ e } b \text{ são independentes)}$$

$$A: \theta_{ij} \neq \frac{n_{i.} n_{.j}}{n^2} \text{ (os eventos } a \text{ e } b \text{ não são independentes)}$$

4.5.2 FBST

Considerando a tabela 4.5.2, testar a hipótese H é equivalente a testar:

$$H: \begin{cases} \theta_{11} = \theta_{1.} \theta_{.1} \\ \theta_{12} = \theta_{1.} \theta_{.2} = \theta_{1.} (1 - \theta_{.1}) \\ \theta_{21} = \theta_{2.} \theta_{.1} = (1 - \theta_{1.}) \theta_{.1} \\ \theta_{22} = \theta_{2.} \theta_{.2} = (1 - \theta_{1.}) (1 - \theta_{.1}) \end{cases}$$

portanto, o vetor de parâmetros $\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ associado ao vetor de dados observados $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$, para este teste, pode ser reescrito na forma $\theta = [(\theta_{1.}, \theta_{.1}, \theta_{1.} (1 - \theta_{.1}), (1 - \theta_{1.}) \theta_{.1}, (1 - \theta_{1.}) (1 - \theta_{.1})]$.

A priori adotada para θ será $D_4(\mathbf{1})$, $\mathbf{x}|\theta$, como já foi mencionado, possui distribuição Multinomial com parâmetros n e θ e, portanto, de acordo com a discussão feita em 2.4, $\theta|\mathbf{x} \sim D_4(\mathbf{1}+\mathbf{x})$.

Ao aplicar em θ a reparametrização $\mathbf{w} = (w_1, w_2, w_3) = \ln\left(\frac{1}{\theta_{22}}(\theta_{11}, \theta_{12}, \theta_{21})\right)$, testar H fica equivalente a testar:

$$H: \mathbf{w} = (w_1, w_2, w_3) = \ln\left(\frac{\theta_{1.} \theta_{.1}}{(1-\theta_{1.})(1-\theta_{.1})}, \frac{\theta_{1.}(1-\theta_{.1})}{(1-\theta_{1.})(1-\theta_{.1})}, \frac{(1-\theta_{1.})\theta_{.1}}{(1-\theta_{1.})(1-\theta_{.1})}\right)$$

Com esta reparametrização e de acordo com a conjectura 2.3, a densidade a posteriori $f(\mathbf{w})$ e a densidade de referência $r(\mathbf{w})$ podem ser aproximadas pela normal com matrizes de médias μ_w e μ_r , e de covariâncias Σ_w e Σ_r , respectivamente, dadas por:

$$\mu_w = E(\mathbf{w}) = \begin{bmatrix} \Psi(1+x_{11}) - \Psi(1+x_{22}) \\ \Psi(1+x_{12}) - \Psi(1+x_{22}) \\ \Psi(1+x_{21}) - \Psi(1+x_{22}) \end{bmatrix} \text{ e } \mu_r = E(\mathbf{r}) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

e

$$\Sigma_w = \begin{bmatrix} \Psi'(1+x_{11}) + \Psi'(1+x_{22}) & \Psi'(1+x_{22}) & \Psi'(1+x_{22}) \\ \Psi'(1+x_{22}) & \Psi'(1+x_{12}) + \Psi'(1+x_{22}) & \Psi'(1+x_{22}) \\ \Psi'(1+x_{22}) & \Psi'(1+x_{22}) & \Psi'(1+x_{21}) + \Psi'(1+x_{22}) \end{bmatrix} \text{ e}$$

$$\Sigma_r = \begin{bmatrix} 2\Psi'(1) & \Psi'(1) & \Psi'(1) \\ \Psi'(1) & 2\Psi'(1) & \Psi'(1) \\ \Psi'(1) & \Psi'(1) & 2\Psi'(1) \end{bmatrix}.$$

Com as densidades a posteriori e referência definidas, pode-se calcular a surpresa relativa:

$$s(\mathbf{w}) = \frac{f(\mathbf{w})}{r(\mathbf{w})} = |\Sigma_w|^{-1/2} |\Sigma_r|^{1/2} \exp\left\{\frac{1}{2} [\mathbf{w}' \Sigma_r^{-1} \mathbf{w} - (\mathbf{w} - \mu_w)' \Sigma_w^{-1} (\mathbf{w} - \mu_w)]\right\}.$$

O teste é aplicado conforme descrito em 3.1.

4.5.3 TRVG

Sob H , $\theta_{ij} = \frac{x_{i.} x_{.j}}{n^2}$. Sob A , a estimativa para θ_{ij} é dada pelo estimador de máxima verossimilhança $\hat{\theta}_{ij} = \frac{x_{ij}}{n}$. Deste modo, as funções de verossimilhança sob H e sob $H \cup A$ são respectivamente:

$$L(\boldsymbol{\theta}_H; \mathbf{x}) = \prod_{i=1}^2 \prod_{j=1}^2 \left(\frac{x_{i.} x_{.j}}{n^2} \right)^{x_{ij}} \text{ e } L(\boldsymbol{\theta}_{H \cup A}; \mathbf{x}) = \prod_{i=1}^2 \prod_{j=1}^2 \left(\frac{x_{ij}}{n} \right)^{x_{ij}}.$$

De modo que a estatística qui-quadrado da razão de verossimilhanças é dada por:

$$Q^2 = -2 \ln \lambda(\mathbf{x}) = -2 \ln \left(\frac{\prod_{i=1}^2 \prod_{j=1}^2 \left(\frac{x_{i.} x_{.j}}{n^2} \right)^{x_{ij}}}{\prod_{i=1}^2 \prod_{j=1}^2 \left(\frac{x_{ij}}{n} \right)^{x_{ij}}} \right).$$

Simplificando $\lambda(\mathbf{x})$, obtém-se:

$$Q^2 = -2 \ln \left(\frac{\prod_{i=1}^2 \prod_{j=1}^2 (x_{i.} x_{.j})^{x_{ij}}}{n^n \prod_{i=1}^2 \prod_{j=1}^2 (x_{ij})^{x_{ij}}} \right),$$

e, aplicando o $\ln$, chega-se a:

$$\begin{aligned} Q^2 &= 2 \left(\sum_{i=1}^2 \sum_{j=1}^2 x_{ij} \ln \left(\frac{x_{ij}}{\hat{\mu}_{ij}} \right) \right) \\ &= 2(x_{11} \ln x_{11} + x_{12} \ln x_{12} + x_{21} \ln x_{21} + x_{22} \ln x_{22} \\ &\quad - x_{11} \ln \hat{\mu}_{11} - x_{12} \ln \hat{\mu}_{12} - x_{21} \ln \hat{\mu}_{21} - x_{22} \ln \hat{\mu}_{22}), \end{aligned}$$

onde $\hat{\mu}_{ij} = \frac{x_i \cdot x_j}{n}$.

Para o caso geral (sob $H \cup A$), o espaço paramétrico é determinado pelas proporções θ_{ij} sujeitas à restrição linear $\sum_{i=1}^2 \sum_{j=1}^2 \theta_{ij} = 1$, portanto a dimensão é $2 \times 2 - 1 = 3$. Sob H , θ_{ij} é determinado por $\theta_{i\cdot}$ e $\theta_{\cdot j}$, portanto a dimensão é $(2 - 1) + (2 - 1) = 2$. A diferença entre as duas dimensões é $3 - 2 = 1$.

Portanto, para amostras grandes, $Q^2 \sim \chi_1^2$ e $p\text{-value} = P(\chi_1^2 > Q^2)$.

4.5.4 Resultados e Comparação

A tabela a seguir apresenta alguns resultados do *e-value* e do *p-value* para diferentes valores do vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ com diferentes tamanhos de amostra:

Tabela 4.5.4.1 - Aplicação do teste de Independência em amostras de tamanhos diferentes

x_{11}	x_{12}	x_{21}	x_{22}	n	<i>e-value</i>	<i>p-value</i>
9	5	5	11	30	0,318	0,068
5	13	6	6	30	0,655	0,217
6	8	5	11	30	0,927	0,510
5	8	7	10	30	0,999	0,880
10	7	5	28	50	0,016	0,002
11	24	8	7	50	0,541	0,147
8	17	11	14	50	0,849	0,381
11	16	10	13	50	0,998	0,845
5	34	16	45	100	0,398	0,099
13	30	10	47	100	0,519	0,137
16	66	6	12	100	0,709	0,217
17	22	20	41	100	0,759	0,277

Com o intuito de verificar se a relação entre *p-value* e *e-value* não se modifica de acordo com o tamanho da amostra observada, foram realizadas simulações com diferentes tamanhos de amostra n , varrendo todo o espaço amostral, ou seja, utilizando todas as combinações possíveis de elementos nas quatro posições do

vetor $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ de modo a se obter soma $x_{11} + x_{12} + x_{21} + x_{22} = n$ e, também, de forma que nenhum $x_{ij} < 5$.

Os resultados podem ser observados nos gráficos do *e-value* em função do *p-value* disponibilizados a seguir:

Figura 4.5.4.1 - Relação entre *e-value* e *p-value* para $n = 30$

Figura 4.5.4.2 - Relação entre *e-value* e *p-value* para $n = 50$

Figura 4.5.4.3 - Relação entre *e-value* e *p-value* para $n = 100$

Em todos os gráficos, pode-se verificar que, independentemente do valor de n , a curva que melhor representa os pontos é sempre igual. A linha vermelha representa a curva da função Beta acumulada com parâmetros $a = 0,7092$ e $b = 3,0751$. Estes valores foram ajustados com base nos pontos amostrais obtidos para $n = 100$, conforme descrito no início do capítulo.

Para a obtenção desta curva, o primeiro passo é a discretização dos pontos (*p-value*, *e-value*) para obtenção dos pontos médios necessários para garantir a unicidade no mapeamento dos pares (*p-value*, *e-value*) e, dessa forma possibilitar o ajuste da curva *spline*:

Figura 4.5.4.4 Discretização dos pontos (p -value, e -value)

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste e os pontos verdes com rótulo “medias” representam as médias dos e -values no intervalo discretizado dos p -values.

Com base nos pontos médios, a *spline* é ajustada e, após mais alguns passos, a curva da Beta acumulada é obtida:

Figura 4.5.4.5 - Ajuste da Beta acumulada pela *spline*

Neste gráfico, os pontos amarelos com rótulo “data” representam os pontos (p -value, e -value) obtidos com os dois métodos de resolução do teste, a linha azul com rótulo “spline” representa a curva *spline* ajustada após a discretização dos dados e, finalmente, a linha vermelha com rótulo “beta” representa a curva da Beta acumulada melhor ajustada aos pontos deste teste.

A tabela a seguir mostra os parâmetros a e b obtidos através do modelo de ajuste da Beta acumulada para os testes de Independência e de McNemar:

Tabela 4.5.4.2 – Comparação entre os parâmetros a e b

Teste	a	b
McNemar	0,6871	3,0189
Independência	0,7092	3,0751

Estes resultados mostram que as duas curvas são extremamente próximas. A diferença na curva causada por estas variações entre os parâmetros é muito sutil e, também, pode-se atribuir esta diferença ao fato de, no caso do FBST, ser feita uma aleatorização para se calcular o e -value mas, como já foi dito, esta diferença é muito sutil e tende a convergir caso os testes sejam repetidos um número grande de vezes. Para estes dois testes, a dimensão do espaço paramétrico original é três, enquanto que sob a hipótese H a dimensão é dois.

5. Considerações Finais

Este trabalho teve por objetivo mostrar ao leitor que os dados são realmente a parte relevante para obterem-se conclusões estatísticas adequadas. Independente do paradigma considerado - Bayesiano e Freqüentista, aqui estudados - as conclusões quando coerentes são equivalentes e dependem realmente do que foi observado e do modelo de probabilidades utilizado.

Os problemas escolhidos para a discussão foram problemas envolvendo distribuições multinomiais, a maioria relacionada com a dupla classificação das tabelas de contingência. Este talvez seja um objetivo desprezioso, mas certamente servirá para esclarecer alguns pontos obscuros que surgem em discussões da literatura.

O fato mais importante que foi mostrado é que a relação entre os valores e (Bayesiano) e p (Freqüentista) é quase uma relação um a um. Assim, o dilema de definir-se qual o valor (de e ou de p) que delimita a tomada de decisões (aceita/rejeita) continua sendo um problema que acompanhará o estatístico que insiste em não usar a teoria de decisões para o suporte de suas decisões. Do ponto de vista do FBST, Madruga, Esteves e Wechsler (2001) apresentam uma forma de se obter o ponto de corte do e -value baseado em funções de perdas. Se este é o caso, definido o corte para o e -value, devido à relação obtida neste trabalho, conseguimos o ponto de corte para o p -value. Se, por outro lado, insistirmos em aceitar os pontos de corte tradicionais (1%, 5% ou 10%) para o p -value, os correspondentes para o e -value também ficam claramente definidos.

Com a discussão acima uma pergunta certamente irá aparecer:

Se existe equivalência entre os valores e e p , por que então ser necessário o uso da alternativa da significância Bayesiana?

A resposta imediata, e sem maiores considerações filosóficas, é que a inferência Bayesiana exige o conhecimento apenas da distribuição a posteriori. Esta é uma função obtida pela adoção de uma priori, descrevendo o conhecimento sobre o verdadeiro valor do parâmetro, e da função de verossimilhança, descrevendo a ligação dos dados ao parâmetro. Conhecer o espaço amostral em sua plenitude não é necessário. Isto é, se não conhecemos a regra de parada na obtenção da amostra, a inferência Bayesiana ainda é possível de ser realizada. Nestes casos, a inferência

Freqüentista fica prejudicada e não deve ser usada, pois todas as qualificações de inferência são baseadas no conhecimento completo do espaço amostral. Pode-se assim aceitar a afirmação preferida dos Bayesianos: *Se não posso ter um desempenho superior ao dos Freqüentistas, no mínimo será equivalente.*

6. Referências Bibliográficas

Aitchison, J., Shen, S.M. (1980). **Logistic-normal distributions: some properties and uses**. *Biometrika*, Vol 67: 261-272.

Berger, R.L., Casella, G. (2002). **Statistical Inference**. Australia, Duxbury.

Madrugá, M.R., Esteves, L.G., Wechsler, S. (2001). **On the Bayesianity of Pereira-Stern Tests**. *TEST*, Vol 10:291 - 299.

Madrugá, M.R., Pereira, C.A.B, Stern, J.M. (2003). **Bayesian evidence test for precise hypotheses**. *Journal of Statistical Planning and Inference*, Vol 117: 185 - 198.

Pereira, C.A.B, Basu, D. (1982). **On the Bayesian analysis of categorical data: the problem of nonresponse**. *Journal of Statistical Planning and Inference*, Vol 6(4): 345 - 362.

Pereira, C.A.B, Viana, M.A.G. (1982). **Elementos de inferência Bayesiana**. São Paulo, Associação Brasileira de Estatística.

Pereira, C.A.B, Stern, J.M. (1999). **Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses**. *Entropy Journal*, Vol 1: 69 - 80.

Rodrigues, W.W., Pereira, C.A.B. (2005). **Teste de significância em tabelas de contingência 2x2 usando modelo logístico - normal**. São Paulo, Instituto de Matemática e Estatística da Universidade de São Paulo.

Wilks, S.S. (1935). **The likelihood test of independence in contingency tables**. *Annals of Mathematics Statistics*, 6, 190 - 196.

Wilks, S.S. (1938). **The large-sample distribution of the likelihood ratio for testing composite hypotheses**. *Annals of Mathematics Statistics*, Vol 9: 60 - 62.

A. Anexo - Programação no MatLab

Todos os programas do MatLab descritos a seguir estão disponíveis em CD na contracapa.

Para rodar os programas inclusos no CD é necessário ter instalado juntamente ao MatLab três *toolbox*, listados abaixo:

- *Spline Toolbox*, para a função spline utilizada no programa “tracamedia.m”;
- *Statistics Toolbox*, para as funções mvnrnd (*multivariate normal random*), chi2cdf (*chi square cumulative density function*), normrnd (*random normal for normal distribution*) e betacdf (*beta cumulative distribution function*);
- *Optimization Toolbox*, para a função fmincon de otimização não linear.

Para todos os problemas aqui estudados, são apresentados os programas:

- para o cálculo do *e-value* para uma única amostra, juntamente com a função de otimização;
- para o cálculo do *p-value* para uma única amostra;
- para calcular tanto *e-value* quanto *p-value* para um número fixo de amostras ou mesmo para todo o espaço amostral. Este programa utiliza o programa “tracamedia.m”, comum a todos os testes.

Os comentários não estão acentuados pois o Matlab não aceita acentuações e aparecem sempre após o símbolo % e na cor verde.

O programa “tracamedia.m” é utilizado para ajustar a *spline*, traçar a curva e encontrar os parâmetros da função Beta acumulada:

A.0 O Ajuste da função Beta Acumulada

```
%tracamedia
```

```
dado = [Pv Ev]; %forma uma matriz com os dados de p-value e e-value ja calculados
```

```
dados = sortrows(dado,1); %organiza os dados na ordem crescente com relação ao valor de p-value
```

```
eps = .01; %variavel que controla o tamanho da discretização para o calculo da spline
```

```

n = length(dados); %tamanho da amostra de p-values e e-values
mini = 0; %guarda a posição atual do p-value no calculo das medias
ii = 0; %variavel de controle de fluxo para as medias

while mini <= 1.0
    %obtem os indices de todos os valores de pvalue entre o intervalo considerado
    idx = find(dados(:,1) >= mini & dados(:,1) < mini+eps);
    if idx > 0
        ii = ii + 1;
        %novo dado com o valor medio de p-value e e-value
        novodado(ii,:) = [mean(dados(idx,1)) mean(dados(idx,2))];
    end
    mini = mini + eps %incrementa o passo
end

eps = 0.005
xx=0:eps:1; %vetor que varia de 0 a 1 com o passo da discretizacao adotado

%valor dos pontos obtidos em função de xx usando a spline calculada
yy = spline(novodado(:,1),novodado(:,2),xx);

ad = 0.1; %discretizacao para o parametro a da beta
bd = 0.1; %discretizacao para o parametro b da beta

amin = 0.5; %valor inicial minimo para a
amax = 3; %valor inicial maximo para a
bmin = 0.5; %valor inicial minimo para b
bmax = 5.0;%valor inicial maximo para b

format long

while ad >= 1e-10 %enquanto a discretização for maior 1e-10
    %gera o vetor para a e b segundo a discretização atual
    a = amin:ad:amax;
    b = bmin:bd:bmax;

    %tamanhos do vetor a e b
    na = length(a);
    nb = length(b);

    %calcula-se a soma da diferença ao quadrado entre os valores obtidos
    %com a spline das medias e a nova spline
    for i = 1:na
        for j = 1:nb%para cada valor do vetor a e b gera-se uma beta.
            EV = betacdf(xx,a(i),b(j));%calcula-se novos pontos obtidos em função de xx
            erro(i,j) = sum((yy-EV).^2);
            %calcula-se a soma da diferença ao quadrado entre os valores obtidos
            %com a spline das medias e a nova spline
        end
    end

    %encontra-se o minimo no intervalo
    [minc,ii] = min(erro);
    [minl,jj] = min(minc);

    a(ii(jj))
    b(jj)

    %calcula-se os novos extremos do intervalo
    amin = a(ii(jj))-ad;

```

```

amax = a(ii(jj))+ad;
bmin = b(jj)-bd;
bmax = b(jj)+bd;

%calcula-se a nova discretização dos parametros
ad = ad/10;
bd = bd/10;

end

fprintf('\nO parametro a da funcao beta vale: %.10f',a(ii(jj)))
fprintf('\nO parametro b da funcao beta vale: %.10f',b(jj))

format short

%plota o grafico dos pontos obtidos com as medias juntamente com todos os
%dados usados para calcula-la
figure(2)
plot(Pv,Ev,'y.')
hold
plot(novodado(:,1),novodado(:,2),'g.')
EV = betacdf(xx,a(ii(jj)),b(jj));
texto = ['Obtenção dos pontos medios com discretizacao ' num2str(eps) ' e a beta experimental'];
title(texto);
xlabel('p-value')
ylabel('e-value')
legend('data','medias',0)

%plota o grafico das medias, da melhor spline e dos dados usados para
%calcular a media
figure(3)
plot(Pv,Ev,'y.')
axis([0 1 0 1])
hold
plot(xx,yy,'b')
axis([0 1 0 1])
plot(xx,EV,'r');
axis([0 1 0 1])
texto = ['Obtenção das spline cubica com discretizacao ' num2str(eps) ' e a beta experimental'];
title(texto);
xlabel('p-value')
ylabel('e-value')
legend('data','spline','beta',0)
axis([0 1 0 1])

```

A.1 Teste para Proporção

A.1.1 FBST

Com o programa abaixo, é possível calcular o e-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

$$e_prop([x \ y],H)$$

onde x e y são os dados observados na amostra e H é a proporção a ser testada.

```

function [evalue] = e_prop(obs,H)

num = 10000;
%numero de amostras aleatorias da distribuição normal multivariada usadas para calcular o e-value
priori = [1 1];%define os parametros da priori
sigmaR = 2 * psi(1,priori(1));
v = log(H/(1-H));%matriz de covariancias da normal para a referencia
param = priori + obs;%define os parametros da posteriori
mu = psi(param(1)) - psi(param(2));%vetor de medias da normal para a posteriori
sigma = psi(1,param(1)) + psi(1,param(2));%matriz de covariancias da normal para a posteriori
f = ((v^2/sigmaR) - ((v-mu)^2/sigma));%funcao surpresa relativa apos reparametrizacao

R = normrnd(mu,sigma^.5,1,num);%gera as amostras aleatorias da distribuição normal multivariada

cont = 0;%variavel de contagem da evidencia (proporção) a favor da hipotese
for i=1:num
    if ((R(i)^2/sigmaR) - ((R(i)-mu)^2/sigma)) < f
        %faz a comparação a favor da hipotese, ou seja calcula se o valor da funcao surpresa relativa
        %e menor para o ponto aleatorizado do que no ponto de maximo
        cont = cont + 1;
    end
end
evalue = cont/num;%calcula a proporcao de pontos a favor da hipotese

```

A.1.2 TRVG

Com o programa abaixo, é possível calcular o p-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

p_prop([x y z w])

onde x e y são os dados observados na amostra e H é a proporção a ser testada.

```

function [pvalue] = p_prop(obs,H)

n = sum(obs);%calcula o tamanho da amostra

chi2 = -2*( obs(1)*log(H) + (n-obs(1))*log(1-H) - obs(1)*log(obs(1)/n) - (n-obs(1)) * log(1 - obs(1)/n ));
%calcula a estatistica do TRVG
pvalue = 1-chi2cdf(chi2,1);%%Aproximacao da estatistica pela Qui Quadrado com 1 grau de liberdade

```

A.1.3 O programa para calcular as duas estatísticas para grandes amostras

Este programa utiliza os programas que fazem os cálculos de *e-value* e *p-value* e o “tracamedia.m” e possibilita a repetição dos testes para um número maior de amostras, dando as seguintes opções na tela:

- Escolha do tamanho das amostras a serem observadas (*n*);
- Escolha da proporção *p* a ser testada pela hipótese (*H*);
- Escolha da quantidade de pontos amostrais: fixa (número a escolher) ou todo o espaço amostral;
- Escolha para o cálculo do ponto de máximo usado na função surpresa relativa: o EMV ou o otimizador;
- Após a finalização do teste, é dada a opção de calcular os parâmetros *a* e *b* da função Beta acumulada que melhor se ajusta aos pontos obtidos.

Se não houver opção para se obter os parâmetros da função Beta, os pontos (*p-value*, *e-value*) serão mostrados automaticamente em um gráfico, junto com a curva da função Beta acumulada com os parâmetros obtidos na ocasião do capítulo 4.

%Proporcao

```
clear %limpa as variaveis do espaço de trabalho
close all %fecha todas as janelas de figura q estejam abertas
clc %limpa a janela de comando

%soma eh a variavel q controla o valor da soma dos elementos dos dados observados
fprintf('Digite o valor da soma dos dados observados: \n');
soma = input('Valor deve ser maior ou igual a 30: ');
while isempty(soma) | soma < 30
    soma = input('Valor deve ser maior ou igual a 30: ');
end

fprintf('\nDigite o valor da proporcao: \n');
H = input('Valor deve estar entre 0 e 1: ');
while isempty(H) | H <= 0 | H >= 1
    H = input('Valor deve estar entre 0 e 1: ');
end

fprintf('\nEscolha uma das opções abaixo para gerar os dados observados para a soma escolhida:\n');
fprintf('1 - Gera uma certa quantidade de dados obseavados randomicamente;\n');
fprintf('2 - Gera todos os dados possiveis.\n');
metodo1 = input('(Opção padrao = 1): ');
if isempty(metodo1) | metodo1 < 1 | metodo1 > 2
    metodo1 = 1;
end

if metodo1 == 1
    fprintf('\nDigite o numero de pontos que devem ser randomizados: \n')
    numero_rand = input('(Numero padrao = 50) :');
```



```

        if isempty(numero_rand)
            numero_rand = 50;
        end
    end

    fid = fopen('compara_proporcao.doc','w'); %abre arquivo para gravação
    fprintf(fid,'e-value p-value observacoes\n\n'); %escreve no arquivo aberto
    if metodo1 == 1
        for i = 1:numero_rand
            i
            flag = 0;
            while (flag == 0) %randomiza uma observacao valida
                flag = 1;
                obs(1) = round(rand*soma);
                if(obs(1) <= 5)
                    flag = 0;
                end
                obs(2) = soma - obs(1);
                if(obs(2) <= 5)
                    flag = 0;
                end
            end

            obs

            pvalue(i) = p_prop(obs,H); %calcula o p-value
            evalue(i) = e_prop(obs,H); %calcula o e-value

            fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
            fprintf(fid,'%d\t',obs(:));
            fprintf(fid,'\n');
        end
    end

    if metodo1 == 2

        i = 1; %variavel de controle de fluxo para os valores de e-value e pvalue

        for k = 5:(soma-5)

            obs = [k soma-k]

            pvalue(i) = p_prop(obs,H); %calcula o e-value
            evalue(i) = e_prop(obs,H); %calcula o p-value
            fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
            fprintf(fid,'%d\t',obs(:));
            fprintf(fid,'\n');
            i = i + 1;
        end
    end

    fclose(fid); %fecha o arquivo aberto para gravação

    Ev = evalue(:);
    Pv = pvalue(:);

    %traça o grafico dos pares p-value e e-value junto com a melhor beta
    figure(1)
    clf
    plot(Pv,Ev,'.')

```

```

hold
PV = 0:.001:1;
EV = betacdf(PV,0.0057,0.9956);
plot(PV,EV,'r');
title('Teste de Proporção')
xlabel('p-value')
ylabel('e-value')
fprintf('\nDeseja obter os parametros da beta media?\n')
traca_media = input('(padrao 0), 1 (um) p/ Sim e 0 (zero) p/ Nao ? ');
if isempty(traca_media) | traca_media < 0 | traca_media > 1
traca_media = 0;
end

if traca_media == 1
tracamedia
end

```

A.2 Teste para Homogeneidade de Proporções

A.2.1 FBST

Com o programa abaixo, é possível calcular o e-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

e_homog([x y z w],opcao)

onde x, y, z e w são os dados observados na amostra e 'opcao' pode ser 1 ou 2 dependendo se deseja-se utilizar o estimador de máxima verossimilhança ou o otimizador para calcular o máximo da função surpresa relativa.

```

%e_homog

function [evalue] = e_homog(obs,otm)

cont = 0; %variavel de contagem da evidencia (proporção) a favor da hipotese
num = 10000;
%numero de amostras aleatorias da distribuição normal multivariada usadas para calcular o e-value

soma1 = obs(1)+obs(2);
soma2 = obs(3)+obs(4);

priori = [1 1 1 1]; %parametros da priori
param = priori + obs;%parametros da posteriori
mu1 = psi(param(1)) - psi(param(2));
%vetor de medias da normal para a posteriori da primeira binomial
mu2 = psi(param(3)) - psi(param(4));
%vetor de medias da normal para a posteriori da segunda binomial
sigma1 = psi(1,param(1)) + psi(1,param(2));
%matriz de covariancias da normal para a posteriori da primeira binomial
sigma2 = psi(1,param(3)) + psi(1,param(4));
%matriz de covariancias da normal para a posteriori da segunda binomial

```

```

sigmaR = psi(1,priori(1)) + psi(1,priori(2)); %matriz de covariancias da normal para a referencia

if otm == 2 %utiliza o otimizador do matlab para calcular o maximo da funcao
    save('dados.mat','mu1','mu2','sigma1','sigma2','sigmaR');
    %salva os dados calculados para serem usados nas iteracoes da otimizacao

    [r,fval] = fmincon(@homogfun,.5,[],[],[],[],.01,.99);
    fval = -fval;
else %utiliza o EMV do vetor teta
    prop = (obs(1)+obs(3))/(soma1+soma2);%EMV sob H

    teta = [prop (1-prop) prop (1-prop)];%vetor teta sob H

    p = log(teta(1)/teta(2));%reparametrizacao do vetor teta

    fval = (p^2/sigmaR + p^2/sigmaR - (p - mu1)^2/sigma1 - (p - mu2)^2/sigma2);
    %funcao surpresa relativa apos reparametrizacao
end

R1 = normrnd(mu1,sigma1^.5,num,1);
%gera as amostras aleatorias da distribuicao normal multivariada para a media e sigma 1
R2 = normrnd(mu2,sigma2^.5,num,1);
%gera as amostras aleatorias da distribuicao normal multivariada para a media e sigma 2

for i=1:num %faz a comparacao a favor da evidencia
    if (R1(i)^2/sigmaR + R2(i)^2/sigmaR - (R1(i) - mu1)^2/sigma1 - (R2(i) - mu2)^2/sigma2) < fval
        %faz a comparacao a favor da hipotese, ou seja calcula e verifica se o valor da funcao
        %surpresa relativa e menor para o ponto aleatorizado do que no ponto de maximo
        cont = cont + 1;
    end
end
evalvalue = cont/num;%calcula a proporcao de pontos a favor da hipotese

```

A.2.1.1 O otimizador para o cálculo do ponto de máximo

Esta função é chamada na função que calcula o *e-value* quando opta-se pelo uso do otimizador no lugar do EMV:

```

%função utilizada nas iteracoes da otimizacao

function [f] = homogfun(x)

load dados.mat %carrega os dados salvos

y = log(x/(1-x));%vetor teta reparametrizado
p = [y y];
f = -(p(1)^2/sigmaR + p(2)^2/sigmaR - (p(1) - mu1)^2/sigma1 - (p(2) - mu2)^2/sigma2);
%funcao surpresa relativa

```

A.2.2 TRVG

Com o programa abaixo, é possível calcular o p-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

`p_homog([x y z w])`

onde x, y, z e w são os dados observados na amostra.

```
%p-homog
```

```
function [pvalue] = p_homog(obs)
```

```
n_1 = obs(1)+obs(2);%tamanho da amostra retirada da primeira populacao
```

```
n_2 = obs(3)+obs(4);%tamanho da amostra retirada da segunda populacao
```

```
n1 = obs(1);
```

```
n3 = obs(3);
```

```
teta = (n1 + n3)/(n_1 + n_2);%EMV sobH
```

```
lambda = ( teta^(n1+n3) * (1-teta)^((n_1 + n_2)-n1-n3) ) / ((n1/n_1)^n1 * (1-n1/n_1)^(n_1-n1) *  
(n3/n_2)^n3 * (1-n3/n_2)^(n_2-n3));% Razao de verossimilhanças
```

```
chi2 = - 2 * log(lambda);%estatística do TRVG
```

```
pvalue = 1 - chi2cdf(chi2,1);%Aproximacao da estatística pela Qui Quadrado com 1 grau de liberdade
```

A.2.3 O programa para calcular as duas estatísticas para grandes amostras

Este programa utiliza os programas que fazem os cálculos de *e-value* e *p-value* e o “tracamedia.m” e possibilita a repetição dos testes para um número maior de amostras, dando as seguintes opções na tela:

- Escolha do tamanho das amostras a serem observadas para as duas populações (n e m);
- Escolha da quantidade de pontos amostrais: fixa (número a escolher) ou todo o espaço amostral;
- Escolha para o cálculo do ponto de máximo usado na função surpresa relativa: o EMV ou o otimizador;
- Após a finalização do teste, é dada a opção de calcular os parâmetros *a* e *b* da função Beta acumulada que melhor se ajusta aos pontos obtidos.

Se não houver opção para se obter os parâmetros da função Beta, os pontos (*p-value*, *e-value*) serão mostrados automaticamente em um gráfico, junto com a curva da função Beta acumulada com os parâmetros obtidos na ocasião do capítulo 4.

```
% comparacao homogeneidade
```

```
clear %limpa as variaveis do espaço de trabalho
```

```
close all %fecha todas as janelas de figura q estejam abertas
```

```
clc %limpa a janela de comando
```

```

%soma eh a variavel q controla o valor da soma dos elementos dos dados observados
fprintf('Digite o tamanho da amostra da primeira população, n: \n');
soma1 = input('Valor deve ser maior ou igual a 15: ');

while isempty(soma1) | soma1 < 15
    soma1 = input('Valor deve ser maior ou igual a 15: ');
    %tamanho da amostra da populacao X1
end

fprintf('\nDigite o tamanho da amostra da segunda população, m: \n');
soma2 = input('Valor deve ser maior ou igual a 15: ');
while isempty(soma2) | soma2 < 15
    soma2 = input('Valor deve ser maior ou igual a 15: ');
    %tamanho da amostra da populacao X2
end

fprintf('\nEscolha uma das opções abaixo para gerar os dados observados para a soma escolhida:\n');
fprintf('1 - Gera uma certa quantidade de dados observados randomicamente;\n');
fprintf('2 - Gera todos os dados possíveis.\n');
metodo1 = input('Opção padrao = 1: ');
if isempty(metodo1) | metodo1 < 1 | metodo1 > 2
    metodo1 = 1;
end

if metodo1 == 1
    fprintf('\nDigite o numero de pontos que devem ser randomizados: \n')
    numero_rand = input('(Numero padrao = 50) :');
    if isempty(numero_rand)
        numero_rand = 50;
    end
end

fprintf('\nEscolha uma das opções abaixo para gerar o valor maximo para a comparação do e-
value:\n');
fprintf('1 - EMV;\n');
fprintf('2 - Otimizador.\n');
metodo2 = input('Opção padrao = 1: ');
if isempty(metodo2) | metodo2 < 1 | metodo2 > 2
    metodo2 = 1;
end

fid = fopen('compara_homogeneidade.doc','w'); %abre arquivo para gravação
fprintf(fid,'e-value p-value observações\n\n'); %escreve no arquivo aberto

if metodo1 == 1
    for i = 1:numero_rand
        i
        flag = 0;

        while (flag == 0) %randomiza uma observacao valida
            flag = 1;
            x1 = round(rand*soma1);
            if(x1 <= 5)
                flag = 0;
            end
            x2 = soma1-x1;
            if(x2 <= 5)

```

```

        flag = 0;
    end
    x3 = round(rand*soma2);
    if(x3 <= 5)
        flag = 0;
    end
    x4 = soma2-x3;
    if(x4 <= 5)
        flag = 0;
    end
end

obs = [x1 x2 x3 x4] %vetor de dados observados

pvalue(i) = p_homog(obs); %calcula o p-value
evalue(i) = e_homog(obs,metodo2); %calcula o e-value

fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
fprintf(fid,'%d\t',obs(:));
fprintf(fid,'\n');
end
end

if metodo1 == 2

    i = 1; %variavel de controle de fluxo para os valores de e-value e pvalue

    for x1 = 5:(soma1-5)
        for x3 = 5:(soma2-5)

            obs = [x1 soma1-x1 x3 soma2-x3]

            pvalue(i) = p_homog(obs); %calcula o e-value
            evalue(i) = e_homog(obs,metodo2); %calcula o p-value

            fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
            fprintf(fid,'%d\t',obs(:));
            fprintf(fid,'\n');
            i = i + 1;
        end
    end
end

fclose(fid); %fecha o arquivo aberto para gravação

Ev = evalue(:);
Pv = pvalue(:);

%traça o grafico dos pares p-value e e-value junto com a melhor beta
figure(1)
clf
plot(Pv,Ev,'g.')
hold
PV = 0:.001:1;
EV = betacdf(PV,.8299,1.9586);
plot(PV,EV,'r');
texto = ['Teste de Homogeneidade n = ' num2str(soma1) ' m = ' num2str(soma2)];
title(texto)
xlabel('p-value')
ylabel('e-value')

```

```

fprintf('\nDeseja obter os parametros da beta media?\n')
traca_media = input('(padrao 0), 1 (um) p/ Sim e 0 (zero) p/ Nao ? ');
if isempty(traca_media) | traca_media < 0 | traca_media > 1
    traca_media = 0;
end
if traca_media == 1
    tracamedia
end

```

A.3 Teste de Homogeneidade de Marginais (O Problema de McNemar)

A.3.1 FBST

Com o programa abaixo, é possível calcular o e-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

```
e_homog_margin([x y z w],opcao)
```

onde x, y, z e w são os dados observados na amostra e 'opcao' pode ser 1 ou 2 dependendo se deseja-se utilizar o estimador de máxima verossimilhança ou o otimizador para calcular o máximo da função surpresa relativa.

```

function [evalue] = e_homog_margin(obs,otm)

cont = 0; %variavel de contagem da evidencia (proporção) a favor da hipotese
priori = [1 1 1 1]; %parametros da priori
num = 10000;
%numero de amostras aleatorias da distribuição normal multivariada usadas para calcular o e-value
param = priori + obs; %parametros da posteriori
teta = param/sum(param); %calcula as proporcoes do vetor de dados observados
m = [psi(param(1));psi(param(2));psi(param(3))] - psi(param(4));
%vetor de medias da normal para a posteriori
sigma = [psi(1,param(1)) 0 0; 0 psi(1,param(2)) 0; 0 0 psi(1,param(3))] + psi(1,param(4));
%matriz de covariancias da normal para a posteriori
sigmaR = [psi(1,priori(1)) 0 0; 0 psi(1,priori(2)) 0; 0 0 psi(1,priori(3))] + psi(1,priori(4));
%matriz de covariancias da normal para a referencia
invsigma = inv(sigma); %inverte a matriz sigma
invsigmaR = inv(sigmaR); %inverte a matriz sigmaR

if otm == 2 %utiliza o otimizador do matlab para calcular o maximo da função
    save('dados.mat','m','invsigma','invsigmaR','teta');

    if teta(2) == teta(3) %fornece o intervalo de variacao para o otimizador quanto teta12 = teta21
        lb = .01;
        ub = .99;
    else %fornece o intervalo de variacao para o otimizador quanto teta12 <> teta21
        lb = min(teta(2),teta(3));
        ub = max(teta(2),teta(3));
    end
end

```

```

[r,fval] = fmincon(@homogmarginfun,lb+ub/2,[],[],[],[],lb,ub);
%otimizador para encontrar o maximo
else
    n = sum(obs);%tamanho da amostra

    prop1 = (obs(2)+obs(3))/(2*n);%EMV sob H

    teta = [obs(1)/n prop1 prop1 obs(4)/n];%vetor teta sob H

    p = log([teta(1);teta(2);teta(3)]/teta(4));%%reparametrizacao do vetor teta

    fval = -( (transp(p) * invsigmaR * p) - (transp(p - m) * invsigma * (p-m)) );
    %funcao surpresa relativa apos reparametrizacao
end

R = mvnrnd(m,sigma,num);%gera as amostras aleatorias da distribuicao normal multivariada
R = R';%transpoe a matriz gerada

for i=1:num
    if ( ( (transp(R(:,i)) * invsigmaR * R(:,i)) - (transp(R(:,i) - m) * invsigma * (R(:,i)-m)) ) ) < -fval
        %faz a comparacao a favor da hipotese, ou seja calcula se o valor da funcao surpresa relativa
        %e menor para o ponto aleatorizado do que no ponto de maximo
        cont = cont + 1;
    end
end
end
evaluate = cont/num;%calcula a proporcao de pontos a favor da hipotese

```

A.3.1.1 O otimizador para o cálculo do ponto de máximo

Esta função é chamada na função que calcula o *e-value* quando opta-se pelo uso do otimizador no lugar do EMV:

%função utilizada nas iterações da otimização para o teste da Homogeneidade Marginal

```

function [f] = homogmarginfun(x)

load dados.mat;%carrega os dados salvos

p = log( 1/teta(4) * [ teta(1); x; x] );% vetor teta reparametrizado
f = -( (transp(p) * invsigmaR * p) - (transp(p - m) * invsigma * (p-m)) ) ;% funcao surpresa relativa

```

A.3.2 TRVG

Com o programa abaixo, é possível calcular o p-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

p_homog_margin([x y z w])

onde x, y, z e w são os dados observados na amostra.

```
function [pvalue] = p_homog_margin(obs)
```



```
chi2 = -2*log( ((obs(2)+obs(3))/2)^(obs(2)+obs(3)) * inv( (obs(2)^obs(2)) * (obs(3)^obs(3)) ) );
% calculo da Estatística do TRVG
pvalue = 1 - chi2cdf(chi2,1);%Aproximacao da estatistica pela Qui Quadrado com 1 grau de liberdade
```

A.3.3 O programa para calcular as duas estatísticas para grandes amostras

Este programa utiliza os programas que fazem os cálculos de *e-value* e *p-value* e o “tracamedia.m” e possibilita a repetição dos testes para um número maior de amostras, dando as seguintes opções na tela:

- Escolha do tamanho das amostras a serem observadas (n);
- Escolha da quantidade de pontos amostrais: fixa (número a escolher) ou todo o espaço amostral;
- Escolha para o cálculo do ponto de máximo usado na função surpresa relativa: o EMV ou o otimizador;
- Após a finalização do teste, é dada a opção de calcular os parâmetros *a* e *b* da função Beta acumulada que melhor se ajusta aos pontos obtidos.

Se não houver opção para se obter os parâmetros da função Beta, os pontos (*p-value*, *e-value*) serão mostrados automaticamente em um gráfico, junto com a curva da função Beta acumulada com os parâmetros obtidos na ocasião do capítulo 4.

%Homogeneidade Marginal

```
clear %limpa as variaveis do espaço de trabalho
close all %fecha todas as janelas de figura q estejam abertas
clc %limpa a janela de comando

%soma eh a variavel q controla o valor da soma dos elementos dos dados observados
fprintf('Digite o valor da soma dos dados observados: \n');
soma = input('Valor deve ser maior ou igual a 30: ');
while isempty(soma) | soma < 30
    soma = input('Valor deve ser maior ou igual a 30: ');
end

fprintf('\nEscolha uma das opções abaixo para gerar os dados observados para a soma escolhida:\n');
fprintf('1 - Gera uma certa quantidade de dados observados randomicamente;\n');
fprintf('2 - Gera todos os dados possíveis.\n');
metodo1 = input('(Opção padrao = 1): ');
if isempty(metodo1) | metodo1 < 1 | metodo1 > 2
    metodo1 = 1;
end

if metodo1 == 1
    fprintf('\nDigite o numero de ponto que devem ser randomizados: \n')
```

```

    numero_rand = input('(Numero padrao = 50) :');
    if isempty(numero_rand)
        numero_rand = 50;
    end
end

fprintf('\nEscolha uma das opções abaixo para gerar o valor maximo para a comparação do e-
value:\n');
fprintf('1 - Estimador;\n');
fprintf('2 - Otimizador.\n');
metodo2 = input('(Opção padrao = 1): ');
if isempty(metodo2) | metodo2 < 1 | metodo2 > 2
    metodo2 = 1;
end

fid = fopen('compara_homogeneidade_marginal.doc','w'); %abre arquivo para gravação
fprintf(fid,'e-value p-value observações\n\n'); %escreve no arquivo aberto

if metodo1 == 1
    for i = 1:numero_rand
        i
        flag = 0;
        while (flag == 0) %randomiza uma observacao valida
            flag = 1;
            kkkk = round(rand*soma);
            if(kkkk <= 5)
                flag = 0;
            end
            kkk = round(rand*(soma-kkkk));
            if(kkk <= 5)
                flag = 0;
            end
            kk = round(rand*(soma-kkkk-kkk));
            if(kk <= 5)
                flag = 0;
            end
            k = soma - kkkk-kkk-kk;
            if(k <= 5)
                flag = 0;
            end
        end

        obs = [kkkk kkk kk k] %vetor de dados observados

        pvalue(i) = p_homog_margin(obs); %calcula o p-value
        evalue(i) = e_homog_margin(obs,metodo2); %calcula o e-value

        fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
        fprintf(fid,'%d\t',obs(:));
        fprintf(fid,'\n');
    end
end

if metodo1 == 2

    i = 1; %variavel de controle de fluxo para os valores de e-value e pvalue

    for kkkk = 5:floor(soma/2)
        for k = 5:(soma-kkkk)

```

```

for kkk = 5:floor((soma-kkk-k)/2)
    kk = soma - kkk - k - kkk;

    if kk > 4
        obs = [kkkk kkk kk k]
        pvalue(i) = p_homog_margin(obs); %calcula o e-value
        evalue(i) = e_homog_margin(obs,metodo2);
        %calcula o p-value

        fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
        fprintf(fid,'%d\t',obs(:));
        fprintf(fid,'\n');
        i = i + 1;
    end
end
end
end

fclose(fid); %fecha o arquivo aberto para gravação

Ev = evalue(:);
Pv = pvalue(:);

%traça o grafico dos pares p-value e e-value junto com a melhor beta
figure(1)
clf
plot(Pv,Ev,'g.')
hold
PV = 0:.001:1;
EV = betacdf(PV,0.6871,3.0189);
plot(PV,EV,'r');
title('Teste de Homogeneidade Marginal')
xlabel('p-value')
ylabel('e-value')

fprintf('\nDeseja obter os parametros da beta media?\n')
traca_media = input('(padrao 0), 1 (um) p/ Sim e 0 (zero) p/ Nao ? ');
if isempty(traca_media) | traca_media < 0 | traca_media > 1
    traca_media = 0;
end

if traca_media == 1
    tracamedia
end

```

A.4 Teste do Equilíbrio Populacional de Hardy-Weinberg

A.4.1 FBST

Com o programa abaixo, é possível calcular o e-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

```
e_hardy([x y z],opcao)
```

onde x, y e z são os dados observados na amostra e 'opcao' pode ser 1 ou 2 dependendo se deseja-se utilizar o estimador de máxima verossimilhança ou o otimizador para calcular o máximo da função surpresa relativa.

```
%e_hardy

function [evaluate] = e_hardy(obs,otm)

priori = [1 1 1]; %parametros da priori
num = 10000;
%numero de amostras aleatorias da distribuicao normal multivariada usadas para calcular o e-value
cont = 0; %variavel de contagem da evidencia (proporcao) a favor da hipotese
param = priori + obs;%parametros da posteriori
m = [psi(param(1));psi(param(2))] - psi(param(3));%vetor de medias da normal para a posteriori
sigma = [psi(1,param(1)) 0 ; 0 psi(1,param(2))] + psi(1,param(3));
%matriz de covariancias da normal para a posteriori
sigmaR = [psi(1,priori(1)) 0 ; 0 psi(1,priori(2))] + psi(1,priori(3));
%matriz de covariancias da normal para a referencia
invsigma = inv(sigma);%inverte a matriz sigma
invsigmaR = inv(sigmaR);%inverte a matriz sigmaR

if otm == 2 %utiliza o otimizador do matlab para calcular o maximo da funcao

    save('dados.mat','m','invsigma','invsigmaR')
    %salva os dados calculados para serem usados nas iteracoes da otimizacao

    %otimizador
    [r,fval] = fmincon(@hardyfun,0.5,[],[],[],[],.01,.99);
    fval = -fval;
else %utiliza o EMV do vetor teta
    n = sum(obs);

    prop = (2*obs(1)+obs(2))/(2*n); %EMV sob H

    teta1 = [prop^2 2*prop*(1-prop) (1-prop)^2];%vetor teta sob H

    p = log([teta1(1);teta1(2)]/teta1(3));%reparametrizacao do vetor teta

    fval = ( (transp(p) * invsigmaR * p) - (transp(p - m) * invsigma * (p-m)) );
    %funcao surpresa relativa apos reparametrizacao
end

R = mvnrnd(m,sigma,num); %gera as amostras aleatorias da distribuicao normal multivariada
R = R'; %transpoe a matriz gerada

for i=1:num
    if (( transp(R(:,i)) * invsigmaR * R(:,i) ) - ( transp(R(:,i) - m) * invsigma * (R(:,i) - m) )) < fval
        %faz a comparacao a favor da hipotese, ou seja calcula se o valor da funcao surpresa relativa
        %e menor para o ponto aleatorizado do que no ponto de maximo
        cont = cont + 1;
    end
end

evaluate = cont/num;%calcula a proporcao de pontos a favor da hipotese
```

4.1.1 O otimizador para o cálculo do ponto de máximo

Esta função é chamada na função que calcula o *e-value* quando opta-se pelo uso do otimizador no lugar do EMV:

%função utilizada nas iterações da otimização para o teste do Equilíbrio de Hardy-Weinberg

```
function [f] = hardyfun(x);
```

load dados.mat %carrega os dados salvos

p = log([x^2/(1-x)^2 ; 2*x/(1-x)]); % vetor teta reparametrizado

f = -((transp(p) * invsigmaR * p) - (transp(p - m) * invsigma * (p - m))); % funcao surpresa relativa

A.4.2 TRVG

Com o programa, para calcular o *p-value* para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

p_hardy([x y z])

onde x, y e z são os dados observados na amostra.

%p_hardy

```
function [pvalue] = p_hardy(obs);
```

n1 = obs(1); %posicao 1 do vetor de observacoes

n2 = obs(2); %posicao 2 do vetor de observacoes

n3 = obs(3); %posicao 3 do vetor de observacoes

C1 = (2*n1+n2);

n = sum(obs); %tamanho da amostra

teta = C1/(2*n); %EMV sob H

temp = ((n1/n)^n1 * (n2/n)^n2 * (n3/n)^n3);

if temp == 0 %se temp == 0 nao se calcula p-value

pvalue = NaN;

else

lambda = (teta^(2*n1) * (2*teta*(1-teta))^n2 * (1-teta)^(2*n3))/temp;

% Razao de verossimilhanças

chi2 = - 2 * log(lambda); % Estatística do TRVG

pvalue = 1-chi2cdf(chi2,1);

%Aproximacao da estatística pela Qui Quadrado com 1 grau de liberdade

end

A.4.3 O programa para calcular as duas estatísticas para grandes amostras

Este programa utiliza os programas que fazem os cálculos de *e-value* e *p-value* e o “tracamedia.m” e possibilita a repetição dos testes para um número maior de amostras, dando as seguintes opções na tela:

- Escolha do tamanho das amostras a serem observadas (*n*);
- Escolha da quantidade de pontos amostrais: fixa (número a escolher) ou todo o espaço amostral;
- Escolha para o cálculo do ponto de máximo usado na função surpresa relativa: o EMV ou o otimizador;
- Após a finalização do teste, é dada a opção de calcular os parâmetros *a* e *b* da função Beta acumulada que melhor se ajusta aos pontos obtidos.

Se não houver opção para se obter os parâmetros da função Beta, os pontos (*p-value*, *e-value*) serão mostrados automaticamente em um gráfico, junto com a curva da função Beta acumulada com os parâmetros obtidos na ocasião do capítulo 4.

%Hardy-Weinberg

```
clear %limpa as variaveis do espaço de trabalho
close all %fecha todas as janelas de figura q estejam abertas
clc %limpa a janela de comando
```

```
%soma eh a variavel q controla o valor da soma dos elementos dos dados
%observados, e o tamanho da amostra
```

```
fprintf('Digite o valor da soma dos dados observados: \n');
```

```
soma = input('Valor deve ser maior ou igual a 30: ');
```

```
while isempty(soma) | soma < 30
    soma = input('Valor deve ser maior ou igual a 30: ');
```

```
end
```

```
fprintf('\nEscolha uma das opções abaixo para gerar os dados observados para a soma escolhida:\n');
```

```
fprintf('1 - Gera uma certa quantidade de dados obseavados randomicamente;\n');
```

```
fprintf('2 - Gera todos os dados possiveis.\n');
```

```
metodo1 = input('(Opção padrao = 1): ');
```

```
if isempty(metodo1) | metodo1 < 1 | metodo1 > 2
    metodo1 = 1;
```

```
end
```

```
if metodo1 == 1
```

```
    fprintf('\nDigite o numero de pontos que devem ser randomizados: \n')
```

```
    numero_rand = input('(Numero padrao = 50) :');
```

```
    if isempty(numero_rand)
```

```

        numero_rand = 50;
    end
end

fprintf('\nEscolha uma das opções abaixo para gerar o maximo da funcao surpresa relativa para a
obtencao do e-value:\n');
fprintf('1 - EMV;\n');
fprintf('2 - Otimizador.\n');
metodo2 = input('(Opção padrao = 1): ');
if isempty(metodo2) | metodo2 < 1 | metodo2 > 2
    metodo2 = 1;
end

fid = fopen('compara_hardy.doc','w'); %abre arquivo para gravação
fprintf(fid,'e-value p-value observações\n\n'); %escreve no arquivo aberto

if metodo1 == 1
    for i = 1:numero_rand
        i
        flag = 0;

        while (flag == 0) %randomiza uma observacao valida
            flag = 1;
            obs(1) = round(rand*soma);
            if(obs(1) <= 5)
                flag = 0;
            end
            obs(2) = round(rand*(soma-obs(1)));
            if(obs(2) <= 5)
                flag = 0;
            end
            obs(3) = soma - obs(1)-obs(2);
            if(obs(3) <= 5)
                flag = 0;
            end
        end

        obs

        pvalue(i) = p_hardy(obs); %calcula o p-value
        evalue(i) = e_hardy(obs,metodo2); %calcula o e-value
        fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
        fprintf(fid,'%d\t',obs(:));
        fprintf(fid,'\n');
    end
end

if metodo1 == 2

    i = 1; %variavel de controle de fluxo para os valores de e-value e pvalue

    for kkk = 5:floor(soma/2)
        for kk = 5:soma
            k = soma - kkk - kk;

            if k > 4

                obs = [kkk kk k]

                pvalue(i) = p_hardy(obs); %calcula o e-value

```

```

        evalue(i) = e_hardy(obs,metodo2); %calcula o p-value
        fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
        fprintf(fid,'%d\t',obs(:));
        fprintf(fid,'\n');
        i = i + 1;
    end
end
end

fclose(fid); %fecha o arquivo aberto para gravação

Ev = evalue(:);
Pv = pvalue(:);

%traça o grafico dos pares p-value e e-value junto com a melhor beta
figure(1)
clf
plot(Pv,Ev,'g.')
hold
PV = 0:.001:1;
EV = betacdf(PV,.8278,1.9751);
plot(PV,EV,'r');
title('Teste de Hardy-Weinberg')
xlabel('p-value')
ylabel('e-value')

fprintf('\nDeseja obter os parametros da beta media?\n')
traca_media = input('(padrao 0), 1 (um) p/ Sim e 0 (zero) p/ Nao ? ')
if isempty(traca_media) | traca_media < 0 | traca_media > 1
    traca_media = 0;
end

if traca_media == 1
    tracamedia
end

```

A.5 Teste de Independência

A.5.1 FBST

Com o programa abaixo, é possível calcular o e-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

```
e_indep([x y z w],opcao)
```

onde x, y, z e w são os dados observados na amostra e 'opcao' pode ser 1 ou 2 dependendo se deseja-se utilizar o estimador de máxima verossimilhança ou o otimizador para calcular o máximo da função surpresa relativa.


```

%e_indep

function [evaluate] = e_indep(obs,otm)

kkkk = obs(1,1);
kkk = obs(1,2);
kk = obs(2,1);
k = obs(2,2);

obs = [kkkk k k k];%dados observados na amostra

priori = [1 1 1 1];%parametros da priori
num = 10000;
%numero de amostras aleatorias da distribuicao normal multivariada usadas para calcular o e-value

param = priori + obs;%parametros da posteriori
teta = param/sum(param);%calcula as proporcoes do vetor de dados observados
m = [psi(param(1));psi(param(2));psi(param(3))] - psi(param(4));
%vetor de medias da normal para a posteriori
sigma = [psi(1,param(1)) 0 0; 0 psi(1,param(2)) 0; 0 0 psi(1,param(3))] + psi(1,param(4));
%matriz de covariancias da normal para a posteriori
sigmaR = [psi(1,priori(1)) 0 0; 0 psi(1,priori(2)) 0; 0 0 psi(1,priori(3))] + psi(1,priori(4));
%matriz de covariancias da normal para a referencia
invsigma = inv(sigma);%inverte a matriz sigma
invsigmaR = inv(sigmaR);%inverte a matriz sigmaR

if otm == 2%utiliza o otimizador do matlab para calcular o maximo da funcao

    save('dados.mat','m','invsigma','invsigmaR');

    %otimizador
    [r,fval] = fmincon(@indepfun,[0.5 0.5],[],[],[],[],[0.01;0.01],[0.99;0.99]);

else%utiliza o EMV do vetor teta

    n = sum(obs);%tamanho da amostra

    prop1 = (obs(1)+obs(2))/n;

    prop2 = (obs(1)+obs(3))/n;

    teta = [prop1*prop2 prop1*(1-prop2) (1-prop1)*prop2 (1-prop1)*(1-prop2)];%%vetor teta sob H

    p = log([teta(1);teta(2);teta(3)]/teta(4));%reparametrizacao do vetor teta

    fval = -( (transp(p) * invsigmaR * p) - (transp(p - m) * invsigma * (p-m)) );
    %funcao surpresa relativa apos reparametrizacao

end
R = mvnrnd(m,sigma,num);%gera as amostras aleatorias da distribuicao normal multivariada
R = R';%transpoe a matriz gerada

cont = 0;
for i=1:num
    if ( ( (transp(R(:,i)) * inv(sigmaR) * R(:,i)) - (transp(R(:,i) - m) * inv(sigma) * (R(:,i)-m)) ) ) < -fval
        %faz a comparacao a favor da hipotese, ou seja calcula se o valor da funcao surpresa relativa
        %e menor para o ponto aleatorizado do que no ponto de maximo
        cont = cont + 1;
    end
end
evaluate = cont/num;%calcula a proporcao de pontos a favor da hipotese

```

A.4.1.1 O otimizador para o cálculo do ponto de máximo

Esta função é chamada na função que calcula o *e-value* quando opta-se pelo uso do otimizador no lugar do EMV:

```
%função utilizada nas iterações da otimização para o teste de Independência  
function [f] = indepfun(x)  
  
load dados.mat%carrega os dados salvos  
  
p = log( 1/((1-x(1))*(1-x(2))) * [ x(1)*x(2); x(1)*(1-x(2)) ; (1-x(1))*x(2) ] );% vetor teta reparametrizado  
f = -( (transp(p) * invsigmaR * p) - (transp(p - m) * invsigma * (p-m)) ) ;% funcao surpresa relativa
```

A.5.2 TRVG

Com o programa abaixo, é possível calcular o p-value para uma única amostra, basta digitar na janela de comandos do MatLab o seguinte texto:

p_homog([x y z w])

onde x, y, z e w são os dados observados na amostra.

```
%p_indep  
function [pvalue] = p_indep(obs)  
  
kkkk = obs(1,1);%dados amostrais  
kkk = obs(1,2);  
kk = obs(2,1);  
k = obs(2,2);  
  
soma = k + kk + kkk + kkkk;%tamanho da amostra  
  
u = [(kkkk + kkk)*(kkkk + kk) (kkkk + kkk)*(kkk + k) ; (kk + k)*(kkkk + kk) (kk + k)*(kkk + k)]/soma;%vetor teta sob H  
chi2 = 0;  
for i = 1:2  
    for j = 1:2  
        chi2 = chi2 + 2 * obs(i,j) * log( obs(i,j)/u(i,j) );% calculo da Estatística do TRVG  
    end  
end  
pvalue = (1-chi2cdf(chi2,1));%Aproximacao da estatística pela Qui Quadrado com 1 grau de liberdade
```

A.5.3 O programa para calcular as duas estatísticas para grandes amostras

Este programa utiliza os programas que fazem os cálculos de *e-value* e *p-value* e o “tracamedia.m” e possibilita a repetição dos testes para um número maior de amostras, dando as seguintes opções na tela:

- Escolha do tamanho das amostras a serem observadas (n);
- Escolha da quantidade de pontos amostrais: fixa (número a escolher) ou todo o espaço amostral;
- Escolha para o cálculo do ponto de máximo usado na função surpresa relativa: o EMV ou o otimizador;
- Após a finalização do teste, é dada a opção de calcular os parâmetros a e b da função Beta acumulada que melhor se ajusta aos pontos obtidos.

Se não houver opção para se obter os parâmetros da função Beta, os pontos (*p-value*, *e-value*) serão mostrados automaticamente em um gráfico, junto com a curva da função Beta acumulada com os parâmetros obtidos na ocasião do capítulo 4.

```
%Independencia
```

```
clear %limpa as variaveis do espaço de trabalho
close all %fecha todas as janelas de figura q estejam abertas
clc %limpa a janela de comando
```

```
%soma eh a variavel q controla o valor da soma dos elementos dos dados observados
```

```
fprintf('Digite o valor da soma dos dados observados: \n');
soma = input('Valor deve ser maior ou igual a 30: ');
while isempty(soma) | soma < 30
    soma = input('Valor deve ser maior ou igual a 30: ');
end
```

```
fprintf('\nEscolha uma das opcoes abaixo para gerar os dados observados para a soma escolhida:\n');
fprintf('1 - Gera uma certa quantidade de dados obseavados randomicamente;\n');
fprintf('2 - Gera todos os dados possiveis.\n');
metodo1 = input('(Opcao padrao = 1): ');
if isempty(metodo1) | metodo1 < 1 | metodo1 > 2
    metodo1 = 1;
end
```

```
if metodo1 == 1
    fprintf('\nDigite o numero de ponto que devem ser randomizados: \n')
    numero_rand = input('(Numero padrao = 50) :');
    if isempty(numero_rand)
        numero_rand = 50;
    end
end
```

```

end

fprintf('\nEscolha uma das opções abaixo para gerar o valor maximo para a comparação do e-
value:\n');
fprintf('1 - Estimador;\n');
fprintf('2 - Otimizador.\n');
metodo2 = input('(Opção padrao = 1): ');
if isempty(metodo2) | metodo2 < 1 | metodo2 > 2
    metodo2 = 1;
end

fid = fopen('compara_independencia.doc','w'); %abre arquivo para gravação
fprintf(fid,'e-value p-value observações\n\n'); %escreve no arquivo aberto

if metodo1 == 1
    for i = 1:numero_rand
        i
        flag = 0;

        obs = zeros(2);
        while (flag == 0) %randomiza uma observacao valida
            flag = 1;
            obs(1,1) = round(rand*soma);
            if(obs(1,1) <= 5)
                flag = 0;
            end
            obs(1,2) = round(rand*(soma-obs(1,1)));
            if(obs(1,2) <= 5)
                flag = 0;
            end
            obs(2,1) = round(rand*(soma-obs(1,1)-obs(1,2)));
            if(obs(2,1) <= 5)
                flag = 0;
            end
            obs(2,2) = soma - obs(1,1)-obs(1,2)-obs(2,1);
            if(obs(2,2) <= 5)
                flag = 0;
            end
        end

        obs

        pvalue(i) = p_indep(obs); %calcula o p-value
        evalue(i) = e_indep(obs,metodo2); %calcula o e-value

        fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
        fprintf(fid,'%d\t',obs(:));
        fprintf(fid,'\n');
    end
end

if metodo1 == 2

    i = 1; %variavel de controle de fluxo para os valores de e-value e pvalue

    for kkkk = 5:floor(soma/2)
        for k = 5:(soma-kkkk)
            for kkk = 5:floor((soma-kkkk-k)/2)
                kk = soma - kkkk - k - kkk;
            end
        end
    end
end

```

```

if kk > 4

    obs = [kkkk kkk; kk k] %vetor de dados observados

    pvalue(i) = p_indep(obs); %calcula o e-value
    evalue(i) = e_indep(obs,metodo2); %calcula o p-value

    fprintf(fid,'%f\t%f\t',evalue(i),pvalue(i));
    fprintf(fid,'%d\t',obs(:));
    fprintf(fid,'\n');
    i = i + 1;
end
end
end
end

fclose(fid); %fecha o arquivo aberto para gravação

Ev = evalue(:);
Pv = pvalue(:);

%traça o grafico dos pares p-value e e-value junto com a melhor beta
figure(1)
clf
plot(Pv,Ev,'g.')
hold
PV = 0:.001:1;
EV = betacdf(PV,0.7092,3.0751);
plot(PV,EV,'r');
title('Teste de Independencia')
xlabel('p-value')
ylabel('e-value')

fprintf('\nDeseja obter os parametros da beta media?\n')
traca_media = input('(padrao 0), 1 (um) p/ Sim e 0 (zero) p/ Nao ? ');
if isempty(traca_media) | traca_media < 0 | traca_media > 1
    traca_media = 0;
end

if traca_media == 1
    tracamedia
end

```