**Teste de significância em tabelas de contingência
2x2 usando o modelo logístico – normal**

William Waiteman Rodrigues

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

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

São Paulo, fevereiro de 2006.

Capítulo 1

Introdução

Embora os objetivos dessa dissertação sejam modestos, consideramos relevantes para o desenvolvimento do FBST (Full Bayesian Significance Test) para dados gerados por processos de Bernoulli. Dependendo da regra de parada do experimento, dados gerados por processos de Bernoulli produzem modelos estatísticos conhecidos como Hipergeométrico Multivariado, Multinomial ou Multinomial Negativa. As análises apresentadas aqui são independentes da regra de parada e somente exigem a permutabilidade do processo multivariado de Bernoulli gerador. A discussão se restringirá ao caso das tabelas de contingência 2x2. Com relação à extensão para dimensões maiores, pensamos ser apenas um exercício algébrico.

O FBST – Full Bayesian Significance Test – é um teste estatístico de significância Bayesiano baseado apenas na distribuição a posteriori e seu principal objetivo é a determinação da *evidência* que os dados carregam em favor de uma *hipótese precisa*. Por hipótese precisa entendemos a hipótese que é definida por um conjunto de valores paramétricos com dimensão menor do que o seu complementar, o qual define a hipótese alternativa.

Estaremos restritos ao seguinte modelo:

1. O vetor de observações do experimento, $X = (x_1, x_2, x_3, x_4)$, onde, para $i = 1, 2, 3, 4$, x_i é um número inteiro não negativo;
2. O modelo gerador dos dados é um processo de Bernoulli $\{U_k\}_{k \geq 1}$ de ordem 4 e dimensão 3. Isto é:

- i) O parâmetro é um vetor $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ onde, para $i = 1, 2, 3, 4$, θ_i é um racional não negativo e tal que $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$;
 - ii) Com $e_1=(1,0,0,0)$, $e_2=(0,1,0,0)$, $e_3=(0,0,1,0)$ e $e_4=(0,0,0,1)$, a distribuição marginal do processo é $P\{U_k = e_i | \theta\} = \theta_i$ para todo inteiro $k \geq 1$ e $i=1,2,3,4$; e
 - iii) Condicionado ao conhecimento de θ , as variáveis U_k são mutuamente independentes e identicamente distribuídas como em ii).
3. Ao longo da dissertação, considerando que $X = U_1 + U_2 + \dots + U_k$ para algum $k \geq 1$, a verossimilhança básica associada a cada um dos problemas discutidos é obtida da seguinte função:

$$L(\theta | X) \propto \prod_{i=1}^4 \theta_i^{x_i}$$

4. A classe de distribuições a priori será restrita a classe das Distribuições de Dirichlet de ordem 4 e dimensão 3. Para um vetor de racionais positivos $a = (a_1, a_2, a_3, a_4)$, a seguinte densidade (de θ) é a que irá gerar todas as densidades a priori utilizadas ao longo da dissertação

$$g(\theta | a) = \Gamma\left(\sum_{j=1}^4 a_j\right) \prod_{i=1}^4 \frac{\theta_i^{a_i-1}}{\Gamma(a_i)},$$

onde $\Gamma(y)$ é a função Gama avaliada no ponto y .

5. Utilizando-se uma notação simplificada, seja $A_i = a_i + x_i$ para $i = 1, 2, 3, 4$. Denotando o parâmetro da posteriori por $A = a + X = (A_1, A_2, A_3, A_4)$, a densidade da distribuição a posteriori será $g(\theta | A)$.

Os problemas que serão discutidos nesta dissertação são estudados em farta literatura tanto freqüentista quanto Bayesiana. Muitos dos artigos descrevendo esses problemas já estão disponíveis por mais de 80 anos. Assim apenas iremos listar esses casos e apresentar os parâmetros de interesse envolvidos. A ordem de apresentação vai do mais simples para o mais difícil:

1. **Teste de Proporção:** Aqui estamos interessados em verificar se o valor de θ_i é uma proporção especificada. Isto é, considere um número no intervalo $[0;1]$, digamos b . O objetivo é testar $H: \theta_i = b$ contra uma alternativa $A: \theta_i \in B$, onde B é um conjunto próprio de $[0;1]$: digamos $B = [0;1]$ ou

mesmo $B = [b; 1]$. Note que o conjunto que define H tem dimensão zero enquanto o que define A tem dimensão 1.

Observação: Note que o caso não singular, onde tanto A quanto H possuem a mesma dimensão do espaço paramétrico, o cálculo da evidência se resume em calcular as probabilidades desses conjuntos. Assim a comparação dos valores de $P(H|X)$ e $P(A|X)$ irão permitir a definição de bons critérios de decisão sobre a aceitação de H ou de A . Esta dissertação se restringe ao caso onde H e A são definidas por conjuntos de diferentes dimensões.

2. **Teste de McNemar:** Aqui estamos interessados em testar $H: \theta_2 = \theta_3$ contra a alternativa $A: \theta_2 \neq \theta_3$. Trata-se novamente de um problema de testar hipótese precisa, pois a dimensão do conjunto que define H (A) é dois (três).
3. **Teste de Homogeneidade:** Consideremos agora que os valores paramétricos das marginais da tabela são conhecidos. Por exemplo, suponha que $\theta_1 + \theta_2 = \lambda$ é conhecido. Definindo os parâmetros $\gamma_1 = \theta_1 \div \lambda$ e $\gamma_2 = \theta_3 \div (1 - \lambda)$, o objetivo é testar a hipótese $H: \gamma_1 = \gamma_2$ contra a alternativa $A: \gamma_1 \neq \gamma_2$. Consequentemente, a dimensão de H (A) é um (dois).
4. **Teste do Equilíbrio de Hardy-Weinberg:** O parâmetro agora é definido por $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ satisfazendo a seguinte relação: $(\gamma_1, \gamma_2, \gamma_3) = (\theta_1, \theta_2 + \theta_3, \theta_4)$. Considerando o parâmetro $\lambda = \gamma_1 + (\frac{1}{2})\gamma_2$, o objetivo é testar a hipótese nula $H: \gamma_1 = \lambda^2$ e $\gamma_3 = (1 - \lambda)^2$ contra a alternativa A definida pelo conjunto complementar da linha que define H em relação ao simplex: $\{(\gamma_1, \gamma_2, \gamma_3) > 0 \wedge \gamma_1 + \gamma_2 + \gamma_3 = 1\}$. Consequentemente, a dimensão de H (A) é um (dois).
5. **Teste de Independência:** Talvez este seja o teste mais celebrado da literatura estatística. O espaço paramétrico é o original e apenas na hipótese H é onde o espaço se restringe. Com as funções $\theta_1 + \theta_2 = \lambda$ e $\theta_1 + \theta_3 = \sigma$, o interesse neste caso é testar as seguintes hipóteses:
 $H: \theta \in \{(\theta_1, \theta_2, \theta_3, \theta_4) > 0 \wedge \theta_1 + \theta_2 + \theta_3 + \theta_4 = 1 \wedge \theta_1 = \lambda\sigma \wedge \theta_4 = (1 - \lambda)(1 - \sigma)\}$ contra
 $A: \theta \in \{(\theta_1, \theta_2, \theta_3, \theta_4) > 0 \wedge \theta_1 + \theta_2 + \theta_3 + \theta_4 = 1\}$.
 Agora a dimensão de H (A) é dois (três).

O principal objetivo desta dissertação é apresentar ao leitor os cálculos do valor da evidência. Evidência que os dados carregam em relação à veracidade da

hipótese H sendo testada. Esse valor de evidência, valor e (***e-value*** ou ***ev***), é obtido no espaço paramétrico utilizando-se apenas a densidade a posteriori do parâmetro do modelo estatístico. Esse cálculo pode ser considerado como o '*dual*' do valor p (***p-value*** ou ***pv***) que é obtido no espaço amostral. Diferentemente dos trabalhos frequentistas históricos, neste trabalho nenhum método de redução de dimensão será empregado. Podemos afirmar que estamos apresentando uma receita de como avaliar evidência contida nos dados sobre uma hipótese H . Não será necessário modificarmos o método conforme o problema apresentado! Note-se que muitas vezes a eliminação do parâmetro '*nuisance*' (incômodo) é necessária na metodologia frequentista. O método de eliminação pode se modificar de problema para problema, sendo necessária às vezes a redução por condicionamento, outras por marginalização. Mais problemático é o fato de o condicionante ter de ser modificado com a mudança da regra de parada. Pereira & Lindley (1987) questionam a razão pela qual duas tabelas (binomial e binomial negativa) com mesmas frequências possam produzir diferentes testes – e assim diferentes valores p – para os mesmos parâmetros. O FBST – ***Full Bayesian Significance Test*** – que produz o *e-value* no lugar do *p-value* é baseado apenas na densidade a posteriori e assim não exigindo alguma mágica de redução de dimensão. Se houverem reduções estas são intrínsecas ao método e não por considerações de conveniência.

Os programas de cálculo do *e-value* serão baseados na distribuição normal. Utilizaremos a parametrização

$$\omega = (\omega_1, \omega_2, \omega_3) = \ln \left[\frac{1}{\theta_4} (\theta_1, \theta_2, \theta_3) \right]$$

Com as distribuições definidas acima, essa parametrização tem uma densidade que se aproxima de uma densidade normal tridimensional com matriz de covariância tendo todas as componentes fora da diagonal iguais. O principal resultado que utilizaremos é o fato de a densidade gama, de acordo com a relação

existente entre seus parâmetros, poder ser aproximada por uma densidade log normal. Com isto pode-se facilmente realizar inferências consistentes sobre o logaritmo da razão de chances (razão dos produtos cruzados). Os programas para o uso dos testes de significância serão apresentados em linguagem **R** ao final da apresentação de cada teste.

O Capítulo 2 descreve as principais propriedades da distribuição de Dirichlet. O Capítulo 3 descreve de maneira sucinta o FBST. O Capítulo 4 lista os programas que foram construídos para o cálculo do valor da evidência e finalmente o Capítulo 5 apresenta uma aplicação para dados reais.

Capítulo 2

Propriedades da distribuição de Dirichlet

Neste capítulo apresentaremos a distribuição de Dirichlet e alguns resultados importantes que serão utilizados no desenvolvimento do cálculo da aproximação normal e dos programas computacionais.

2.1 Definição: Um vetor aleatório $\theta \in S_3 = \{\theta \in R^4 \mid \theta > 0 \wedge 1' \theta = 1\}$ tem distribuição de Dirichlet com parâmetro positivo $a \in (R^4)^+$, e escrevemos $\theta | a \sim Di_4(a)$, se sua distribuição tiver a seguinte densidade

$$P(\theta | a) = \frac{1}{B(a)} \prod_{k=1}^4 \theta_k^{a_k - 1},$$

onde

$$B(a) = \int_{\theta \in S_3} \prod_{k=1}^4 \theta_k^{a_k - 1} d\theta = \frac{1}{\Gamma(\sum_{j=1}^4 a_j)} \prod_{k=1}^4 \Gamma(a_k)$$

é o fator de normalização da distribuição.

Antes de enunciar a primeira propriedade, lembremos da densidade Gama. Uma variável aleatória real positiva, y , tem distribuição Gama com parâmetros reais de locação e escala $\alpha > 0$ e $\beta > 0$ - denota-se por $y | (\alpha, \beta) \sim G(\alpha, \beta)$ - se a densidade for a seguinte função:

$$f(y | \alpha, \beta) = \beta^\alpha [y^{\alpha-1} \exp(-\beta y)] \{\Gamma(\alpha)\}^{-1}.$$

2.2 Teorema: CARACTERIZAÇÃO DA DIRICHLET PELA GAMA

Consideremos que as componentes do vetor aleatório $z = (z_1, z_2, z_3, z_4)$ são mutuamente independentes e possuindo distribuições Gama com mesmo parâmetro escala $b > 0$. Isto é, se $a = (a_1, a_2, a_3, a_4)$ é um vetor de componentes reais positivas e $z_i | (a, b) \sim G(a_i; b)$, então:

- i) Com $t = 1'z = z_1 + z_2 + z_3 + z_4$, $t | (a; b) \sim G(1'a; b)$;
- ii) Se $\theta = z t^{-1}$, então $\theta | a \sim Di_4(a)$ e
- iii) Fixado o vetor a , θ e t são independentes.
- iv) A média e a matriz de covariância de θ são respectivamente:

$$m = E\{\theta | a\} = a t^{-1} \text{ e } C = \{Diag(m) - m'm\}(t+1)^{-1}.$$

(Lembremos que m é um vetor linha, $Diag(m)$ é uma matriz diagonal de ordem 4 cujos elementos da diagonal são as componentes de m e $m'm$ é uma matriz quadrada de ordem 4 obtida pela multiplicação da transposta de m (vetor coluna) pelo vetor linha m .)

Demonstração: Vamos considerar a transformação

$$\theta = \frac{1}{t} z, \quad t = 1'z \Rightarrow z = t\theta.$$

Com essa transformação, uma das variáveis, por digamos $z_4 = t(1 - \theta_1 - \theta_2 - \theta_3)$, torna-se redundante e a matriz Jacobiana da transformação é

$$J = \frac{\partial(z_1, z_2, z_3, z_4)}{\partial(\theta_1, \theta_2, \theta_3, t)} = \begin{bmatrix} t & 0 & 0 & \theta_1 \\ 0 & t & 0 & \theta_2 \\ 0 & 0 & t & \theta_3 \\ -t & -t & -t & 1 - \theta_1 - \theta_2 - \theta_3 \end{bmatrix}$$

Realizando algumas operações com matrizes, obtemos a fatoração desta matriz como $J = AB$ onde, onde A e B são as matrizes triangulares apresentadas a seguir:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{bmatrix} \text{ e } B = \begin{bmatrix} t & 0 & 0 & \theta_1 \\ 0 & t & 0 & \theta_2 \\ 0 & 0 & t & \theta_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Como A e B são matrizes triangulares, os determinantes são iguais à produtória dos elementos da diagonal principal, portanto $|J| = |A||B| = t^3$.

A distribuição conjunta de z é

$$f(z|a, b) = \prod_{k=1}^4 f(z_k|a_k, b) = \prod_{k=1}^4 \frac{b^{a_k}}{\Gamma(a_k)} e^{-bz_k} (z_k)^{a_k-1}$$

e a densidade conjunta de t e θ é

$$\begin{aligned} g(\theta, t|a, b) &= |J| f(z^{-1}[\theta, t]) = t^3 \prod_{k=1}^4 \frac{b e^{-bt\theta_k}}{\Gamma(a_k)} (bt\theta_k)^{a_k-1} = \\ &= \left(\prod_{k=1}^4 \frac{\theta_k^{a_k-1}}{\Gamma(a_k)} \right) b^{1'a} e^{-bt} t^{1'a-1}. \end{aligned}$$

A densidade marginal de θ é obtida como

$$g(\theta|a, b) = \int_{t=0}^{\infty} g(\theta, t|a, b) dt = \left(\prod_{k=1}^4 \frac{\theta_k^{a_k-1}}{\Gamma(a_k)} \right) b^{1'a} \int_{t=0}^{\infty} e^{-bt} t^{1'a-1} dt$$

O último fator desse produto é a função Gama avaliada no ponto (α, β) onde $\alpha = 1'a$ e $\beta = b$. Isto é, $\Gamma(\alpha) \div \beta$. Finalmente temos

$$g(\theta|a, b) = g(\theta|a) = \left(\prod_{k=1}^4 \frac{\theta_k^{a_k-1}}{\Gamma(a_k)} \right) \Gamma(1'a) = \frac{1}{B(a)} \prod_{k=1}^4 \theta_k^{a_k-1}$$

Isto é, $\theta|a$ tem distribuição $Di_4(a)$.

Para o cálculo da média e da matriz de covariância basta lembrar que $\theta_1|a$ e

$(\theta_1 + \theta_2)|a$ têm distribuição beta com médias $m_1 = a_1 \div \alpha$ e $M = m_1 + m_2 =$

$(a_1 + a_2) \div \alpha$, onde $\alpha = 1'a$.

As variâncias são dadas por

$$v_1 = m_1(1 - m_1) \div (\alpha + 1) \text{ e } V = M(1 - M) \div (\alpha + 1).$$

Como $V = v_1 + v_2 + \text{Cov}(\theta_1, \theta_2)$ obtemos assim os elementos da diagonal da matriz de covariância, $\text{Cov}(\theta_i, \theta_j|a)$. ■

Alguns problemas importantes são definidos por testes de hipóteses baseados em somas parciais dos elementos de θ . Assim, essas somas se tornam relevantes e suas distribuições são apresentadas no resultado enunciado a seguir.

2.3 Lema: PARTIÇÃO DA DISTRIBUIÇÃO DE DIRICHLET

Consideremos que $\theta | a \sim \text{Di}_4(a)$, a parametrização definida por

$$\theta^{(1)} = \frac{1}{\theta_1 + \theta_2}(\theta_1, \theta_2), \quad \theta^{(2)} = \frac{1}{\theta_3 + \theta_4}(\theta_3, \theta_4) \text{ e } w = \begin{bmatrix} \theta_1 + \theta_2 \\ \theta_3 + \theta_4 \end{bmatrix},$$

possui as seguintes propriedades:

- i) Independência: $\theta^{(1)} \perp\!\!\!\perp \theta^{(2)} \perp\!\!\!\perp w$.
- ii) Está na mesma classe de distribuições:
 $\theta^{(1)} \sim \text{Di}_2(a_1, a_2), \quad \theta^{(2)} \sim \text{Di}_2(a_3, a_4) \text{ e } w \sim \text{Di}_2(a_1 + a_2, a_3 + a_4).$

Demonstração: Pelo Teorema 2.2 podemos imaginar que θ é obtido da normalização de um vetor z com componentes Gamas como descrito em 2.2.

$$\theta = \frac{1}{1'z} z, \quad 1'z \sim G(1'a, b) \text{ e } \theta \perp\!\!\!\perp 1'z$$

Consideremos os vetores $\theta^{(1)}$ e $\theta^{(2)}$ que são distribuídos como

$$\theta^{(1)} = \frac{1}{\theta_1 + \theta_2}(\theta_1, \theta_2) = \frac{1}{z_1 + z_2}(z_1, z_2) \sim \text{Di}_2(a_1, a_2) \text{ e}$$

$$\theta^{(2)} = \frac{1}{\theta_3 + \theta_4}(\theta_3, \theta_4) = \frac{1}{z_3 + z_4}(z_3, z_4) \sim \text{Di}_2(a_3, a_4).$$

onde $\theta^{(1)} \perp\!\!\!\perp \theta^{(2)} \perp\!\!\!\perp w | a$ e além disso, fixado a , são independentes das somas parciais (independentes por construção) apresentadas a seguir:

$$z_1 + z_2 | (a, b) \sim G(a_1 + a_2, b) \quad \text{e} \quad z_3 + z_4 | (a, b) \sim G(a_3 + a_4, b)$$

Utilizando novamente a caracterização para as duas novas variáveis Gama, temos que:

Como $w = \frac{1}{1'z}(z_1 + z_2; z_3 + z_4)$, $w | (a, b) \sim Di_2(a_1 + a_2, a_3 + a_4)$ ■

A discussão que apresentamos a seguir é o coração desta dissertação pois aqui descreveremos a distribuição do logaritmo da razão de chances. Com a utilização das propriedades aqui descritas, acreditamos que os programas construídos para análise de dados discretos, provenientes de processos de Bernoulli, devam se tornar bastante eficientes.

Não custa ressaltar que todos os resultados que apresentamos acima e os que vamos desenvolver na sequência, são bastante gerais e são facilmente estendidos para quaisquer dimensões com diferentes partições.

A partir desse ponto iremos discutir as distribuições da razão de chances e dos seus logaritmos. A transformação logística aplicada a uma distribuição normal p-dimensional sob um simplex p-dimensional é conhecida como distribuição logístico – normal, segundo Aitchison e Shen (1980). A definição a seguir não é restrita a escolha do denominador. Com qualquer elemento do vetor θ no denominador a definição pode ser modificada apenas permutando-se os índices.

2.4 Definição: Se $\theta|a \sim Di_4(a)$, então o vetor ϕ definido como

$\phi = (\theta_1, \theta_2, \theta_3) \div \theta_4$ é dito ter distribuição de Dirichlet do segundo tipo, de ordem 4 com parâmetro a .

2.5 Teorema: DENSIDADE DIRICHLET DO SEGUNDO TIPO

A densidade de $\phi|a$ é dada pela seguinte função:

$$f(\phi | a) = \frac{\prod_{k=1}^3 \phi_k^{a_k - 1}}{B(a)(1 + 1'\phi)^{1'a}}$$

Demonstração: A demonstração é imediata se lembrarmos os passos da demonstração do Teorema 2.2 e o fato de

$$\phi = (\phi_1, \phi_2, \phi_3) = \frac{1}{z_4} (z_1, z_2, z_3),$$

Como antes, para a e b fixados, z_1, z_2, z_3 , e z_4 são mutuamente independentes e, para $k=1,2,3,4$, $z_k | a, b \sim G(a_k; b)$.

Diversos autores já mencionaram que o logaritmo de uma variável Gama é bem aproximado por uma variável Normal, veja Aitchison e Shen (1980) e Lindley (1964) para alguns detalhes. Em outras palavras se z tem distribuição gama então existe uma variável y com distribuição log normal equivalente a z . Isto é, $z \approx y$, indicando que as densidades de z e y são aproximadamente iguais. Se esta afirmação for válida, os procedimentos computacionais poderão ser baseados em distribuições normais com matrizes de covariância bastante simples. Isto permite a criação de procedimentos computacionais mais eficientes, que discutiremos no próximo capítulo.

Para o cálculo dos momentos da distribuição log-gama usaremos as funções digama $\psi(\alpha)$ e trigama $\psi'(\alpha)$, definidas como

$$\psi(\alpha) = \frac{\partial}{\partial \alpha} \ln \Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \text{ e } \psi'(\alpha) = \frac{\partial}{\partial \alpha} \psi(\alpha)$$

2.6 Lema: Suponha que $z|(a,b) \sim G(a,b)$. A esperança e a variância de $\ln(z)$ para a e b fixados são, respectivamente,

$$E\{\ln(z)|a,b\} = \psi(a) - \ln(b) \text{ e } V\{\ln(z)|a,b\} = \psi'(a).$$

Demonstração: Por simplicidade, e sempre que estiver implícito, vamos escrever $E\{\ln(z)\}$ e $V\{\ln(z)\}$ no lugar de $E\{\ln(z)|a,b\}$ e de $V\{\ln(z)|a,b\}$. Usando a densidade de z , temos:

$$\int_0^{\infty} f(z) dx = \int_0^{\infty} \frac{b^a}{\Gamma(a)} z^{a-1} e^{-bz} dz = 1$$

derivando com relação a a

$$\int_0^{\infty} \frac{b^a}{\Gamma(a)} \ln(bz) z^{a-1} e^{-bz} dz - \frac{\Gamma'(a)}{\Gamma(a)} \int_0^{\infty} \frac{b^a}{\Gamma(a)} z^{a-1} e^{-bz} dz = 0$$

$$E(\ln(bz)) - \frac{\Gamma'(a)}{\Gamma(a)} \int_0^{\infty} f(z) dz = 0$$

$$E(\ln(bz)) = \frac{\Gamma'(a)}{\Gamma(a)} = \psi(a) \Rightarrow E(\ln(b)) + E(\ln(z)) = \psi(a)$$

$$E(\ln(z)) = \psi(a) - \ln(b).$$

Da demonstração anterior sabemos que $E(\ln(bz)) = \frac{\Gamma'(a)}{\Gamma(a)} = \psi(a)$, e se derivarmos novamente com relação ao parâmetro a

$$\begin{aligned} \psi'(a) &= \frac{\partial}{\partial a} E(\ln(bz)) = \frac{\partial}{\partial a} \int_0^{\infty} \ln(bz) z^{a-1} e^{-bz} \frac{b^a}{\Gamma(a)} dz = \\ &= \int_0^{\infty} \ln(bz)^2 z^{a-1} e^{-bz} \frac{b^a}{\Gamma(a)} dz - \frac{\Gamma'(a)}{\Gamma(a)} \int_0^{\infty} \left(\ln(bz) z^{a-1} e^{-bz} \frac{b^a}{\Gamma(a)} \right) dz = \\ &= E(\ln(bx)^2) - \frac{\Gamma'(a)}{\Gamma(a)} E(\ln(bx)) = E(\ln(bx)^2) - E(\ln(bx))^2 = \\ &Var(\ln(bx)) \Rightarrow Var(\ln(b) + \ln(x)) = V\{\ln(z)\} = \psi'(a) \end{aligned}$$

A demonstração fica assim concluída. ■

Para a verificação da qualidade da aproximação entre a distribuição gama e a distribuição log-normal, realizamos simulações das distribuições para diferentes valores de a e b , que representam os parâmetros da distribuição gama, com os resultados apresentados nas figuras seguintes.

Analisando inicialmente as distribuições com valores dos parâmetros a e b iguais (Figura 1), observamos que, com o crescimento de a (e b), as distribuições gama e log-normal vão se tornando bastante próximas. O valor $a=10$ já fornece

uma aproximação bastante razoável e no caso $a=50$ já não é possível distinguir diferenças importantes entre as densidades.

O teorema 2.2 mostra que na construção da Dirichlet existe plena liberdade na escolha do valor b . Então podemos considerar o valor de b que traga a melhor aproximação entre as distribuições. Com isso, nas simulações a seguir buscamos determinar se existe uma relação entre a e b que melhora esta aproximação mesmo para valores pequenos de a .

As simulações foram realizadas considerando o parâmetro b um múltiplo de a onde $a < b$ (Figura 2) e $a > b$ (Figura 3). Nestes dois casos, podemos observar que as diferentes relações entre os parâmetros não trouxeram mudanças nas diferenças de distribuições se comparadas às já encontradas nos casos onde $a=b$. Como a variância da distribuição gama é dada por a/b^2 , observamos apenas que com o crescimento de $b > 1$ a variabilidade das duas densidades reduz-se consideravelmente.

Com isso consideramos que a aproximação da densidade da gama para a densidade log-normal ocorre já para valores pequenos do parâmetro a e que o parâmetro b afeta a variabilidade das distribuições,

Com as considerações acima, observamos que ao trabalharmos com a densidade $g(\omega | A)$ e com a caracterização da Dirichlet a partir das variáveis gama de parâmetros originados do vetor $A=X+a$, relativamente grandes, afirmamos que a qualidade da aproximação, independente do parâmetro b (que desaparece no processo) utilizado é bastante apropriada.

Figura 1. Comparativo entre as distribuições Gama(a,b) e log-Normal com $a=b$

Figura 2. Comparativo entre as distribuições Gama(a,b) e log-Normal com $a < b$

Figura 3. Comparativo entre as distribuições Gama(a,b) e log-Normal com $a > b$

Com o resultado acima, aliado à Caracterização das distribuições Dirichlet e Dirichlet do segundo tipo por distribuições Gama, pode-se trabalhar com a aproximação normal para a reparametrização da distribuição de Dirichlet.

2.7 Conjectura: Seja $\theta | a \sim Di_4(a)$. O vetor

$$\omega = (\omega_1, \omega_2, \omega_3) = \ln \left[\frac{1}{\theta_4} (\theta_1, \theta_2, \theta_3) \right]$$

tem distribuição aproximadamente normal tridimensional com média

$$\mu_\omega = E(\omega) = \begin{bmatrix} \psi(a_1) - \psi(a_4) \\ \psi(a_2) - \psi(a_4) \\ \psi(a_3) - \psi(a_4) \end{bmatrix}$$

e matriz de covariância

$$\Sigma_\omega = \begin{bmatrix} \psi'(a_1) + \psi'(a_4) & \psi'(a_4) & \psi'(a_4) \\ \psi'(a_4) & \psi'(a_2) + \psi'(a_4) & \psi'(a_4) \\ \psi'(a_4) & \psi'(a_4) & \psi'(a_3) + \psi'(a_4) \end{bmatrix}.$$

Com os resultados apresentados, podemos descrever a distribuição do logaritmo da razão de produtos cruzados. Essa medida de associação tem grande utilização na análise de tabelas de contingência 2x2 e é base para diversos testes de hipóteses, entre eles alguns que serão discutidos no capítulo 4.

2.8 Conjectura: LOGARITMO DA RAZÃO DE PRODUTOS CRUZADOS

Seja $\theta | a \sim Di_4(a)$ e

$$\rho = \ln \left[\frac{\theta_1 \theta_4}{\theta_2 \theta_3} \right]$$

então a distribuição de ρ fixado o parâmetro a é aproximadamente normal com as seguintes média e variância

$$\mu = \psi(a_1) - \psi(a_2) - \psi(a_3) + \psi(a_4) \text{ e}$$

$$\sigma^2 = \psi'(a_1) + \psi'(a_2) + \psi'(a_3) + \psi'(a_4).$$

Capítulo 3

O FBST – Teste de Significância Genuinamente Bayesiano

O FBST – Full Bayesian Significance Test - foi apresentado originalmente por Pereira & Stern (1999) como um teste de significância bayesiano coerente e intuitivo. Apresentaremos aqui a versão generalizada do teste, descrito por Madruga *et al* (2003), que utiliza a densidade referência $r(\theta)$ em Θ .

A escolha da densidade referência é feita no espaço das densidades sobre o espaço paramétrico original, aquele onde definimos nossa priori. Normalmente a escolha de $r(\theta)$ recai sobre a densidade que descreve a “menor” informação sobre θ .

Seja $H: \theta \in \Theta_H$ uma hipótese precisa e $g(\theta)$ uma densidade para θ . Definimos a *surpresa relativa* $s(\theta)$ por

$$s(\theta) = \left\{ \frac{g(\theta)}{r(\theta)} \right\}$$

e com isso

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

Definimos também, no espaço paramétrico Θ , o conjunto Θ^* de pontos $\theta \in \Theta$ com surpresa relativa $s(\theta)$ maior do que qualquer ponto em Θ_H , isto é

$$\Theta^* = \left\{ \theta \in \Theta \mid \frac{g(\theta)}{r(\theta)} \geq s^* \right\}$$

A evidência contra H é dada pela probabilidade posteriori em Θ^*

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

e o valor da evidência em favor de H é definido como $ev = 1 - \overline{ev}$

3.1 Definição: O FBST é o procedimento que rejeita a hipótese nula quando ev for pequeno.

O cálculo do FBST é realizado em dois passos listados a seguir:

Otimização numérica: Tem por objetivo localizar o argumento θ^* que maximiza a surpresa relativa s^* sob a hipótese nula

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

Integração numérica: Tem por objetivo calcular a probabilidade do parâmetro pertencer a região onde a densidade é maior que s^*

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

O programa **R** possui diversas rotinas de otimização global, além de outras funções que serão utilizadas nos programas descritos no próximo capítulo.

Como a classe de distribuições a priori está restrita a classe das Distribuições de Dirichlet de ordem 4, a escolha natural para $r(\theta)$ é a densidade Dirichlet com vetor de parâmetros formado por 1 nas quatro posições.

Ao aplicarmos a reparametrização

$$\omega = (\omega_1, \omega_2, \omega_3) = \ln \left[\frac{1}{\theta_4} (\theta_1, \theta_2, \theta_3) \right]$$

no vetor θ , trabalhamos agora com um novo espaço paramétrico Ω , o que exige uma transformação na hipótese H e nas densidades apresentadas anteriormente.

Seja agora $H: \omega \in \Omega_H$ uma hipótese precisa, $f(\omega)$ uma densidade para ω e $q(\omega)$ a nova função de referência sob a reparametrização. A nova função surpresa é dada por

$$h(\omega) = \frac{f(\omega)}{q(\omega)}, \text{ e com isso}$$

$$\omega^* = \arg \max_{\omega \in \Omega_H} \left\{ \frac{f(\omega)}{q(\omega)} \right\} \quad e \quad h^* = \max_{\omega \in \Omega_H} \left\{ \frac{f(\omega)}{q(\omega)} \right\} = \left\{ \frac{f(\omega^*)}{q(\omega^*)} \right\}.$$

No novo espaço paramétrico Ω , definimos Ω^* como o conjunto de pontos $\omega \in \Omega$ com surpresa relativa $h(\omega)$ maior do que qualquer ponto em Ω_H

$$\Omega^* = \left\{ \omega \in \Omega \left| \frac{f(\omega)}{q(\omega)} \geq h^* \right. \right\}$$

e a evidência contra H é dada por

$$\overline{ev} = \int_{\Omega^*} f(\omega) d\omega$$

A evidência contra H definida em 3.1 continua válida para a parametrização.

As densidades da posteriori e da referência, pela conjectura 2.7, estão definidas. A densidade da posteriori é normal com vetor de médias

$$\mu_\omega = \begin{bmatrix} \psi(A_1) - \psi(A_4) \\ \psi(A_2) - \psi(A_4) \\ \psi(A_3) - \psi(A_4) \end{bmatrix}$$

e matriz de covariância

$$\Sigma_{\omega} = \begin{bmatrix} \psi'(A_1) + \psi'(A_4) & \psi'(A_4) & \psi'(A_4) \\ \psi'(A_4) & \psi'(A_2) + \psi'(A_4) & \psi'(A_4) \\ \psi'(A_4) & \psi'(A_4) & \psi'(A_3) + \psi'(A_4) \end{bmatrix}$$

A densidade de referência $q(\omega)$ é normal com vetor de médias

$$\mu_q = \begin{bmatrix} \psi(1) - \psi(1) \\ \psi(1) - \psi(1) \\ \psi(1) - \psi(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

e matriz de covariância

$$\Sigma_q = \begin{bmatrix} 2\psi'(1) & \psi'(1) & \psi'(1) \\ \psi'(1) & 2\psi'(1) & \psi'(1) \\ \psi'(1) & \psi'(1) & 2\psi'(1) \end{bmatrix}.$$

A surpresa relativa será então

$$h(\omega) = \frac{f(\omega)}{q(\omega)} = |\Sigma_{\omega}|^{-1/2} |\Sigma_q|^{1/2} \exp \left[\frac{1}{2} \left(\omega' \Sigma_q^{-1} \omega - (\omega - \mu_{\omega})' \Sigma_{\omega}^{-1} (\omega - \mu_{\omega}) \right) \right].$$

Alguns testes de hipóteses são particularizações dos resultados gerais apresentados acima. Como exemplo podemos citar o caso do teste de proporção, cujo vetor de parâmetros $\theta = (\theta_1, \theta_2)$ tem dimensão menor que três. No Capítulo 4 apresentamos essas particularizações e os programas desenvolvidos para cada teste de hipótese considerando a reparametrização ω .

Capítulo 4

Programas

Neste capítulo apresentaremos os programas desenvolvidos para os testes de hipótese apresentados no capítulo 1.

O processo de cálculo do FBST a ser utilizado nos programas terá dois passos:

- ✓ A **otimização numérica** será utilizada nos casos onde a dimensão de $\mathbf{H}$ for diferente de zero e, com o valor ω^* determinado, será obtida a surpresa relativa h^* ;
- ✓ A **integração numérica** da probabilidade a posteriori será efetuada em três passos:
 - simulação de n pontos com distribuição igual a $f(\omega | \mathbf{A})$;
 - contagem c de quantos pontos da simulação possuem surpresa relativa menor que h^*

$$c = \sum_{i=1}^n \mathbf{I} \left\{ \frac{f(\omega_i)}{q(\omega)} < h^* \right\} \text{ e};$$

- a proporção c/n é a aproximação de ev .

Com o objetivo de se obter uma maior estabilidade numérica, os cálculos serão realizados com uma função proporcional a $\ln(h(\omega))$. Para a simulação de vetores com distribuição normal multivariada, utilizaremos a biblioteca *MASS*, Venables & Ripley (1999). A instalação e ativação desta biblioteca no programa **R** são necessárias para o cálculo do FBST dos testes de McNemar, Hardy-Weinberg e Independência.

Para cada teste de hipótese discutido apresentaremos um exemplo onde comparamos os *e-values* obtidos pelo programa FBST desenvolvido nesta

dissertação aos p -values obtidos nos testes qui-quadrado tradicionais, sempre considerando a distribuição priori uniforme. Quando possível, os e -values obtidos serão comparados aos e -values calculados por Pereira & Stern (1999) diretamente no espaço paramétrico original, portanto sem a reparametrização e a utilização da aproximação normal.

1.1. Teste de proporção

Neste teste consideramos uma variável aleatória x com distribuição binomial de parâmetro θ e tamanho de amostra n . O teste de nosso interesse é um teste preciso do tipo $H: \theta = b$ contra uma alternativa $A: \theta \in B$, onde b é um escalar e B é um conjunto próprio de $[0;1]$. Com a reparametrização ω teremos como nova hipótese de teste

$$H : \omega = \ln\left(\frac{\theta}{1-\theta}\right) = \ln\left(\frac{b}{1-b}\right) = v$$

A densidade a posteriori $f(\omega)$ e a densidade de referência $q(\omega)$ são normais com médias e variâncias dadas por

$$\begin{aligned}\mu_{\omega} &= \psi(A_1) - \psi(A_2) \\ \sigma_{\omega}^2 &= \psi'(A_1) + \psi'(A_2) \text{ e} \\ \sigma_q^2 &= 2\psi'(1)\end{aligned}$$

Como a dimensão de H é zero, conseguimos obter os valores máximos do parâmetro ω e da surpresa relativa no espaço paramétrico nulo h^* de forma direta

$$\omega^* = \arg \max_{\omega \in \Omega_H} \left\{ \frac{f(\omega)}{q(\omega)} \right\} = v, \text{ bem como } h^* = \max_{\omega \in \Omega_H} \left\{ \frac{f(\omega)}{q(\omega)} \right\} = \left\{ \frac{f(v)}{q(v)} \right\}.$$

A função surpresa de ω é dada por

$$h(\omega) = \frac{\sigma_q}{\sigma_{\omega}} \exp \left[\frac{1}{2} \left(\frac{\omega^2}{\sigma_q^2} - \frac{(\omega - \mu_{\omega})^2}{\sigma_{\omega}^2} \right) \right]$$

Para realizarmos um comparativo entre a medida de evidência fornecida pelo FBST com o p -value, consideramos uma amostra de $n=20$ tentativas e x_1 é o número de sucessos obtidos. A hipótese a ser testada é $H: \theta = 0,5$. Para todos os

valores possíveis de x_1 , a tabela a seguir fornece os resultados das duas medidas. O valores de evidência obtidos com a aproximação normal não apresentam diferenças nas duas primeiras casas decimais aos valores de evidência obtidos no FBST calculado no espaço paramétrico original.

x_1	Ev	pv
0	0,00	0,00
1	0,00	0,00
2	0,00	0,00
3	0,00	0,00
4	0,01	0,01
5	0,03	0,02
6	0,07	0,06
7	0,17	0,16
8	0,36	0,35
9	0,65	0,64
10	1,00	1,00

O programa em R que calcula o FBST com a reparametrização para o teste da proporção está apresentado abaixo. Como dados de entrada, ele exige o parâmetro a da densidade priori, os dados observados e o valor da proporção b que deseja ser testada.

Programa 1. Programa para o teste de proporção

```
num.simul<-100000
n<-sum(obs)
conta<-0
param<- priori+ obs
## matrizes de médias e covariâncias
mu<- digamma(param[1])-digamma(param[2])
sigma<-sqrt(trigamma(param[1])+trigamma(param[2]))
sigma.r<- sqrt(2* trigamma(1))
dens.l <- function(v) {
  (v/sigma.r)^2-((v-mu)/sigma)^2
}
dens.H<- dens.l(log(H/(1-H)))
simul.valor<-rnorm(num.simul, mean=mu, sd=sigma)
dens.valor<- dens.l(simul.valor)
for (j in 1:num.simul) {
  if(dens.valor[j]< dens.H) conta<-conta+1
}
cat('Full Bayesian Significance Test', "\n")
cat('Test for binomial', "\n", "\n")
cat('Null Hypothesis: H=', H, "\n", "\n")
cat('Priori and Observed Values provided: ', "\n")
print(cbind(X=rbind(Priori =priori[1:2],
Observed=obs[1:2]),Total=c(sum(priori),sum(obs))))
cat(' ', "\n")
cat('Probabilities: ', "\n")
```

```
print(cbind(X=rbind('Null Hypothesis'= round(c(H, 1-H),3), Posterior=
round((priori[1:2]+obs[1:2])/(sum(priori)+sum(obs)),3))))
cat(' ', "\n")
cat("Evidence for", obs[1], "successes:", round(conta/num.simul, 3), "\n")
```

1.2. Teste de McNemar

Neste teste consideramos uma variável aleatória x com distribuição multinomial de parâmetro $\theta=(\theta_1, \theta_2, \theta_3, \theta_4)$ e tamanho de amostra n . O teste de interesse é

$$H: \theta_2 = \theta_3 \text{ contra a alternativa } A: \theta_2 \neq \theta_3.$$

Com a reparametrização de θ , obtemos o novo teste de hipótese

$$H: \omega_2 = \omega_3$$

com a distribuição posteriori, função de referência e surpresa relativa como as apresentadas anteriormente no Capítulo 3

$$h(\omega) = |\Sigma_\omega|^{-1/2} |\Sigma_q|^{1/2} \exp \left[\frac{1}{2} \left(\omega' \Sigma_q^{-1} \omega - (\omega - \mu_\omega)' \Sigma_\omega^{-1} (\omega - \mu_\omega) \right) \right].$$

O comparativo do *e-value* com o *p-value* foi realizado alterando-se o valor de x_2 no vetor $x = (x_1, x_2, x_3, x_4) = (20 - x_2, x_2, 10, 10)$, com resultados apresentados na tabela abaixo.

x_2	ev	pv
10	1,00	1,00
11	0,99	1,00
12	0,97	0,83
13	0,93	0,67
14	0,87	0,54
15	0,79	0,42
16	0,71	0,32
17	0,60	0,25
18	0,49	0,18
19	0,38	0,14
20	0,25	0,10

O programa em **R** que calcula o FBST com aproximação normal para o teste de McNemar está apresentado abaixo. Como dados de entrada, ele exige o parâmetro a da densidade priori e os dados observados.

Programa 2. Programa para o teste de McNemar

```

num.simul<-10000
conta<-0
param<- priori+obs
teta<-param/sum(param)
par.H<-c(teta[1], mean(teta[2:3]), mean(teta[2:3]), teta[4])
mu<- c(digamma(param[1]), digamma(param[2]), digamma(param[3]))-digamma(param[4])
sigma<-
diag(c(trigamma(param[1]), trigamma(param[2]), trigamma(param[3]))+trigamma(param[4])
sigma.r <-
diag(c(trigamma(priori[1]), trigamma(priori[2]), trigamma(priori[3]))+trigamma(priori[4])
dens.l <- function(v) {
  (t(v)%*%solve(sigma.r)%*%v) - (t(v - mu)%*%solve(sigma)%*%(v - mu))
}
f.otim <- function(x) {
  p<-log((1/(teta[4]))* c(teta[1],x,x))
  dens.l(p)
}
result<-optimize(f.otim, interval
=c(min(teta[2],teta[3]),max(teta[2],teta[3])+0.0001), maximum = TRUE)
dens.sn<- result$objective
set.seed=33
valor<-mvrnorm(num.simul, mu=mu, Sigma=sigma)
for (j in 1:num.simul) {
  if(dens.l(valor[j,])<dens.sn) conta<-conta+1
}

cat('Full Bayesian Significance Test',"\\n")
cat('Test for McNemar', "\\n", "\\n")
cat('Data Provided: ', "\\n")
cat('Priori: ', "\\n")
print(cbind(X=rbind(A=priori[1:2], B=priori[3:4], Total=
c(priori[1]+priori[3], priori[2]+priori[4])), Total=c(sum(priori[1:2]), sum(priori[3:4]), sum(priori)))
cat(' ', "\\n")
cat('Observed values:', "\\n")
print(cbind(X=rbind(A=obs[1:2], B=obs[3:4], Total=
c(obs[1]+obs[3], obs[2]+obs[4])), Total=c(sum(obs[1:2]), sum(obs[3:4]), sum(obs)))
cat(' ', "\\n")
cat('Posteriori Probabilities:', "\\n")
print(cbind(X=rbind(A=round(teta[1:2],3), B=round(teta[3:4],3), Total=
round(c(teta[1]+teta[3], teta[2]+teta[4]),
3)), Total=round(c(sum(teta[1:2]), sum(teta[3:4]), sum(teta)),3)))
cat(' ', "\\n")
cat("Evidence = ", format(conta/num.simul, digits=3), "\\n")

```

1.3. Teste de homogeneidade

Para o teste de homogeneidade vamos considerar que o valor dos parâmetros de uma das marginais da tabela, por exemplo λ , é conhecido.

Categoria 2	Categoria 1		Total
	1	2	
1	θ_1	θ_2	λ
2	θ_3	θ_4	$(1-\lambda)$

Definindo os parâmetros $\gamma_1 = \theta_1/\lambda$ e $\gamma_2 = \theta_3/(1-\lambda)$, o objetivo é testar a hipótese **H**: $\gamma_1 = \gamma_2$ contra a alternativa **A**: $\gamma_1 \neq \gamma_2$

A partir do lema 2.3, podemos afirmar que os novos parâmetros γ_1 e γ_2 também possuem distribuição de Dirichlet independentes

$$\begin{aligned}\gamma_1 | A &\sim Di_2(A_1, A_2) \\ \gamma_2 | A &\sim Di_2(A_3, A_4)\end{aligned}$$

que, com a reparametrização ω aplicada em γ_1 e γ_2 teremos como nova hipótese de teste

$$\begin{aligned}\omega_i &= \ln\left(\frac{\gamma_i}{1-\gamma_i}\right), \quad i = 1, 2 \\ H : \gamma_1 = \gamma_2 &\Rightarrow \ln\left(\frac{\gamma_1}{1-\gamma_1}\right) = \ln\left(\frac{\gamma_2}{1-\gamma_2}\right) \Rightarrow \omega_1 = \omega_2\end{aligned}$$

As densidades a posteriori e a função de referência de ω são normais com médias e variâncias dadas por

$$\begin{aligned}\mu_1 &= \psi(A_1) - \psi(A_2) \\ \sigma_1^2 &= \psi'(A_1) + \psi'(A_2) \\ \mu_2 &= \psi(A_3) - \psi(A_4) \\ \sigma_2^2 &= \psi'(A_3) + \psi'(A_4) \text{ e} \\ \sigma_q^2 &= 2\psi'(1)\end{aligned}$$

A função surpresa de ω é dada por

$$h(\omega) = \frac{f_1(\omega)}{q(\omega)} \cdot \frac{f_2(\omega)}{q(\omega)} = \frac{\sigma_q^2}{\sigma_1\sigma_2} \exp\left[\frac{1}{2}\left(\frac{\omega_1^2}{\sigma_q^2} - \frac{(\omega_1 - \mu_1)^2}{\sigma_1^2} + \frac{\omega_2^2}{\sigma_q^2} - \frac{(\omega_2 - \mu_2)^2}{\sigma_2^2}\right)\right]$$

Uma outra forma de representar a hipótese de teste é através do logaritmo da razão de produtos cruzados

$$H : \omega_1 = \omega_2 \Rightarrow \rho = 0.$$

A comparação do *e-value* obtido pelo FBST com o *p-value* foi realizado considerando as marginais conhecidas com $n=20$ e alterando-se o valor de x_2 no vetor na forma no vetor $x = (x_1, x_2, x_3, x_4) = (x_1, 20 - x_1, x_3, 20 - x_3)$, com resultados apresentados na tabela abaixo.

x_1	x_3	ev	pv	x_1	x_3	ev	pv
5	0	0,07	0,02	10	6	0,41	0,20
5	1	0,18	0,08	10	7	0,61	0,34
5	2	0,42	0,21	10	8	0,81	0,53
5	3	0,70	0,43	10	9	0,95	0,75
5	4	0,92	0,71	10	10	1,00	1,00
5	5	1,00	1,00	12	0	0,00	0,00
5	6	0,93	0,72	12	1	0,00	0,00
5	7	0,77	0,49	12	2	0,00	0,00
5	8	0,57	0,31	12	3	0,01	0,00
5	9	0,39	0,18	12	4	0,09	0,01
5	10	0,25	0,10	12	5	0,07	0,03
10	0	0,01	0,00	12	6	0,15	0,06
10	1	0,01	0,00	12	7	0,27	0,11
10	2	0,02	0,01	12	8	0,43	0,21
10	3	0,05	0,02	12	9	0,62	0,34
10	4	0,12	0,05	12	10	0,81	0,53
10	5	0,24	0,10				

O programa em R que calcula o FBST com aproximação normal para o teste de homogeneidade está apresentado abaixo. Como dados de entrada, ele exige o parâmetro a da densidade priori, os dados observados e o percentual p para o cálculo do intervalo de credibilidade de $p\%$ para γ_1 , γ_2 e ρ .

Programa 3. Programa para o teste de homogeneidade

```

num.simul<-10000
conta<-0
param<- priori+obs
teta<-c(param[1:2]/sum(param[1:2]),param[3:4]/sum(param[3:4]))
mu.1<- digamma(param[1]) - digamma(param[2])
mu.2<- digamma(param[3]) - digamma(param[4])
sigma.1<-sqrt(trigamma(param[1])+trigamma(param[2]))
sigma.2<-sqrt(trigamma(param[3])+trigamma(param[4]))
sigma.r<-sqrt(2*trigamma(1))
dens.s <- function(v) {
  w1<-v[1]
  w2<-v[2]
  ((w1/sigma.r)^2) - (((w1-mu.1)/sigma.1)^2) + ((w2/sigma.r)^2) - (((w2-
mu.2)/sigma.2)^2)
}
f.otim <- function(x) {
  y<-log(x/(1-x))
  p<- c(y,y)

```

```

dens.s(p)
}
result<-optimize(f.otim, interval =c(0,1), maximum = TRUE)
dens.sn<- result$objective
set.seed=33
valor<-
matrix(c(rnorm(num.simul,mu.1,sigma.1),rnorm(num.simul,mu.2,sigma.2)),ncol=2)
for (j in 1:num.simul) {
  if(dens.s(valor[j,])<dens.sn) conta<-conta+1
}
cat('Full Bayesian Significance Test','\n')
cat('Test for homogeneity', "\n", "\n")
cat('Data Provided: ', "\n")
cat('Priori: ', "\n")
print(cbind(X=rbind(A=priori[1:2], B=priori[3:4], Total=
c(priori[1]+priori[3],priori[2]+priori[4])),Total=c(sum(priori[1:2]),sum(priori[3
:4]), sum(priori))))
cat('', "\n")
cat('Observed values:', "\n")
print(cbind(X=rbind(A=obs[1:2], B=obs[3:4], Total=
c(obs[1]+obs[3], obs[2]+obs[4])),Total=c(sum(obs[1:2]),sum(obs[3:4]), sum(obs))))
cat('', "\n")
cat('Posteriori Probabilities:', "\n")
print(cbind(X=rbind(A=round(teta[1:2],3), B=round(teta[3:4],3))))
cat('', "\n", "\n")
cat("Evidence = ", round(conta/num.simul,3), "\n")

```

1.4. Teste do Equilíbrio de Hardy – Weinberg

Neste teste o parâmetro agora é definido por $\gamma=(\gamma_1, \gamma_2, \gamma_3)$. Trabalhando com o parâmetro $\lambda = \gamma_1 + (\frac{1}{2})\gamma_2$, o objetivo é testar $H: \gamma_1 = \lambda^2$ e $\gamma_3 = (1-\lambda)^2$ contra a alternativa A definida pelo conjunto complementar da linha que define H em relação ao simplex: $\{(\gamma_1, \gamma_2, \gamma_3) > 0 \wedge \gamma_1 + \gamma_2 + \gamma_3 = 1\}$: (γ_1, γ_3) pertence ao quadrado unitário e tal que $\gamma_1 + \gamma_3 \leq 1$. A reparametrização do vetor de parâmetros original γ para ω nos fornece

$$\omega = \ln \begin{bmatrix} \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_3 \end{bmatrix}$$

e a hipótese de teste após a reparametrização é

$$H : \omega_1 = \ln \left[\frac{\lambda}{1-\lambda} \right].$$

Por termos um modelo trinomial, devemos fazer uma redução nas dimensões das matrizes médias e covariâncias da distribuição posteriori e da função de referência. Como essa atividade é apenas um exercício algébrico,

apresentaremos diretamente os resultados das matrizes para a densidade posteriori

$$\mu_{\omega} = E(\omega) = \begin{bmatrix} \psi(a_1) - \psi(a_3) \\ \psi(a_2) - \psi(a_3) \end{bmatrix} e$$

$$\Sigma_{\omega} = Var(\omega) = \begin{bmatrix} \psi'(a_1) + \psi'(a_3) & \psi'(a_3) \\ \psi'(a_3) & \psi'(a_1) + \psi'(a_3) \end{bmatrix},$$

e da matriz de covariâncias da função de referência $s(\omega)$ é

$$\Sigma_q = \begin{bmatrix} 2\psi'(1) & \psi'(1) \\ \psi'(1) & 2\psi'(1) \end{bmatrix}$$

A função surpresa $h(\omega)$ é dada por

$$h(\omega) = |\Sigma_{\omega}|^{-1/2} |\Sigma_q|^{1/2} \exp \left[\frac{1}{2} (\omega' \Sigma_q^{-1} \omega - (\omega - \mu_{\omega})' \Sigma_{\omega}^{-1} (\omega - \mu_{\omega})) \right].$$

A tabela abaixo apresenta o comparativo do *e-value* e do *p-value* considerando $x = (x_1, x_2, x_3) = (x_1, 20 - x_1 - x_3, x_3)$.

x_1	x_3	ev	pv	x_1	x_3	ev	pv
1	2	0,01	0,00	5	1	0,09	0,04
1	3	0,02	0,01	5	2	0,29	0,14
1	4	0,05	0,02	5	3	0,61	0,34
1	5	0,10	0,04	5	4	0,89	0,65
1	6	0,18	0,08	5	5	1,00	1,00
1	7	0,29	0,15	5	6	0,90	0,66
1	8	0,42	0,26	5	7	0,68	0,39
1	9	0,66	0,39	5	8	0,43	0,20
1	10	0,83	0,57	5	9	0,24	0,09
1	11	0,95	0,77	5	10	0,11	0,04
1	12	1,00	0,99	9	0	0,20	0,09
1	13	0,95	0,78	9	1	0,56	0,39
1	14	0,82	0,55	9	2	0,97	0,91
1	15	0,61	0,33	9	3	0,88	0,59
1	16	0,40	0,16	9	4	0,52	0,26
1	17	0,22	0,05	9	5	0,24	0,09
1	18	0,10	0,00	9	6	0,09	0,03
5	0	0,03	0,01	9	7	0,03	0,01

O programa em **R** que calcula o FBST com aproximação normal para o teste de equilíbrio de Hardy-Weinberg está apresentado abaixo. Como dados de entrada, ele exige o parâmetro a da densidade priori e os dados observados.

Programa 4. Programa para o teste de equilíbrio de Hardy - Weinberg

```

num.simul<-10000
conta<-0
param<- priori+obs
mu<- c(digamma(param[1]),digamma(param[2]))-digamma(param[3])
sigma<-diag(c(trigamma(param[1]),trigamma(param[2])))+trigamma(param[3])
sigma.r <-diag(c(trigamma(priori[1]),trigamma(priori[2])))+trigamma(priori[3])
dens.l <- function(v) {
  (t(v)%*%solve(sigma.r)%*%v) - (t(v - mu)%*%solve(sigma)%*%(v - mu))
}
f.otim <- function(x) {
  p<- log(c(x^2,2*x*(1-x))/(1-x)^2)
  dens.l(p)
}
result<-optimize(f.otim, interval=c(0, 1), maximum = TRUE)
dens.sn<- result$objective
set.seed=33
valor<-mvrnorm(num.simul, mu=mu, Sigma=sigma)
for (j in 1:num.simul) {
  if(dens.l(valor[j,])<dens.sn) conta<-conta+1
}
cat('Full Bayesian Significant Test',"\\n")
cat('Test for Hardy-Weinberg equilibrium', "\\n", "\\n")
cat('Priori and Observed Values provided: ', "\\n")
print(cbind(X=rbind(Priori =priori, Observed=obs),Total=c(sum(priori),sum(obs))))
cat('','\\n")
cat("Evidence = ", round(conta/num.simul,3), "\\n")

```

1.5. Teste de independência

Para este teste, definimos as funções $\theta_1+\theta_2=\lambda$ e $\theta_1+\theta_3=\sigma$, como ilustrado na tabela abaixo.

Categoria 2	Categoria 1		Total
	1	2	
1	θ_1	θ_2	λ
2	θ_3	θ_4	$(1-\lambda)$
Total	σ	$(1-\sigma)$	1

O nosso interesse é testar a hipótese

$$H: \theta_1 = \lambda\sigma \text{ e } \theta_4 = (1-\lambda)(1-\sigma).$$

que, após a reparametrização, é dada por

$$H : \omega = (\omega_1, \omega_2, \omega_3) = \ln \left[\frac{\lambda\sigma}{(1-\lambda)(1-\sigma)}; \frac{\lambda}{(1-\lambda)}; \frac{\sigma}{(1-\sigma)} \right]$$

Assim como no teste de homogeneidade, aqui novamente a hipótese de teste pode ser apresentada baseada no logaritmo da razão de chances

$$H : \rho = 0,$$

A tabela abaixo apresenta um comparativo entre *e-value* e o *p-value* para diferentes valores do vetor $x = (x_1, x_2, x_3, x_4)$. A comparação entre os resultados obtidos pelo programa com os resultados obtidos pelo FBST analisado no espaço paramétrico original não apresentou diferenças até a segunda casa decimal.

x_1	x_2	x_3	x_4	ev	p _v
12	6	95	35	0,96	0,57
48	25	9	10	0,54	0,14
96	50	18	20	0,24	0,04
18	5	39	30	0,31	0,06
36	10	78	60	0,08	0,01

O programa em R que calcula o FBST com aproximação normal para o teste de independência está apresentado abaixo. Como dados de entrada, ele exige o parâmetro a da densidade priori, os dados observados e o percentual p para o cálculo do intervalo de credibilidade de $p\%$ para $\gamma_1, \gamma_2, \lambda$ e ρ .

Programa 5. Programa para o teste de independência

```
num.simul<-10000
conta<-0
param<- priori+obs
teta<-param/sum(param)
mu<- c(digamma(param[1]),digamma(param[2]),digamma(param[3]))-digamma(param[4])
sigma<-
diag(c(trigamma(param[1]),trigamma(param[2]),trigamma(param[3])))+trigamma(param[
4])
sigma.r<-
diag(c(trigamma(priori[1]),trigamma(priori[2]),trigamma(priori[3])))+trigamma(pri
ori[4])
c<-c(1,-1,-1,1)
mu.cr<-t(digamma(param)) %*% c
sigma.cr<-sqrt(sum(trigamma(param)))

dens.l <- function(v) {
  det(sigma.r)/det(sigma)* exp(0.5*(t(v)%*%solve(sigma.r)%*%v) - (t(v -
mu)%*%solve(sigma)%*%(v - mu)))
}

f.otim <- function(x) {
  p<-log((1/((1-x[1])*(1-x[2])))* c(x[1]*x[2],x[1]*(1-x[2]),(1-x[1])*x[2]))
  - dens.l(p)
}
result<-optim(c(sum(teta[1:2]),teta[1]+teta[3]), f.otim)
dens.sn<- abs(result$value)

set.seed=33
```

```

valor<-mvrnorm(num.simul, mu=mu, Sigma=sigma)
for (j in 1:num.simul) {
  if(dens.l(valor[j,])<dens.sn) conta<-conta+1
}
cat('Full Bayesian Significance Test', "\n")
cat('Test for independency', "\n", "\n")
cat('Data Provided: ', "\n")
cat('Priori: ', "\n")
print(cbind(X=rbind(A=priori[1:2], B=priori[3:4], Total=
c(priori[1]+priori[3],priori[2]+priori[4])),Total=c(sum(priori[1:2]),sum(priori[3:4]), sum(priori))))
cat('', "\n")
cat('Observed values: ', "\n")
print(cbind(X=rbind(A=obs[1:2], B=obs[3:4], Total=
c(obs[1]+obs[3],obs[2]+obs[4])),Total=c(sum(obs[1:2]),sum(obs[3:4]), sum(obs))))
cat('', "\n")
cat('Posteriori Probabilities: ', "\n")
print(cbind(X=rbind(A=teta[1:2], B=teta[3:4], Total=
c(teta[1]+teta[3],teta[2]+teta[4])),Total=c(sum(teta[1:2]),sum(teta[3:4]),
sum(teta))))
cat('', "\n", "\n")
cat('Estimatives and Credible sets: ', "\n")
cat(" gama 1 = ", param[1]/sum(param[1:2]),
" [, qbeta(cred,param[1],param[2]), "; ", qbeta(1-cred,param[1],param[2]), " ] ", "\n")
cat(" gama 2 = ", param[3]/sum(param[3:4]),
" [, qbeta(cred,param[3],param[4]), "; ", qbeta(1-cred,param[3],param[4]), " ] ", "\n")
cat("lambda = ", sum(teta[1:2]),
" [, qbeta(cred,sum(param[1:2]),sum(param[3:4])), "; ", qbeta(1-cred,sum(param[1:2]),sum(param[3:4])), " ] ", "\n")
cat(" ratio = ", log(teta[1]*teta[4]/(teta[2]*teta[3])),
" [, qnorm(cred,mu.cr,sigma.cr), "; ", qnorm(1-cred,mu.cr,sigma.cr), " ] ", "\n", "\n")
cat('', "\n")
cat("Evidence = ", conta/num.simul, "\n")

```

Capítulo 5

Aplicação em tabelas de contingência 2x2x2

Neste capítulo faremos uma extensão dos programas apresentados para tabelas de contingência 2x2, com a aplicação num conjunto de dados reais oriundos de um projeto analisado pelo Centro de Estatística Aplicada – CEA – USP em 2005.

O projeto tinha como um dos objetivos determinar qual tipo de parasita entre carrapatos, ácaros e malófagos apresentava a maior prevalência na população das aves de cerrado e mata de galeria. Para cada ave capturada na amostra, anotava-se a observação de um vetor (c, a, m) cujos elementos assumiram os valores 0 ou 1 de acordo com a ausência ou presença de carrapatos (c), ácaros (a) e malófagos (m).

Considerações iniciais devem ser observadas para que o modelo a ser utilizado possa ser validado. A primeira delas é a de que toda ave do ambiente tenha a mesma chance de ser capturada. A segunda é a de que não exista preferência de espécies no que diz respeito ao ataque dos diversos tipos de parasita.

A última, e talvez a mais rigorosa, é a de que a captura das aves foi feita de forma independente. Isto é, o fato de um tipo de ave ser selecionado não aumenta ou diminui a chance de uma outra ave, de mesma espécie, ser selecionada. Em resumo, a coleta é feita segundo um modelo de seleção independente e identicamente distribuída.

A notação utilizada é:

1. As freqüências amostrais são denotadas por x_{cam} . Isto é, o número de aves com vetor de incidência (c,a,m). Por exemplo, y_{011} é o número de unidades amostrais com ausência de carrapatos e presença de ácaro e malófagos. A tabela 1 descreve todas as freqüências observadas. O vetor das observações será então $X = (x_{000}, x_{001}, x_{010}, x_{100}, x_{011}, x_{101}, x_{110}, x_{111})$.
2. As proporções populacionais de aves em cada uma das 8 possíveis categorias, (c,a,m), são denotadas por θ_{cam} . Assim, o vetor populacional das proporções nas respectivas categorias é representado por

$$\Theta = (\theta_{000}, \theta_{001}, \theta_{010}, \theta_{100}, \theta_{011}, \theta_{101}, \theta_{110}, \theta_{111}).$$

Tabela 1. Freqüências observadas

C	A	M	Θ	X	Freqüência
0	0	0	θ_{000}	x_{000}	208
0	0	1	θ_{100}	x_{100}	79
0	1	0	θ_{010}	x_{010}	276
1	0	0	θ_{001}	x_{001}	91
0	1	1	θ_{110}	x_{110}	140
1	0	1	θ_{011}	x_{011}	15
1	1	0	θ_{101}	x_{101}	108
1	1	1	θ_{111}	x_{111}	36

Consideraremos a densidade a priori como uma Dirichlet de ordem 8 e com vetor de parâmetros $\mathbf{a}$ formado por 1 nas oito posições. Com isso temos que a densidade posteriori será uma Dirichlet com vetor $\mathbf{A}$, onde $\mathbf{A} = \mathbf{a} + \mathbf{X}$.

Baseado no objetivo do estudo, os parâmetros de interesse são as proporções de aves que apresentam cada uma das parasitas. Isto é, os parâmetros a serem comparados são os seguintes:

$$\theta_{1..} = \theta_{111} + \theta_{110} + \theta_{101} + \theta_{100} ,$$

$$\theta_{.1.} = \theta_{111} + \theta_{110} + \theta_{011} + \theta_{010} e$$

$$\theta_{..1} = \theta_{111} + \theta_{011} + \theta_{101} + \theta_{001} .$$

As hipóteses de teste são

Hipótese H: todos os tipos de parasitas têm a mesma prevalência;

Hipótese H_{CA} : Carrapatos e Ácaros possuem a mesma prevalência;

Hipótese H_{AM} : Ácaros e Malófagos possuem a mesma prevalência e

Hipótese H_{CM} : Carrapatos e Malófagos possuem a mesma prevalência.

Com os parâmetros definidos acima, as hipóteses são

$$H : \theta_{1..} = \theta_{.1.} = \theta_{..1}$$

$$H_{CA} : \theta_{1..} = \theta_{.1.} \Leftrightarrow H_{CA} : \theta_{101} + \theta_{100} = \theta_{011} + \theta_{010},$$

$$H_{AM} : \theta_{..1} = \theta_{.1.} \Leftrightarrow H_{AM} : \theta_{101} + \theta_{001} = \theta_{110} + \theta_{010},$$

$$H_{CM} : \theta_{..1} = \theta_{.1.} \Leftrightarrow H_{CM} : \theta_{011} + \theta_{001} = \theta_{110} + \theta_{100}.$$

A simulação foi realizada para cada um dos testes de hipótese e os resultados da evidência apontam que as prevalências de cada tipo de parasitas são diferentes.

Hipótese	Ev
H	<0,001%
H_{CA}	<0,001%
H_{AM}	<0,001%
H_{CM}	<0,001%

Analisando-se as freqüências observadas, podemos afirmar que a ordem da maior prevalência para a menor é a seguinte: Ácaros, Malófagos e Carrapatos.

Referências Bibliográficas

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

Lindley, D.V. (1964). **The Bayesian Analysis of Contingency Tables.** *The Annals of Mathematical Statistics*, vol.35, n 4, 1622-1643.

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

Pereira, C.A.B., Lindley, D. (1987), **Examples Questioning The Use Of Partial Likelihood.** *The Statistician*, 36, p. 15-20.

Pereira, C.A.B., Stern, J.M., (1999), **Evidence and credibility: full Bayesian significance test for precise hypotheses.** *Entropy Journal*, 1, 69-80.

Pereira, C.A.B., Stern, J.M., Wechsler, S. (2005). **Can a significance test be genuinely Bayesian?** *Technical Report*, MAE 05/15.

Venables, W.N., Ripley, B.D., (1999). **Modern Applied Statistics with S-Plus, Third Edition.** New York, Springer.