ANÁLISE DE DADOS CATEGÓRICOS USANDO

TEORIA DA INFORMAÇÃO

MARIA DAS MERCÊS SANTANA NUNESMAIA

DISSERTAÇÃO APRESENTADA AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA A OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR:

Prof.Dr. CLÓVIS DE ARAÚJO PERES

- SÃO PAULO, JULHO DE 1977 -

Aos meus pais,

Olegário e Iraci

Ao

Gil

AGRADECIMENTOS

Este trabalho é resultado da ajuda recebida de muitos, de diversas formas e nas diversas etapas de sua elaboração. Queremos a todos externar os nossos agradecimentos, em especial

- Ao Prof. Clóvis de Araújo Peres, pela sugestão do tema, pela dedicação na orientação deste trabalho, e por nos ter proporcionado uma formação básica com a riqueza de seu convívio humano e científico.
- À Profa. Lolita Carneiro Dantas, pelo seu apoio constante e pela força do seu estímulo sem os quais não seria possível a realização deste trabalho.
- Ao Prof. José Figueiredo Leal de Araújo, Diretor do Instituto de Matemática da Universidade Federal da Bahia, pela compreensão demonstrada durante o nosso Curso de Pós-Graduação.
- Ao Prof. Wilton de Oliveira Bussab pelas sugestões apresentadas durante a elaboração deste trabalho.
- Aos colegas do Curso de Pós-Graduação do Instituto de Matemática e Estatística da Universidade de São Paulo, pela convivência amiga.
- Aos colegas dos Departamentos de Matemática e de Estatística do Instituto de Matemática da Universidade Federal da Bahia.
- Aos Professores do Departamento de Estatística do Instituto de Matemática e Estatística da Universidade de São Pau

lo.

- Ao sr. João Baptista Esteves de Oliveira pela eficiência de seu trabalho datilográfico.

- 0 -

ÍNDICE

PREFÁCIO 1

CAPÍTULO I - INTRODUÇÃO. 3

CAPÍTULO II - INFORMAÇÃO ESTATÍSTICA 6

2.1 - Informação: Definição e Propriedades. 6

2.2 - Informação Discriminante mínima 12

2.3 - Informação Estatística. 17

2.4 - Testes de Hipóteses e Distribuição Estatísticas Associadas. 20

CAPÍTULO III - DISTRIBUIÇÃO MULTINOMIAL. 25

3.1 - Distribuição Multinomial. 25

3.2 - Testes de Hipóteses Simples 29

3.3 - Teste de Homogeneidade. 39

CAPÍTULO IV - TABELAS DE CONTINGÊNCIA COMPLETAS. 47

4.1 - Tabela de Contingência: conceituação. 47

4.2 - Tabela de Contingência de Dupla Entrada: Teste de Independência 48

4.3 - Tabela de Contingência de Três Entradas: Teste de Independência 55

4.4 - Homogeneidade Entre Tabelas de Contingência de Dupla Entrada 62

CAPÍTULO V - TABELAS DE CONTINGÊNCIA INCOMPLETAS 68

5.1 - Definições Preliminares 68

5.2 - Análise de Algumas Tabelas de Contingência Incompletas 73

5.3 - Interação 87

BIBLIOGRAFIA 92

B

PREFÁCIO

Neste trabalho procuramos estudar de modo resumido as novas tendências da literatura sobre a Análise de Tabelas de Contingência.

No capítulo I apresentamos uma nota histórica sobre Análise de Dados de Classificação Cruzada com a finalidade de situar o nosso objeto de estudo. O conteúdo do capítulo II é constituído por uma revisão sobre Teoria da Informação, procurando destacar o uso desta teoria em teste de hipóteses e o seu relacionamento com o teste da Razão de Verossimilhança. No capítulo III estudamos populações com distribuição multinomial em termos da Teoria da Informação. A-analisamos tabelas de contingência completas, utilizando a Teoria da Informação no capítulo IV. Finalmente no capítulo V analisamos Tabelas de Contingência Incompletas, utilizando a estatística da Razão de Verossimilhança, por apresentar algumas vantagens sobre outras estatísticas existentes na literatura e pelo seu relacionamento com a Teoria da Informação.

Lembramos que nos capítulos II, III e IV seguimos a notação utilizada por Kullback (1968) e no capítulo V a no

tação utilizada por Leo Goodman (1968), com o objetivo de facilitar possíveis consultas das referências.

- ¤ -

CAPÍTULO I

INTRODUÇÃO

A Análise de Dados de Classificação Cruzada foi amplamente estudada por Fisher no período compreendido entre 1922 e 1962, principalmente em problemas ligados à áreas biológicas. Vários autores apresentaram contribuições isoladas, em diversos aspectos, à Análise de Dados de Classificação Cruzada. Uma teoria geral, usando Teoria da Informação, que engloba de uma certa maneira o que foi visto antes, foi formalizada em 1967 por Kullback em seu livro "Information Theory and Statistics". Antes disto algumas medidas de informação eram conhecidas. Fisher em 1925 introduziu em seu trabalho sobre Teoria de Estimação a definição técnica de informação. Shannon e Wiener em 1948 publicaram, independentemente, trabalhos sobre medidas logarítmicas de informação para aplicação em Teoria da Comunicação. De acordo com os objetivos de nosso trabalho iremos tratar da Teoria da Informação como um ramo da Teoria da Probabilidade e da Estatística Matemática que pode ser aplicado nos mais diversos campos da ciência. Esta teoria além de ter se tornado básica para

a Estatítica Matemática é essencial na Inferência Estatística.

A Análise de Dados de Classificação Cruzada apesar de amplamente estudada, adquiriu uma nova dimensão, mostrando-se como objeto muito rico para a investigação nos últimos anos, primeiro devido a utilização da Teoria da Informação na Análise das Tabelas de Contigência Completa e segundo devido ao conceito de quase-independência e aos métodos desenvolvidos por Leo Goodman em 1968 para analisar Tabelas de Contigência Incompletas.

Segundo Goodman (1968) as Tabelas de Contigência Incompletas (ou Truncadas) são aquelas em que as frequências em algumas das caselas da Tabela são omitidas da análise por terem sido perdidas, vazias, irrealizáveis ou restritas de um certo modo, enquanto que as Tabelas de Contigência Completas são aquelas em que nenhuma das caselas da Tabela é omitida da análise.

O estudo das Tabelas de Contigência Incompletas pode parecer à primeira vista um tópico de interesse limitado mas elas podem ser utilizadas numa variedade muito grande de situações reais e mais ainda, a sua análise fornece valiosa ajuda à análise das Tabelas de Contigência Completas.

Numerosos trabalhos têm sido publicados (Bishop e Fienberg, 1969; Mantel, 1970; Fienberg, 1970; Hocking e Oxspring, 1971; Grizzle e Willians, 1972; Chen e Fienberg,1974 e 1976, etc.) focalizando aspectos de casos particulares das Tabelas de Contigência Incompletas. No nosso estudo daremos ênfase às abordagens oferecidas por Kullback (1968) para a Análise de Tabelas de Contigência Completas e por Goodman (1968) para a Análise de Tabelas de Contigência Incompletas.

Considerando a importância prática do tema, o objetivo do nosso trabalho é apresentá-lo de modo didático e atualizado.

- ¤ -

CAPÍTULO II

INFORMAÇÃO ESTATISTÍCA

Com o objetivo de facilitar o entendimento de capítulos subsequentes, descreveremos sucintamente os principais aspectos da "Teoria da Informação Estatística" conforme foi desenvolvida por Kullback (1968). Inicialmente apresentaremos noções básicas de informação.

2.1 - INFORMAÇÃO: DEFINIÇÃO E PROPRIEDADES

Seja X uma variável aleatória com função de densidade de probabilidade dada por $f_1(x)$, quando X assume valores na população P_1 e $f_2(x)$, quando X assume valores na população P_2.

Seja H_i, $i=1,2$ a hipótese de que X tenha função de densidade de probabilidade $f_i(x)$, $i=1,2$ e seja x um valor especificado de X. Então segue do Teorema de Bayes que

$$(2.1.1) \qquad P(H_i|x) = \frac{P(H_i)f_i(x)}{P(H_1)f_1(x)+P(H_2)f_2(x)}, \qquad i=1,2.$$

Se todas as probabilidades definidas em (2.1.1) forem não nulas, obtemos:

$$(2.1.2) \qquad \frac{P(H_1|x)}{P(H_2|x)} = \frac{P(H_1)f_1(x)}{P(H_2)f_2(x)},$$

onde $P(H_i)$ e $P(H_i|x)$ são probabilidades "a priori" e "a posteriori" de H_i, respectivamente.

Aplicando logaritmo neperiano a ambos os membros de (2.1.2) temos:

$$(2.1.3) \qquad \log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1|x)}{P(H_2|x)} - \log \frac{P(H_1)}{P(H_2)}.$$

A diferença expressa no segundo membro de (2.1.3) pode ser considerada como a informação resultante da observação $X = x$.

<u>DEFINIÇÃO 1</u> - O logaritmo da razão de verossimilhança,

$$\log \frac{f_1(x)}{f_2(x)},$$

é a informação contida em $X = x$ para discriminar H_1 de H_2, quando H_2 é a hipótese nula e representaremos por $I(1:2;x)$.

<u>EXEMPLO 1</u> - Seja

$$f_i(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2}},$$

$i=1,2$. Neste caso as hipóteses consideradas acima são equi-

valentes a:

$$\begin{cases} H_2: \mu = \mu_1 \\ H_1: \mu = \mu_2. \end{cases}$$

Temos:

$$I(1:2;x) = \log \frac{f_1(x)}{f_2(x)} = \frac{1}{2}[(x-\mu_2)^2-(x-\mu_1^2)] = x(\mu_1-\mu_2)+\frac{1}{2}(\mu_2^2-\mu_1^2).$$

<u>EXEMPLO 2</u> - Seja

$$f_i(x) = \binom{n}{x} p_i^x q_i^{n-x},$$

com $q_i=1-p_i$, $i=1,2$. Neste caso temos:

$$I(1:2;x) = \log \frac{p_1^x q_1^{n-x}}{p_2^x q_2^{n-x}} = x \log \frac{p_1}{p_2} + (n-x)\log \frac{q_1}{q_2}.$$

Quando consideramos todos os valores que a variável aleatória X pode assumir definimos informação média por observação do seguinte modo:

<u>DEFINIÇÃO 2</u> - Informação média por observação para determinar H_1 de H_2 (hipótese nula) é dada por:

$$I(1:2;X) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)}\, dx$$

para variáveis aleatórias contínuas.

$$I(1:2;X) = \sum_{i=1}^{n} f_1(x_i)\log \frac{f_1(x_i)}{f_2(x_i)}$$

para variáveis aleatórias discretas.

OBSERVAÇÃO - Adotaremos a notação I(1:2) no lugar de I(1:2;X) quando definimos apenas uma variável aleatória.

EXEMPLO 3 - Considere a situação do exemplo 1. Então

$$I(1:2) = \int f_1(x)\log \frac{f_1(x)}{f_2(x)}\, dx =$$

$$= (\mu_1-\mu_2) \int x f_1(x)dx + \frac{1}{2}(\mu_2^2-\mu_1^2) \int f_1(x)dx.$$

Portanto

$$I(1:2) = (\mu_1-\mu_2)\mu_1 + \frac{1}{2}(\mu_2^2-\mu_1^2) = \frac{1}{2}(\mu_1-\mu_2)^2 =$$

Observemos que a informação média por observação é maior quanto maior for a diferença entre μ_1 e μ_2 e é igual a zero se $\mu_1 = \mu_2$.

EXEMPLO 4 - Considere o exemplo 2. Então

$$I(1:2) = \sum_{x=0}^{n} f_1(x)\log \frac{f_1(x)}{f_2(x)} = \sum_{x=0}^{n} f_1(x)(x \log \frac{p_1}{p_2} + (n-x)\log \frac{q_1}{q_2}) =$$

$$= \log \frac{p_1}{p_2} \sum_{x=0}^{n} x f_1(x) + \log \frac{q_1}{q_2} \sum_{x=0}^{n} (n-x) f_1(x).$$

Portanto

$$I(1:2) = n(p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2}).$$

Observemos que $I(1:2)$ é maior quanto maior for a diferença entre p_1 e p_2 e é igual a zero se $p_1 = p_2$.

PROPRIEDADES:

P.1) $I(1:2) \geq 0$, sendo $I(1:2) = 0$ se e somente se $f_1(x) = f_2(x)$.

PROVA

Verifica-se por meio de estudo de funções que

$$\log U > 1 - \frac{1}{U}, \quad U > 0 \text{ e } U \neq 1.$$

Usando este fato e fazendo

$$U = \frac{f_1(x)}{f_2(x)}$$

temos que

$$I(1:2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \, dx > \int f_1(x) \left(1 - \frac{f_2(x)}{f_1(x)}\right) dx = 0.$$

Portanto

$$I(1:2) > 0 \text{ para } \frac{f_1(x)}{f_2(x)} \neq 1 \text{ e}$$

$$I(1:2) = 0 \text{ se e somente se } \frac{f_1(x)}{f_2(x)} = 1.$$

P.2) $I(1:2;X,Y) = I(1:2;X) + I(1:2;Y)$, onde X e Y são variáveis aleatórias independentes.

A verificação desta propriedade é imediata desde que

$$I(1:2;X,Y) = \iint f_1(x,y)\log \frac{f_1(x,y)}{f_2(x,y)}\,dxdy.$$

Em decorrência desta propriedade temos que a informação obtida de n variáveis aleatórias independentes e identicamente distribuídas é dada por:

$$I(1:2;X_1,X_2,\ldots,X_n) = nI(1:2;X_1).$$

Este resultado será útil nos capítulos subsequentes.

P.3)
$$\begin{aligned} I(1:2;X,Y) &= I(1:2;X) + I(1:2;Y|X) = \\ &= I(1:2;Y) + I(1:2;X|Y) \end{aligned}$$

onde X e Y são variáveis aleatórias quaisquer.

Vejamos,

$$\begin{aligned} I(1:2;X,Y) &= \iint f_1(x,y)\log \frac{f_1(x,y)}{f_2(x,y)}\,dxdy = \\ &= \int g_1(x)\log \frac{g_1(x)}{g_2(x)}\,dx + \int g_1(x)\left[h_1(y|x)\log \frac{h_1(y|x)}{h_2(y|x)}\,dy\right]dx, \end{aligned}$$

onde

$$g_i(x) = \int f_i(x,y)dy, \quad h_i(y|x) = \frac{f_i(x,y)}{g_i(x)}, \quad i=1,2.$$

Se fizermos

$$\int h_1(y|x)\log \frac{h_1(y|x)}{h_2(y|x)}\, dy = I(1:2,Y|X=x)$$

e

$$\int g_1(x)I(1:2,Y\ X=x)dx = I(1:2;Y|X)$$

obteremos o resultado em P.3, como queríamos verificar.

2.2 - INFORMAÇÃO DISCRIMINANTE MÍNIMA

Sejam, $f_2(x)$ um elemento genérico de um conjunto H_2 de funções densidades de probabilidades e $f_1(x)$ um elemento do conjunto H_1 de funções densidades de probabilidades, definido por: $H_1 = \{f_1 | \int T(x)f_1(x)dx = \theta\}$, onde $T(x)$ é uma Estatística e θ é um parâmetro populacional. Nestas condições, para uma dada $f_2(x)$ do conjunto H_2, consideremos o problema de determinar $f_1(x)$ de H_1 tal que minimize a informação

$$I(1:2) = \int f_1(x)\log \frac{f_1(x)}{f_2(x)}\, dx$$

sujeita a restrição

$$\int T(x) f_1(x) dx = \theta.$$

Usando multiplicadores de Lagrange, o problema é resolvido minimizando a expressão.

$$(2.2.1) \qquad \int (f_1(x) \log \frac{f_1(x)}{f_2(x)} + KT(x) f_1(x) + Lf_1(x)) dx$$

onde K e L são constantes.

O valor mínimo de (2.2.1) ocorre quando

$$(2.2.2) \qquad f(x) = f^*(x) = f_2(x) e^{-KT(x)-L-1}.$$

Por conveniência computacional faremos $-K = \tau$. Então as constantes L e τ são determinadas por meio das equações de restrição, isto é,

$$e^{-L-1} \log M_2(\tau) = 1 \Longleftrightarrow \int f^*(x) dx = 1$$

$$\frac{d}{d\tau} \log M_2(\tau) = \theta \Longleftrightarrow \int T(x) f^*(x) dx = \theta$$

onde

$$M_2(\tau) = \int f_2(x) e^{\tau T(x)} dx.$$

Nestas condições o valor mínimo de I(1:2) é dado por:

(2.2.3) $$I(^*:2) = \theta\tau - \log M_2(\tau).$$

A função $f^*(x)$ em (2.2.2) é chamada *distribuição conjugada de* $f_2(x)$ e pode ser representada em termos de $M_2(\tau)$ da seguinte forma:

(2.2.4) $$f^*(x) = \frac{f_2(x)e^{\tau T(x)}}{M_2(\tau)}.$$

OBSERVAÇÕES:

1) $f^*(x) = \dfrac{f_2(x)e^{\tau T(x)}}{M_2(\tau)}$ constitui uma uma família exponencial.

2) A estatística $Y = T(x)$ é uma estatística suficiente para famílias do tipo $f^*(x)$.

3) $I(^*:2) \geq 0$ e $I(^*:2) = 0$ se e somente se $\tau = 0$ para todo θ.

4) $I(^*:2)$ é monotonicamente crescente para $\theta \geq \theta_0$, onde θ_0 é determinado por:

$$\left[\frac{d}{d\tau} \log M_2(\tau)\right]_{\tau=0} = \theta_0.$$

<u>EXEMPLO 5</u> - Consideremos a função densidade de probabilidade $f_2(x)$ dada por

$$f_2(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2}}.$$

Seja H_1 um conjunto de funções densidades de probabilidades

$f_1(x)$, tal que $\int T(x) f_1(x) dx = \theta$, onde θ é um parâmetro populacional e $T(x)$ é uma estatística.

A fim de obtermos a informação discriminante mínima dada por (2.2.4) precisamos calcular τ e $M_2(\tau)$. Para isto consideremos a situação em que $\theta = E(x)$ e portanto $T(x) = x$.

Então temos:

$$M_2(\tau) = \int f_2(x) e^{\tau T(x)} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2}} e^{\tau x} dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x^2-2\mu_2 x+\mu_2^2-2\tau x)}{2}} dx =$$

$$= e^{\frac{2\mu_2\tau+\tau^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{[x-(\mu_2+\tau)]^2}{2}} dx = e^{\mu_2\tau+\tau^2/2}.$$

Por outro lado sabemos que

$$\theta = \frac{\partial}{\partial\tau}[\log M_2(\tau)].$$

Portanto

$$\theta = \mu_2 + \tau$$

ou seja $$\tau = \theta - \mu_2.$$

Substituindo os valores de $M_2(\tau)$ e τ em (2.2.3), temos

$$(2.2.5) \qquad I(*:2) = (\theta-\mu_2)\theta - (\mu_2\tau + \frac{\tau^2}{2}) = \frac{1}{2}(\theta-\mu_2)^2.$$

EXEMPLO 6 - Consideremos a função densidade de probabilidade, $f_2(x)$, dada por

$$f_2(x) = \binom{n}{x} p_2^x q_2^{n-x}, \quad q_2 = 1-p_2,$$

como no exemplo 5 vamos admitir que $\theta = E(x)$ e portanto $T(x) = x$. Neste caso teremos:

$$M_2(\tau) = \sum f_2(x) e^{\tau x} = \sum_{x=0}^{n} \binom{n}{x} p_2^x q_2^{n-x} e^{\tau x}$$

$$= \sum_{x=0}^{n} (p_2 e^{\tau})^x q_2^{n-x} = (p_2 e^{\tau} + q_2)^n$$

e

$$\theta = \frac{\partial}{\partial \tau} \log M_2(\tau) = \frac{n p_2 e^{\tau}}{p_2 e^{\tau} + q_2}$$

ou seja

$$e^{\tau} = \frac{\theta q_2}{p_2(n-\theta)}$$

donde

$$\tau = \frac{\theta q_2}{p_2(n-\theta)}.$$

Substituindo $M_2(\tau)$ e τ em (2.2.3) temos:

$$I(^*:2) = \theta \log \frac{\theta q_2}{p_2(n-\theta)} - \log\left(p_2 \frac{\theta q_2}{p_2(n-\theta)} + q_2\right)^n$$

isto é,

$$(2.2.6) \qquad I(^*:2) = \theta \log \frac{\theta}{n-\theta} + \theta \log \frac{q_2}{p_2} - n \log \frac{nq_2}{n-\theta}.$$

Como $\theta = np_1$ temos de (2.2.6) que

$$(2.2.7) \qquad I(^*:2) = n[p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2}].$$

2.3 - INFORMAÇÃO ESTATÍSTICA

Na seção anterior introduzimos o conceito de informação discriminante mínima, $I(^*:2)$, como o valor mínimo de

$$I(1:2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx,$$

para uma dada $f_2(x)$ e para toda $f_1(x)$ tal que

$$\theta = \int T(x) f_1(x) dx.$$

Vimos que o valor mínimo $I(^*:2) = \theta\tau - \log M_2(\tau)$ ocorre quando

$$f_1(x) = f^*(x) = \frac{e^{\tau T(x)} f_2(x)}{M_2(\tau)}, \quad M_2(\tau) = \int f_2(x) e^{\tau T(x)} dx \text{ e}$$

$$\theta = \frac{d}{d\tau} \log M_2(\tau).$$

Seja O_n uma amostra de n observações independentes com função densidade de probabilidade $f_2(x)$. Com base nesta amostra seja $\hat{\theta} = T(x)$ um estimador não viesado do parâmetro θ, tal que

$$\theta = \int T(x) f^*(x) dx,$$

onde $f^*(x)$ é definida por (2.2.4).

<u>DEFINIÇÃO 1</u> - Denomina-se Informação Estatística à estimativa de $I(^*:2)$ obtida substituindo em sua expressão, θ por $\hat{\theta}$. Portanto se a denotarmos por $\hat{I}(^*:2;O_n)$ podemos escrever:

$$\hat{I}(^*:2,O_n) = \hat{\theta}(x)\hat{\tau}(x) - \log M_2(\hat{\tau}(x)),$$

onde $\hat{\tau}(x)$ é determinado por meio da relação

$$(2.3.1) \qquad T(x) = \hat{\theta}(x) = \left[\frac{d}{d\tau} \log M_2(\tau)\right]_{\tau=\hat{\tau}(x)}.$$

Nos casos em que diferentes funções das observações são estimadores não viesados de θ, usaremos aquela que produzir o maior valor de $\hat{I}(^*:2;O_n)$.

Notamos que $\hat{I}(^*:2,O_n)$ é uma medida da discriminação entre a amostra e $f_2(x)$ de tal modo que: quanto maior for o valor de $\hat{I}:(^*:2;O_n)$ maior será a divergência entre a amostra e $f_2(x)$.

A Informação Estatística, $\hat{I}(^*:2;O_n)$, foi determinada considerando uma dada função densidade de probabilidade $f_2(x)$. Em aplicações é comum termos um conjunto H_2 de funções densidade de probabilidades f_2 e não uma única. Neste caso temos

DEFINIÇÃO 2 - Denomina-se Informação Estatística mínima, $\hat{I}(^*:H_2;O_n)$, ao valor mínimo de $\hat{I}(^*:2;O_n)$ quando $f_2(x)$ variar em H_2, isto é,

$$\hat{I}(^*:H_2;O_n) = \min\{\hat{I}(^*:2;O_n) \mid f_2(x) \in H_2\}.$$

EXEMPLO 7 - Consideremos uma amostra, O_n, de n observações independentes de uma população com função densidade de probabilidade definida como no exemplo 5 da seção anterior. Então a Informação Estatística é dada por

$$\hat{I}(^*:2;O_n) = n\hat{I}(^*:2),$$

onde $\hat{I}(^*:2)$ é dada por (2.2.5) com θ substituído por $\hat{\theta} = \bar{x}$, isto é,

$$\hat{I}(^*:2;O_n) = \frac{n}{2}(\bar{x}-\mu_2)^2.$$

EXEMPLO 8 - Consideremos o caso em que temos uma amostra, O_n, de n observações independentes de uma população com função densidade de probabilidade $f_2(x)$ definida por:

$$f_2(x) = p_2^x q_2^{1-x}, \quad q_2 = 1-p_2; \quad x = 0,1.$$

Então a Informação Estatística é obtida quando substituímos em (2.2.6) da seção anterior θ por $\hat{\theta} = n\hat{p}_1 = x$, isto é,

$$\hat{I}(^*:2;O_n) = x \log \frac{x}{np_2} + (n-x)\log \frac{n-x}{q_2}$$

Podemos observar que $\hat{I}(^*:2;O_n)$ nos dois exemplos é uma medida da discrepância entre a amostra e $f_2(x)$.

2.4 - TESTES DE HIPÓTESES E DISTRIBUIÇÃO DAS ESTATÍSTICAS ASSOCIADAS

As medidas $\hat{I}(^*:2;O_n)$ e $\hat{I}(^*:H_2;O_n)$ estão associadas à construção de testes de hipóteses da maneira especificada abaixo:

TABELA 2.1

TIPO DE HIPÓTESE	MEDIDA DE INFORMAÇÃO ASSOCIADA
1 $\begin{cases} H_2: f \in H_2 \\ H_1: f \in H_1 \end{cases}$	$\hat{I}(^*:H_2;O_n) - \hat{I}(^*:H_1;O_n)$
2 $\begin{cases} H_2: f = f_2 \\ H_1: f = f_1 \end{cases}$	$\hat{I}(^*:2;O_n) - \hat{I}(^*:1;O_n)$
3 $\begin{cases} H_2: f \in H_2 \\ H_1: f \in (H-H_2) \end{cases}$	$\hat{I}(^*:H_2;O_n)$
4 $\begin{cases} H_2: f = f_2 \\ H_1: f \neq f_2 \end{cases}$	$\hat{I}(^*:2;O_n)$

Verificamos facilmente que as situações 2, 3 e 4 são casos particulares da situação 1. Em cada situação a Estatística usada para os testes são funções convinientes das especificados acima. Estas funções são obtidas com o objetivo de se conseguir estatísticas com distribuições conhecidas.

A seguir iremos estudar as distribuições destas estatísticas.

Inicialmente verificaremos que $\hat{\tau}(x)$ determinado pela equação (2.3.1) é o estimador de máxima verossimilhança de τ considerado como parâmetro de

$$f^* = \frac{f_2(x)e^{\tau T(x)}}{M_2(\tau)}$$

De fato,

$$\log f^*(x) = \log f_2(x)+\tau T(x)-\log M_2(\tau)$$

e

$$\frac{d}{d\tau} \log f^*(x) = T(x) - \frac{d}{d\tau} \log M_2(\tau).$$

Portanto o estimador de máxima verossimilhança de τ é determinado pela equação $T(x) = \frac{d}{d\tau} \log M_2(\tau)$, a qual coincide com a equação (2.3.1).

Com base no resultado acima verifica-se facilmente que as medidas de informação especificadas na tabela 2.1 estão relacionadas com o teste da razão de verossimilhança para testar as respectivas hipóteses (Kullback, 1968) do seguinte modo:

1. $$\hat{I}(^*:H_2;O_n)-\hat{I}(^*:H_1;O_n) = -\log \frac{\max\limits_{f\in H_2} f(x)}{\max\limits_{f\in H_1} f(x)} = -\log \lambda^*$$

onde λ^* é a Razão de Verossimilhança estudada por Chernoff (1954).

2. $$\hat{I}(^*:2;O_n)-\hat{I}(^*:1;O_n) = -\log \frac{f_2(x)}{f_1(x)} = \log \frac{f_1(x)}{f_2(x)}$$

onde $\frac{f_1(x)}{f_2(x)}$ é a Razão de Verossimilhança de Neyman-Pearson para a construção de testes UMP.

3. $$\hat{I}(^*:H_1;O_n) = -\log \frac{\max\limits_{f\in H_2} f(x)}{\max\limits_{f\in H} f(x)} = -\log \lambda,$$

onde λ é a Razão de Verossimilhança de Neyman-Pearson.

4. $$\hat{I}(^*:2;O_n) = -\log \frac{f_2(x)}{\max\limits_{f\in H} f(x)} = -\log \lambda.$$

Com relação a λ sabemos, da literatura, que em condições gerais de regularidades $-2 \log \lambda$ tem distribuição limite, sob H_2, de χ^2_ϕ onde ϕ é a diferença entre a dimensão

do espaço paramétrico sob o modelo geral H e a dimensão do espaço paramétrico sob a hipótese H_2.

Com relação a λ^* Chernoff (1954) mostrou que $-2\log\lambda^*$ algumas vezes tem distribuição limite de χ^2 e outras vezes tem distribuição limite degenerada.

<u>EXEMPLO 9</u> - Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu, 1)$.

Nosso objetivo é testar a seguinte hipótese:

$$\begin{cases} H_2: \mu = \mu_2 \\ H_1: \mu \neq \mu_2 \end{cases}$$

Neste caso a estatística para o teste é $2\hat{I}(^*:2;0_n)$.

Verificamos na seção anterior que

$$\hat{I}(^*:2;0_n) = \frac{n}{2}(\bar{x}-\mu_2)^2$$

e portanto $2\hat{I}(^*:2;0_n) = n(\bar{x}-\mu_2)^2$ a qual tem distribuição de χ_1^2.

Rejeita-se H_2 para valores grandes de $2\hat{I}(^*:2;0_n)$.

<u>EXEMPLO 10</u> - Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu;1)$.

Consideremos a hipótese:

$$\begin{cases} H_2: \ \mu > \mu_2 \\ H_1: \ \mu \leq \mu_2. \end{cases}$$

Neste caso a estatística para o teste é

$$2\hat{I}(^*:H_2;O_n) = 2 \min\{\hat{I}(^*:2;O_n) \mid \mu \leq \mu_2\}.$$

Vimos no exemplo 7 da seção anterior que

$$\hat{I}(^*:2;O_n) = \frac{n}{2}(\bar{x}-\mu)^2,$$

então

$$2\hat{I}(^*:H_2;O_n) = 2 \min\{\frac{n}{2}(\bar{x}-\mu)^2 \mid \mu \leq \mu_2\}.$$

Portanto

$$2\hat{I}(^*:H_2;O_n) = \begin{cases} 0 & \text{se } \bar{x} \leq \mu_2 \\ \frac{n}{2}(\bar{x}-\mu_2)^2 & \text{se } \bar{x} > \mu_2. \end{cases}$$

Neste caso a hipótese H_2 é rejeitada se $\frac{n}{2}(\bar{x}-\mu_2)^2 > C$.

Os problemas práticos correspondentes às hipóteses do primeiro e segundo tipo não têm uma solução geral única como podemos verificar pelo estudo da distribuição, comentado no início desta seção.

- ¤ -

CAPÍTULO III

DISTRIBUIÇÃO MULTINOMIAL

Neste capítulo apresentaremos aplicações da Teoria da Informação desenvolvida no capítulo precedente no caso em que a população tem distribuição multinomial.

3.1 - DISTRIBUIÇÃO MULTINOMIAL

Suponha que ao realizarmos um experimento, um e somente um dos eventos, mutuamente exclusivos, $A_1, A_2, \ldots, A_c$ ocorra com probabilidades $p_1, p_2, \ldots, p_c$ respectivamente tais que $p_i > 0$ e $\sum_i p_i = 1$. Seja x_i, $i=1,2,\ldots,c$, o número de vezes que A_i ocorre em n realizações independentes do experimento com $\sum_i x_i = n$. Então a variável aleatória $X = (x_1, x_2, \ldots, x_c)$ tem distribuição multinomial dada pela expressão

$$P(X) = P(x_1, x_2, \ldots, x_c) = n! \prod_{i=1}^{c} \frac{p_i^{x_i}}{x_i!}.$$

A fim de facilitar o desenvolvimento das seções subsequentes passaremos agora a estender as definições do capítulo anterior ao caso da distribuição multinomial.

Consideremos duas hipóteses simples especificadas abaixo:

$$H_1: P = (P_{11}, P_{12}, \ldots, P_{1c}) \quad \sum_{i=1}^{c} P_{1i} = 1$$

$$H_2: P = (P_{21}, P_{22}, \ldots, P_{2c}) \quad \sum_{i=1}^{c} P_{2i} = 1.$$

A informação média por observação para discriminar H_1 de H_2, quando H_2 é hipótese nula, é

$$I(1:2) = p_{11} \log \frac{p_{11}}{p_{21}} + p_{12} \log \frac{p_{12}}{p_{22}} + \ldots + p_{1c} \log \frac{p_{1c}}{p_{2c}}.$$

A informação discriminante média de uma amostra aleatória O_n é dada por:

$$I(1:2;O_n) = n \sum_{i=1}^{c} p_{1i} \log \frac{p_{1i}}{p_{2i}}.$$

A fim de obtermos a Informação discriminante mínima precisamos determinar a função conjugada de uma dada distribuição multinomial. Para isto, seja $P_2(X)$ uma dada distribuição multinomial com parâmetros n e $P_2 = (p_{21}, p_{22}, \ldots, p_{2c})$ e $P(X)$ uma distribuição multinomial genérica com parâmetros n e $P = (p_1, p_2, \ldots, p_c)$. A distribuição conjugada de $P_2(X)$ é aquela que minimiza

$$\sum_{(x_1+x_2+\ldots+x_c=n)} P(X) \log \frac{P(X)}{P_2(X)}$$

sujeita à restrição $nP = \theta$, onde θ é um vetor de constantes de dimensão c-1.

De acordo com definições do capítulo II a função conjugada de $P_2(X)$ é dada por:

$$P^*(X) = \frac{P_2(X)e^{\tau' T(X)}}{M_2(\tau)}, \tag{3.1.1}$$

onde $\tau' = (\tau_1, \tau_2, \ldots, \tau_c)$ é um vetor de constantes, $T(X) = (X_1, X_2, \ldots, X_c)$ é um estimador não viesado de θ e

$$M_2(\tau) = \sum_{(X_1+X_2+\ldots+X_c=n)} P_2(X)e^{\tau' T(X)} = (P_{21}e^{\tau_1}+P_{22}e^{\tau_2}+\ldots+P_{2c}e^{\tau_c})^n.$$

Fazendo as substituições em (3.1.1), obtemos que $P^*(X)$ tem distribuição multinomial dada por:

$$P^*(X) = n! \prod_{i=1}^{c} \frac{(P_i^*)^{x_i}}{x_i!},$$

onde

$$P_i^* = \frac{P_{2i}e^{\tau_i}}{\sum_{j=1}^{c} P_{2j}e^{\tau_j}}, \qquad i=1,2,\ldots,c$$

e τ_i, $i=1,2,\ldots,c$, é um parâmetro real determinado pela equação

$$\theta_i = \frac{\partial}{\partial \tau_i} \log M_2(\tau).$$

Outra relação entre θ_i e τ_i pode ser obtida, quando fizermos $\theta_i = np_i^*$, isto é

$$\theta_i = np_i^* = n \frac{p_{2i}e^{\tau_i}}{\sum_{j=1}^{c} p_{2j}e^{\tau_j}}$$

ou seja

$$\tau_i = \log \frac{\theta_i}{np_{2i}} + \log\left(\sum_{j=1}^{c} p_{2j}e^{\tau_j}\right), \qquad i=1,2,\ldots,c.$$

Com o conhecimento de $P^*(X)$ obtemos a informação discriminante mínima

$$(3.1.2) \quad I(^*:2;O_n) = \sum P^*(X) \log \frac{P^*(X)}{P_2(X)} = \left(\sum_{i=1}^{c} \tau\, \theta_i\right) - n \log\left(\sum_{i=1}^{c} p_{2i}e^{\tau_i}\right).$$

Substituindo o valor de τ_i, encontrado acima, em (3.1.2), obtemos:

$$(3.1.3) \qquad I(^*:2;O_n) = \sum_{i=1}^{c} \theta_i \log \frac{\theta_i}{np_{2i}}.$$

Observemos que $I(^*:2;O_n)$ não depende do valor de $(p_{21}e^{\tau_1}+\ldots+p_{2c}e^{\tau_c})$, portanto podemos fazê-lo igual a 1 e obtemos:

$$\tau_i = \log \frac{\theta_i}{np_i}, \qquad i=1,2,\ldots,c.$$

3.2 - TESTE DE HIPÓTESES SIMPLES

Sejam $x_1, x_2, \ldots, x_c$, com $\sum x_i = n$, observações de uma distribuição multinomial $P(X)$ com parâmetros

$$n \quad e \quad P = (p_1, p_2, \ldots, p_c).$$

Vamos considerar o problema de testar

$$H_2: P = (p_1, p_2, \ldots, p_c) \text{ contra } H_1: P \neq (p_1, p_2, \ldots, p_c).$$

Neste caso a informação discriminante mínima é

$$I(^*:2;O_n) = n \sum_{i=1}^{c} p_i^* \log \frac{p_i^*}{p_i}$$

onde $p_1^*, p_2^*, \ldots, p_c^*$, são os parâmetros da distribuição conjugada de $P(X)$.

A informação estatística é obtida substituindo p_i^* por $\hat{p}_i^* = \frac{x_i}{n}$.

Portanto

$$\hat{I}(^*:2;O_n) = \sum_{i=1}^{c} x_i \log \frac{x_i}{np_i}.$$

Conforme vimos na seção 2.4 do capítulo anterior, $2\hat{I}(^*:2;O_n)$ tem distribuição limite de χ^2 com c-1 graus de liberdade.

Concluímos então que podemos testar H_2 contra H_1 em termos da estatística $2\hat{I}(^*:2;O_n)$, rejeitando H_2 para valores grandes de $\hat{I}(^*:2;O_n)$.

Um outro teste muito usado nestes casos é o qui-quadrado de Pearson cuja estatística é dada por:

$$\sum_i \frac{(x_i-np_i)^2}{np_i}.$$

Verifica-se que as duas estatísticas acima assumem valores aproximadamente iguais (Kullback, 1968).

2.2.1 - Partições de $\hat{I}(^*:2;O_n)$

Quando a hipótese H_2: $P=(p_1,p_2,\ldots,p_c)$ é rejeitada podemos decompor $\hat{I}(^*:2;O_n)$ aditivamente para testar sub-hipóteses de interesse.

De um modo geral as sub-hipóteses de interesse prático podem ser resumidas nos tipos de partição expressos nas tabelas 3.1 e 3.2, onde cada componente tem distribuições de χ^2 com os respectivos graus de liberdade.

A construção da tabela 3.1 baseia-se nos seguintes fatos:

1) A distribuição multinomial, P(X), pode ser expressa como o produto de uma binomial e uma multinomial condicio-

TABELA 3.1

COMPONENTE	INFORMAÇÃO	G.L.
Dentro de categorias c-1 a $c/x_1,x_2,\ldots,x_{c-2}$	$2[x_{c-1}\log\frac{x_{c-1}q_{c-2}}{n_{c-2}p_{c-1}}+x_c\log\frac{x_cq_{c-2}}{n_{c-2}p_c}]$	1
Entre categorias c-2 e $(c-1)+c/x_1,x_2,\ldots,x_{c-3}$	$2[x_{c-2}\log\frac{x_{c-2}q_{c-3}}{n_{c-3}p_{c-2}}+(n_{c-3}-x_{c-2})\log\frac{(n_{c-3}-x_{c-2})q_{c-3}}{n_{c-3}q_{c-2}}]$	1
⋮	⋮ ⋮	
Dentro de categorias 3 a $c/x_1,x_2$	$2[x_3\log\frac{x_3q_2}{n_2p_3}+x_4\log\frac{x_4q_2}{n_2p_4}+\ldots+x_c\log\frac{x_cq_2}{n_2p_2}]$	c-3
Entre categorias 2 e $(3+4+\ldots+c)/x_1$	$2[x_2\log\frac{x_2q_1}{n_1p_2}+(n_1-x_2)\log\frac{(n_1-x_2)q_1}{n_1q_2}]$	1
Dentro de categorias 2 a c/x_1	$2[x_2\log\frac{x_2q_1}{n_1p_2}+x_3\log\frac{x_3q_1}{n_1p_3}+\ldots+x_c\log\frac{x_cq_1}{n_1p_c}]$	c-2
Entre categorias 1 e $(2+3+\ldots+c)$	$2[x_1\log\frac{x_1}{np_1}+(n-x_1)\log\frac{(n-x_1)}{n(1-p_1)}]$	1
TOTAL $2\hat{I}(^*:2;0_n)$	$2[x_1\log\frac{x_1}{np_1}+x_2\log\frac{x_2}{np_2}+\ldots+x_c\log\frac{x_c}{np_c}]$	c-1

nal, isto é,

$$P(X)=\frac{n!}{x_1!x_2!\ldots x_c!}p_1^{x_1}p_2^{x_2}\ldots p_c^{x_c}=\frac{n!}{x_1!(n-x_1)!}p_1^{x_1}(1-p_1)^{n-x_1}\times$$

$$\times\frac{(n-x_1)!}{x_2!\ldots x_c!}\left(\frac{p_2}{q_1}\right)^{x_2}\left(\frac{p_3}{q_1}\right)^{x_3}\ldots\left(\frac{p_c}{q_1}\right)^{x_c}=\frac{n!}{x_1!(n-x_1)!}p_1^{x_1}(1-p_1)^{n-x_1}\times$$

$$\times\frac{(n-x_1)!}{x_2!(n-x_1-x_2)!}\left(\frac{p_2}{q_1}\right)^{x_2}\left(1-\frac{p_2}{q_1}\right)^{n-x_1-x_2}\times\frac{(n-x_1-x_2)!}{x_3!\ldots x_c!}\left(\frac{p_3}{q_2}\right)^{x_3}\ldots\left(\frac{p_c}{q_2}\right)^{x_c}$$

onde $n_1 = n-x_1$; $n_2 = n-x_1-x_2$; $q_1 = 1-p_1$; $q_2 = 1-p_1-p_2 = q_1-q_2$.

$$\sum_{i=2}^{c} \frac{p_i}{q_1} = 1 = \sum_{i=3}^{c} \frac{p_i}{q_2}.$$

De um modo geral temos:

$n_i = n-x_1-x_2-\ldots-x_i$ e $q_i = 1-p_1-p_2-\ldots-p_i$, $i=1,2,\ldots,r-1$ e

$$\sum_{i=j+1}^{c} \frac{p_i}{q_j} = 1, \qquad j=1,2,\ldots,c-1.$$

2) A hipótese H_2: $P = (p_1,p_2,\ldots,p_c)$ é equivalente à interseção das sub-hipóteses $H_{21},H_{22},\ldots,H_{2,(c-1)}$, onde H_{21} é a hipótese de que a probabilidade de ocorrência da primeira categoria seja p_1, H_{22} é a hipótese de que a probabilidade de ocorrência da segunda categoria seja p_2 dado que a da primeira tenha sido p_1, H_{23} é a hipótese de que a probabilidade de ocorrência da terceira categoria seja p_3 dado que as probabilidades de ocorrência das duas primeiras categorias tenham sido respectivamente p_1 e p_2, e assim por diante.

O fato 2 mostra que a partição permite detectar as categorias que têm probabilidade de ocorrência diferente daquela especificada pela hipótese. Para que este objetivo seja atingido por meio desta partição sugerimos o seguinte procedimento prático:

após a rejeição de H_2: $p = (p_1,p_2,\ldots,p_c)$ ordenamos as cate-

gorias em ordem crescente de acordo com os afastamentos $p_i - \hat{p}_i$, e em seguida, nesta ordem, aplicamos a partição.

Como vimos, as partições expressas na tabela 3.1 permitem detectar o afastamento de cada categoria com relação à hipótese. A seguir verificamos situações em que o interesse é dividir as categorias em dois grupos. Por exemplo, quando as i primeiras categorias apresentam características em comum e as c-1 restantes apresentam outras características em comum.

TABELA 3.2 - COMPARAÇÃO ENTRE DOIS GRUPOS DE CATEGORIAS

COMPONENTE	INFORMAÇÃO	G.L.
Entre categorias 1+2+...+i e categorias (i+1)+...+c	$2[y_1 \log \frac{y_1}{np_{11}} + y_2 \log \frac{y_2}{np_{22}}]$	1
Dentro de categorias (i+1) a c	$2[x_{i+1} \log \frac{x_{i+1}p_{22}}{y_2\, p_{i+1}} + \ldots + x_c \log \frac{x_c p_{22}}{y_2 p_c}]$	c-i-1
Dentro de categorias 1 a i	$2[x_1 \log \frac{x_1 p_{11}}{y_1 p_1} + \ldots + x_i \log \frac{x_i p_{11}}{y_1 p_i}]$	i-1
TOTAL $2\hat{I}(^*:2;O_n)$	$2[x_1 \log \frac{x_1}{np_1} + \ldots + x_c \log \frac{x_c}{np_c}]$	c-1

A construção da tabela 3.2 baseia-se no seguinte:

1) A distribuição multinomial pode ser expressa por:

$$P(X) = \frac{n!}{x_1!x_2!\ldots x_c!} \prod_{i=1}^{c} p_i^{x_i} = \frac{n!}{y_1!y_2!} p_{11}^{y_1} p_{22}^{y_2} \frac{y_1!}{x_1!x_2!\ldots x_i!} \times$$

$$\times \left(\frac{p_1}{p_{11}}\right)^{x_1} \left(\frac{p_2}{p_{11}}\right)^{x_2} \ldots \left(\frac{p_i}{p_{11}}\right)^{x_i} \frac{y_2!}{x_{i+1}!\ldots x_c!} \left(\frac{p_{i+1}}{p_{22}}\right)^{x_{i+1}} \ldots \left(\frac{p_c}{p_{22}}\right)^{x_c}$$

onde $y_1 = x_1+x_2+\ldots+x_i$; $y_2 = x_{i+1}+x_{i+2}+\ldots+x_c$

$p_{11} = p_1+p_2+\ldots+p_i$; $p_{22} = p_{i+1}+p_{i+2}+\ldots+p_c$.

2) A hipótese H_2: $P = (p_1,p_2,\ldots,p_c)$ é equivalente à interseção das sub-hipóteses H_{21}, H_{22}, H_{23}, onde (se considerarmos o vetor P com c componentes particionado em $P^{(1)}$ com i componentes e $P^{(2)}$ com $c-i$ componentes) H_{21} é a hipótese de que $P^{(1)} = (p_1,p_2,\ldots,p_i)$, H_{22} é a hipótese de que $P^{(2)} = (p_{i+1},p_{i+2},\ldots,p_c)$ e H_{23} é a hipótese de que $P_{11} = p_{22}$, onde

$$P_{11} = \sum_{j=1}^{c} p_j \quad \text{e} \quad P_{22} = \sum_{j=i+1}^{c} p_j.$$

<u>EXEMPLO 1</u> - De acordo com a segunda lei de Mendel em genética dizemos que duas características genéticas examinadas simultaneamente são independentes quando os fenótipos encontrados estão na proporção 9:3:3:1.

Consideremos o caso em que as características examinadas são:

Cor de semente de ervilha: Amarela (A) e verde (a).

Forma de semente de ervilha: lisa (B) e rugosa (b).

Sabemos por experimentos realizados com cada característica isolada que, a cor amarela é dominante sobre a cor verde e que a forma lisa é dominante sobre a forma rugosa. Então as proporções esperadas, em cruzamento do tipo AaBb ×AaBb são:

9/16: lisa, amarela (A_1)

3/16: lisa, verde (A_2)

3/16: rugosa, amarela (A_3)

1/16: rugosa, verde (A_4)

Para os dados na tabela abaixo queremos verificar se a segunda lei de Mendel é verdadeira ou não:

DADOS:

FENÓTIPO	FREQUÊNCIA OBSERVADA
AMARELA, LISA	75
VERDE, LISA	23
AMARELA, RUGOSA	22
VERDE, RUGOSA	1
TOTAL	121

HIPÓTESE:

$$H_2: P = \begin{pmatrix} 3/16 \\ 3/16 \\ 3/16 \\ 1/16 \end{pmatrix} \qquad H_1: P \neq \begin{pmatrix} 9/16 \\ 3/16 \\ 3/16 \\ 1/16 \end{pmatrix}$$

A estatística para o teste é:

$$2\hat{I}(*:2,O_n) = 2\sum_{i=1}^{c} x_i \log \frac{x_i}{np_i} = 9.1894.$$

O valor crítico do χ^2 com 3 graus de liberdade é 1.815 ao nível $\alpha = 0,05$. Portanto a hipótese H_2 é rejeitada ao nível de significância, $\alpha = 0,05$. Neste caso podemos decompor $2\hat{I}(*:2;O_n)$, como mostra a tabela 3.1, para verificarmos quais dos fenótipos são responsáveis pela rejeição da hipótese. Para isto recorremos ao procedimento prático sugerido nesta seção e escrevemos P da seguinte forma:

$$P = \begin{pmatrix} 3/16 \\ 3/16 \\ 1/16 \\ 9/16 \end{pmatrix} = \begin{pmatrix} P_2 \\ P_3 \\ P_4 \\ P_1 \end{pmatrix}$$

Então testamos a hipótese

$$\begin{cases} H_2: p_2 = 3/16 \\ H_1: p_2 \neq 3/16. \end{cases}$$

De acordo com a tabela 3.1 a estatística para testar a hipótese H_{21} é

$$2(x_1 \log \frac{x_1}{np_1} + (n-x_1)\log \frac{(n-x_1)}{n(1-p_1)}) = 0,006.$$

O valor crítico do χ^2 com 1 grau de liberdade é .. 3.841 ao nível $\alpha = 0,05$. Portanto a hipótese H_{21} não é rejeitada ao nível $\alpha = 0,05$.

Dado que $p_2 = 3/16$, vamos testar a hipótese:

$$\begin{cases} H_{22}: \begin{pmatrix} p_3 \\ p_4 \\ p_1 \end{pmatrix} = \begin{pmatrix} 3/16 \\ 1/16 \\ 9/16 \end{pmatrix} \\ \\ H_1' : \begin{pmatrix} p_3 \\ p_4 \\ p_1 \end{pmatrix} \neq \begin{pmatrix} 3/16 \\ 1/16 \\ 9/16 \end{pmatrix} \end{cases}$$

A estatística para testar esta hipótese (ver tabela 3.1) é:

$$2\sum_i x_i \log \frac{x_i q_1}{np_i} = 9.7835.$$

O valor crítico do χ^2 com 2 graus de liberdade é 5.991. Portanto rejeitamos a hipótese H_{22} ao nível de significância $\alpha = 0,05$.

Uma vez que a hipótese H_{22} foi rejeitada devemos particionar a Informação $2\sum x_i \log \frac{x_i q_1}{np_i}$ em duas outras componentes,

$$2(x_2 \log \frac{x_2 q_1}{n_1 p_2} + (n_1 - x_2) \log \frac{(n_1 - x_2)}{n_1 q_2}) \text{ e } 2(x_3 \log \frac{x_3 q_2}{n_2 p_3} + x_4 \log \frac{x_4 q_2}{n_2 p_4})$$

as quais testam, respectivamente, as hipóteses:

$$H_{23}: p_3 = 3/16 \quad \text{dado } p_2 = 3/16$$

$$H_1'' : p_3 = 3/16$$

$$\begin{cases} H_{24}: \begin{pmatrix} p_4 \\ p_1 \end{pmatrix} = \begin{pmatrix} 1/16 \\ 9/16 \end{pmatrix} \quad \text{dados} \quad P_1 = 3/16 \text{ e } P_2 = 3/16 \\ H_1''': \begin{pmatrix} p_4 \\ p_1 \end{pmatrix} \neq \begin{pmatrix} 1/16 \\ 9/16 \end{pmatrix} \end{cases}$$

O valor observados para as estatísticas correspondentes ao teste das hipóteses H_{23} e H_{24} foram: 0,02028 e 9,76322, respectivamente. Comparando estes resultados com o valor crítico do χ^2 com 1 grau de liberdade, concluímos que a hipótese H_{23} não é rejeitada e H_{24} é rejeitada ao nível de significância $\alpha = 0,05$.

Para ficar mais claro vamos apresentar estes resultados na tabela a seguir:

COMPONENTE	INFORMAÇÃO	G.L.
Dentro de fenótipos A_4 e A_1/x_2 e x_3	9,76322	1
Entre fenótipos A_3 e A_4+A_1/x_2	0,02028	1
Dentro de fenótipos A_3, A_4 e A_1/x_2	9,7835	2
Entre fenótipos A_2 e $A_3+A_4+A_1$	0,0059	1
$2\hat{I}(*:2;0_n)$	9,7894	3

CONCLUSÃO - Analisando os resultados da tabela verificamos que as frequências observadas correspondentes aos fenótipos A_1 e A_4 são as que se afastam significativamente do esperado pela lei de Mendel.

3.3 - TESTE DE HOMOGENEIDADE

Consideremos r populações multinomiais com c categorias.

Sejam, $x_{i1}, x_{i2}, \ldots, x_{ic}$ observações da i-ésima população multinomial, $P_i(X)$, com parâmetros:

$$n_i \quad e \quad P_i = (p_{i1}, p_{i2}, \ldots, p_{ic}).$$

O problema agora consiste em testar a hipótese

$$\begin{cases} H_2: \ p_{ij} = p_j & i=1,2,\ldots,r \quad j=1,2,\ldots,c \quad \sum_{j=1}^{c} p_j = 1 \\ H_1: \ p_{ij} \neq p_j & \text{para algum } j. \end{cases}$$

A informação discriminante mínima neste caso é dada por:

$$(3.3.1) \qquad I(^*:2;O_n) = \sum_{i=1}^{r} n_i \sum_{j=1}^{c} p^*_{ij} \log \frac{p^*_{ij}}{p_j},$$

onde p^*_{ij}, $j=1,2,\ldots,c$, são os parâmetros da distribuição conjugada de $P_i(X)$.

Substituindo p^*_{ij} em (3.3.1) por $\hat{p}^*_{ij} = \frac{x_{ij}}{n_i}$, obtemos a informação estatística

$$\hat{I}(^*:2;O_n) = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{n_i p_j}$$

a qual testa se as r amostras vieram de uma mesma população multinomial com parâmetros $P = (p_1, p_2, \ldots, p_c)$. A distribuição limite de $2\hat{I}(^*:2;O_n)$ sob H_2 é χ^2 com $r(c-1)$ graus de liberdade.

Quando H_2 é rejeitada podemos decompor $\hat{I}(^*:2;O_n)$ em

duas componentes aditivas como mostra a tabela 3.3.

TABELA 3.3

COMPONENTE	INFORMAÇÃO	G.L.
$2\hat{I}(\hat{p}:p)$	$2\sum_{j=1}^{c} x_j \log \frac{x_i}{np_j^0}$	c-1
$2\hat{I}(p^*:\hat{p})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{nx_{ij}}{n_i x_j}$	(r-1)(c-1)
TOTAL $2\hat{I}(*:2;0_n)$	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{n_i p_j^0}$	r(c-1)

A componente $2\hat{I}(P^*:\hat{P})$ testa a hipótese H_{21}: "As populações são homogêneas". A componente $2\hat{I}(\hat{p}:p)$ testa a hipótese H_{22}: "O parâmetro comum das r populações homogêneas é $P = (p_1^0, p_2^0, \ldots, p_c^0)$."

Em muitas situações práticas o parâmetro $P = (p_1, p_2, \ldots, p_c)$ é desconhecido, então o teste de homogeneidade entre as r populações é feito com base na estatística $2\hat{I}(p^*,\hat{p})$, a qual, sob a hipótese de homogeneidade, tem distribuição limite de χ^2 com (r-1)(c-1) graus de liberdade.

Quando esta hipótese é rejeitada podemos decompor $2\hat{I}(p^*:\hat{p})$ em componentes aditivas de duas maneiras diferentes, de acordo com as tabelas 3.4 e 3.5, onde cada componen

te tem distribuição de χ^2 com seus respectivos graus de liberdade.

TABELA 3.4

COMPONENTE	INFORMAÇÃO	G.L.
Dentro das populações 1 e 2	$2\sum_{i=1}^{2}\sum_{j=1}^{c} x_{ij} \log \frac{M_2 x_{ij}}{n_i y_{2j}}$	c-1
Entre às populações 3 e as populações 1 e 2	$2\sum_{j=1}^{c}\left(x_{3j} \log \frac{M_3 x_{3j}}{n_3 y_{3j}} + y_{2j} \log \frac{M_3 y_{2j}}{M_2 y_{3j}}\right)$	c-1
⋮	⋮ ⋮	
Dentro das populações 1 a r-2	$2\sum_{i=1}^{r-2}\sum_{j=1}^{c} x_{ij} \log \frac{M_{r-2} x_{ij}}{n_i y_{r-2,j}}$	(r-3)(c-1)
Entre as populações r-1 e populações 1 a r-2	$2\sum_{j=1}^{c}\left(x_{r-1,j} \log \frac{M_{r-1} x_{r-1,j}}{n_{r-1} y_{r-1,j}} + y_{r-2,j} \log \frac{M_{r-1} y_{r-2,j}}{M_{r-2} y_{r-1,j}}\right)$	c-1
Dentro das populações 1 a r-1	$2\sum_{i=1}^{r-1}\sum_{i=1}^{c} x_{ij} \log \frac{M_{r-1} x_{ij}}{n_i y_{r-1,j}}$	(r-2)(c-1)
Entre a população r e as populações 1 a r-1	$2\sum_{j=1}^{c}\left(x_{rj} \log \frac{n x_{rj}}{n_r x_j} + y_{r-1,j} \log \frac{n y_{r-1,j}}{M_{r-1} x_j}\right)$	c-1
TOTAL $2\hat{I}(P^*:\hat{P})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{n x_{ij}}{n_i x_j}$	(r-1)(c-1)

A construção da tabela 3.4 baseia-se no seguinte:

1) A distribuição de duas populações multinomiais independentes pode ser expressa pelo produto de uma distribuição marginal das populações combinadas e uma distribuição condicional de cada população dada a população combinada, isto

é,

$$P(X) = \frac{n_1!}{x_{11}!\ldots x_{1c}!} p_1^{x_{11}} \times \ldots \times p_c^{x_{1c}} \frac{n_2!}{x_{21}!\ldots x_{2c}!} p_1^{x_{21}} \times \ldots \times p_c^{x_{2c}} =$$

$$= \frac{(n_1+n_2)!}{y_{21}!\ldots y_{2c}!} p_1^{y_{21}} \times \ldots \times p_c^{y_{2c}} \frac{n_1!n_2!}{(n_1+n_2)!} \cdot \frac{y_{21}!\ldots y_{2c}!}{x_{11}!\ldots x_{1c}!x_{21}!\ldots x_{2c}!}$$

onde $y_{2j} = x_{1j}+x_{2j}$, $j=1,2,\ldots,c$ e $n_1+n_2 = M_2 = y_{21}+y_{22}+\ldots+y_{2c}$.

No caso de termos r populações independentes o procedimento para expressarmos P(X) é análogo ao acima só que as populações combinadas serão indicadas por

$$y_{ij} = x_{ij}+x_{2j}+\ldots+x_{ij}, \qquad i=1,2,\ldots,r-1 \text{ e } j=1,2,\ldots,c$$

e

$$y_{i1}+y_{i2}+\ldots+y_{ic} = n_1+n_2+\ldots+n_i = M_i.$$

2) A hipótese de homogeneidade H_2 é equivalente à interseção das sub-hipóteses $H_2(1,2)$, $H_2(1+2,3)$,...,$H_2(1+2+\ldots+r-1,r)$, onde $H_2(1,2)$ é a hipótese de que as populações 1 e 2 sejam homogêneas, $H_2(1+2,3)$ é a hipótese de que a população 3 e a população formada pela combinação das populações 1 e 2 sejam homogêneas e assim por diante.

Para a partição indicada na tabela 3.5, consideramos as r populações em dois grupos. O grupo 1 constituído das populações 1 a r, e o grupo 2 das populações r_1+1 a r.

Definamos

$$z_{1j} = x_{1j}+x_{2j}+\ldots+x_{r_1 j} \qquad j=1,2,\ldots,c$$
$$z_{2j} = x_{r_1+1j}+x_{r_1+2j}+\ldots+x_{rj}$$

$$T_1 = \sum_{j=1}^{c} z_{1j} \qquad T_2 = \sum_{j=1}^{c} z_{2j} \qquad T_1+T_2 = n.$$

TABELA 3.5

COMPONENTE	INFORMAÇÃO	G.L.
Entre conjuntos 1 e 2	$2\sum_{j=1}^{c}\left(z_{ij}\log\frac{nz_{1j}}{T_1x_j}+z_{2j}\log\frac{nz_{1j}}{T_2x_j}\right)$	c-1
Dentro do conjunto 2	$2\sum_{i=r_1+1}^{r}\sum_{j=1}^{c}x_{ij}\log\frac{T_2x_{ij}}{n_iz_{2j}}$	$(r-r_1-1)(c-1)$
Dentro do conjunto 1	$2\sum_{i=1}^{r_1}\sum_{j=1}^{c}x_{ij}\log\frac{T_1x_{ij}}{n_iz_{1j}}$	$(r_1-1)(c-1)$
TOTAL $2\hat{I}(P^*:\hat{P})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}\log\frac{nx_{ij}}{n_ix_j}$	(r-1)(c-1)

OBSERVAÇÃO - As r populações podem ser divididas em dois ou mais grupos, por conviniência consideramos apenas dois grupos.

EXEMPLO 2 - Os dados da Tabela 3.6 representam a produção de 5 tipos de máquinas, classificada em defeituosa e perfeita. Queremos saber se as máquinas são homogêneas com relação à produção.

TABELA 3.6

MÁQUINAS

	A	B	C	D	E
Defeituosa	26	72	61	29	135
Perfeita	172	169	142	36	542
TOTAL	198	241	203	65	677

HIPÓTESE:

$$\begin{cases} H_2: \ p_{ij} = p_j \text{ para todo } j \\ H_1: \ p_{ij} \neq p_j \text{ para algum } j \end{cases} \quad j=1,2; \ i=1,2,\ldots,5.$$

De acordo com a tabela 3.4 a estatística para testar a hipótese acima é:

$$2\hat{I}(P^*:\hat{P}) = 2\sum_{i=1}^{5}\sum_{j=1}^{2} x_{ij} \log \frac{nx_{ij}}{n_i x_j} =$$

$$= 2[\sum_{i=1}^{5}\sum_{j=1}^{2} x_{ij} \log x_{ij} - \sum_{j=1}^{2} x_j \log x_j + n \log n - \sum_{i=1}^{5} n_i \log n_i].$$

Dos dados obtemos $2\hat{I}(P^*:\hat{P}) = 42,04940$ que comparado com o valor crítico do χ^2 com 4 graus de liberdade, ao nível de significância $\alpha = 0,05$, (9,488) nos leva a rejeitar a hipótese que as máquinas são homogêneas com relação à produção.

Então podemos decompor $2\hat{I}(P^*:\hat{P})$ em componentes aditivas para que possamos analisar com mais detalhes a hipóte

se. Para isto vamos considerar a partição da tabela 3.4, que de acordo com os dados do problema se resume na Tabela 3.7.

TABELA 3.7

COMPONENTE	INFORMAÇÃO	G.L.
Dentro A e B	2.142564	1
Entre C e (A+B)	4.350176	1
Dentro A,B e C	6.49274	2
Entre D e (A+B+C)	10.84138	1
Dentro de A,B,C e D	17.334126	3
Entre E e (A+B+C+D)	24.715274	1
TOTAL	42.049400	4

Os valores críticos dos χ_1^2, χ_2^2 e χ_3^2, ao nível 0,05, são respectivamente 3.841, 5.991 e 7.815. Comparando os valores observados com seus respectivos valores críticos concluímos que as máquinas A e B são homogêneas entre si e diferentes de C, que as máquinas A, B e C são diferentes de D e que as máquinas A, B, C e D são diferentes de E.

- ¤ -

CAPÍTULO IV

TABELAS DE CONTINGÊNCIA COMPLETA

Neste capítulo analisaremos Tabelas de Contingência Completa aplicando Informação Estatística, como foi realizado para Distribuição Multinomial.

Analisaremos Tabelas de Contingência de Duas e de Três Entradas. A extensão da análise para Tabelas de ordem mais alta, embora não apresente problemas do ponto de vista conceitual, envolve um algebrismo muito complexo.

4.1 - TABELA DE CONTINGÊNCIA: CONCEITUAÇÃO

Suponhamos que os elementos de uma população sejam classificados segundo dois atributos R e C com r e c categorias respectivamente. Representemos por n o tamanho de uma amostra dessa população e por n_{ij} o número de elementos da amostra que pertença à i-ésima categoria do atributo R $(i=1,2,\ldots,r)$ e à j-ésima categoria do atributo C $(j=1,2,\ldots,c)$.

Os elementos, classificados desse modo podem ser representados na tabela 4.1, a qual é denominada *Tabela de Con-*

tigência.

Quando temos dois atributos denominaremos *Tabela de Contigência de Dupla Entrada*, três atributos, Tabela de Contingência de Três Entradas, K atributos, Tabela de Contigência de K Entradas.

TABELA 4.1

C

		C_1	C_2	...	C_c	TOTAL
	R_1	x_{11}	x_{12}	...	x_{1c}	$x_{1.}$
	R_2	x_{21}	x_{22}	...	x_{2c}	$x_{2.}$
R	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
	R_r	x_{r1}	x_{r2}	...	x_{rc}	$x_{r.}$
	TOTAL	$x_{.1}$	$x_{.2}$	...	$x_{.c}$	

4.2 - TABELA DE CONTINGÊNCIA DE DUPLA ENTRADA: TESTE DE INDEPENDÊNCIA

Consideremos uma Tabela de Contingência de Dupla Entrada como definida anteriormente e representemos por p_{ij} a probabilidade de ocorrência da casela (i,j) i=1,2,...r e j= =1,2,...,c.

O problema consiste em testar a independência entre dois atributos. Nestes casos formulamos a seguinte hipótese:

$$\begin{cases} H_2: p_{ij} = p^0_{i.}p^0_{.j} \text{ para toda casela } (i,j) \\ H_1: p_{ij} \neq p^0_{i.}p^0_{.j} \text{ para alguma casela } (i,j) \end{cases} \quad i=1,2,\ldots,r;\ j=1,2,\ldots,c$$

onde

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1,\ p_{ij} > 0 \quad e \quad \sum_{i=1}^{r} p^0_{i.} = 1 = \sum_{j=1}^{c} p^0_{.j}.$$

O expoente 0 indica que o valor do parâmetro é especificado.

Como uma tabela de contingência de ordem r×c é essencialmente uma amostra de uma população multinomial com rc categorias, omitiremos detalhes já desenvolvidos no capítulo anterior.

A informação discriminante mínima, neste caso, é dada por

$$I(^*:2;O_n) = n \sum_{i=1}^{r}\sum_{j=1}^{c} p^*_{ij} \log \frac{p^*_{ij}}{p^0_{i.}p^0_{.j}}.$$

Ao substituirmos p^*_{ij} por $\hat{p}^*_{ij} = \frac{x_{ij}}{n}$, obtemos a informação estatística

$$\hat{I}(^*:2;O_n) = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{np^0_{i.}p^0_{.j}}.$$

Sob a hipótese de independência $2\hat{I}(^*:2;O_n)$ tem distribuição limite de χ^2 com rc-1 graus de liberdade.

Podemos decompor $\hat{I}(^*:2,O_n)$ em componentes aditivas de modo que possamos analisar com mais detalhes a hipótese H_2: $p_{ij} = p^0_{i.}p^0_{.j}$, como mostra a tabela 4.2.

TABELA 4.2

COMPONENTE	INFORMAÇÃO	G.L.
$H_2(R)$	$2\sum_{i=1}^{r} x_{i.} \log \frac{x_{i.}}{np^0_{i.}}$	r-1
$H_2(C)$	$2\sum_{j=1}^{c} x_{.j} \log \frac{x_{.j}}{np^0_{.j}}$	c-1
$2\hat{I}(H_1:H_{23})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{nx_{ij}}{x_{i.}x_{.j}}$	(r-1)(c-1)
TOTAL $2\hat{I}(^*:2;O_n)$	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{np^0_{i.}p^0_{.j}}$	rc-1

As componentes $H_2(R)$, $H_2(C)$ e $2\hat{I}(H_1;H_{23})$ testam, respectivamente, as hipóteses:

H_{21}: $p_{i.}=p^0_{i.}$, $i=1,2,\ldots,r$; H_{22}: $p_{.j}=p^0_{.j}$, $j=1,2,\ldots,c$ e

H_{23}: $p_{ij}=p_{i.}p_{.j}$ para todo (ij).

4.2.1 - Partição da Componente $\hat{I}(H_1:H_{23})$

Valores significantes para $\hat{I}(H_1:H_{23})$ da Tabela 4.2 podem sugerir dependência entre linhas e colunas de um de-

terminado subconjunto da tabela. Nestes casos podemos particionar a Tabela de Contingência em subtabelas convinientes para testarmos essas possíveis dependências.

Suponhamos, por exemplo, que seja de interesse particionar uma Tabela de Contingência de ordem r×c em quatro subtabelas, agrupando as linhas em dois subconjuntos de r_1 e r_2 linhas respectivamente, $r_1 + r_2 = r$ e as colunas em dois subconjuntos de c_1 e c_2 colunas respectivamente, $c_1 + c_2 = c$, como mostram as tabelas 4.2.1, 4.2.2, 4.2.3 e 4.2.4.

TABELA 4.3.1

	1	2	...	c_1	TOTAL
1	x_{11}	x_{12}	...	x_{1c_1}	$x^{11}_{1.}$
2	x_{21}	x_{22}	...	x_{2c_1}	$x^{11}_{2.}$
⋮	⋮	⋮	...	⋮	⋮
r_1	$x_{r_1 1}$	$x_{r_1 2}$	...	$x_{r_1 c_1}$	$x^{11}_{r_1.}$
TOTAL	$x^{11}_{.1}$	$x^{11}_{.2}$	...	$x^{11}_{.c_1}$	n_{11}

TABELA 4.3.2

	c_1+1	c_1+2	...	c_1+c_2	TOTAL
1	x_{1c_1+1}	x_{1c_1+2}	...	$x_{1c_1+c_2}$	$x^{12}_{1.}$
2	x_{2c_1+1}	x_{2c_1+2}	...	$x_{2c_1+c_2}$	$x^{12}_{2.}$
⋮	⋮	⋮	...	⋮	⋮
r_1	$x_{r_1c_1+1}$	$x_{r_1c_1+2}$	...	$x_{r_1c_1+c_2}$	$x^{12}_{r_1.}$
TOTAL	$x^{12}_{.c_1+1}$	$x^{12}_{.c_1+2}$	...	$x^{12}_{.c_1+c_2}$	n_{12}

TABELA 4.3.3

	1	2	...	c_1	TOTAL
r_1+1	x_{r_1+11}	x_{r_1+12}	...	$x_{r_1+1c_1}$	$x^{21}_{r_1+1\cdot}$
r_1+2	x_{r_1+21}	x_{r_1+22}	...	$x_{r_1+2c_1}$	$x^{21}_{r_1+2\cdot}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
r_1+r_2	$x_{r_1+r_21}$	$x_{r_1+r_22}$	...	$x_{r_1+r_2c_1}$	$x^{21}_{r_1+r_2\cdot}$
TOTAL	$x^{21}_{\cdot 1}$	$x^{21}_{\cdot 2}$	...	$x^{21}_{\cdot c_1}$	n_{21}

TABELA 4.3.4

	c_1+1	c_1+2	...	c_1+c_2	TOTAL
r_1+1	$x_{r_1+1c_1+1}$	$x_{r_1+1c_1+2}$	...	$x_{r_1+1c_1+c_2}$	$x^{22}_{r_1+1\cdot}$
r_1+2	$x_{r_1+2c_1+1}$	$x_{r_1+2c_1+2}$	...	$x_{r_1+2c_1+c_2}$	$x^{22}_{r_1+2\cdot}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
r_1+r_2	$x_{r_1+r_2c_1+1}$	$x_{r_1+r_2c_1+2}$	...	$x_{r_1+r_2c_1+c_2}$	$x^{22}_{r_1+r_2\cdot}$
TOTAL	$x^{22}_{\cdot c_1+1}$	$x^{22}_{\cdot c_1+2}$	...	$x^{22}_{\cdot c_1+c_2}$	n_{22}

Neste caso a partição de $\hat{I}(H_1:H_{23})$ se resume na tabela 4.3.

TABELA 4.3

COMPONENTE	INFORMAÇÃO	G.L.
9	$2\sum_{\alpha=1}^{2}\sum_{\beta=1}^{2} n_{\alpha\beta} \log \frac{nn_{\alpha\beta}}{n_{\alpha\cdot}n_{\cdot\beta}}$	1
8	$2\sum_{j=c_1+1}^{c_1+c_2}\left(x_{\cdot j}^{12} \log \frac{n_{\cdot 2}x_{\cdot j}^{12}}{n_{12}x_{\cdot j}^{\cdot 2}} + x_{\cdot j}^{22} \log \frac{n_{\cdot 2}x_{\cdot j}^{22}}{n_{22}x_{\cdot j}^{\cdot 2}}\right)$	c_2-1
7	$2\sum_{j=1}^{c_1}\left(x_{\cdot j}^{11} \log \frac{n_{\cdot 1}x_{\cdot j}^{11}}{n_{11}x_{\cdot j}^{\cdot 1}} + x_{\cdot j}^{21} \log \frac{n_{\cdot 1}x_{\cdot j}^{21}}{n_{2}\ x_{\cdot j}^{\cdot 1}}\right)$	c_1-1
6	$2\sum_{i=r_1+1}^{r_1+r_2}\left(x_{i\cdot}^{21} \log \frac{n_{2\cdot}x_{i\cdot}^{21}}{n_{21}x_{i\cdot}^{2\cdot}} + x_{i\cdot}^{22} \log \frac{n_{2\cdot}x_{i\cdot}^{22}}{n_{22}x_{i\cdot}^{2\cdot}}\right)$	r_2-1
5	$2\sum_{i=1}^{r_1}\left(x_{i\cdot}^{11} \log \frac{n_{1\cdot}x_{i\cdot}^{11}}{n_{11}x_{i\cdot}^{1\cdot}} + x_{i\cdot}^{12} \log \frac{n_{1\cdot}x_{i\cdot}^{12}}{n_{12}x_{i\cdot}^{1\cdot}}\right)$	r_1-1
4	$2\sum_{i=r_1+1}^{r_1+r_2}\sum_{j=c_1+1}^{c_1+c_2} x_{ij} \log \frac{n_{22}x_{ij}}{x_{i\cdot}^{22}x_{\cdot j}^{22}}$	$(r_2-1)(c_2-1)$
3	$2\sum_{i=r_1+1}^{r_1+r_2}\sum_{j=1}^{c_1} x_{ij} \log \frac{n_{21}x_{ij}}{x_{i\cdot}^{21}x_{\cdot j}^{21}}$	$(r_2-1)(c_1-1)$
2	$2\sum_{i=1}^{r_1}\sum_{j=c_1+1}^{c_1+c_2} x_{ij} \log \frac{n_{12}x_{ij}}{x_{i\cdot}^{12}x_{\cdot j}^{12}}$	$(r_1-1)(c_2-1)$
1	$2\sum_{i=1}^{r_1}\sum_{j=1}^{c_1} x_{ij} \log \frac{n_{11}x_{ij}}{x_{i\cdot}^{11}x_{\cdot j}^{11}}$	$(r_1-1)(c_1-1)$
TOTAL $2\hat{I}(H_1:H_{23})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \log \frac{nx_{ij}}{x_{i\cdot}x_{\cdot j}}$	$(r-1)(c-1)$

As componentes 1, 2, 3 e 4 testam a hipótese de independência nas sub-tabelas 4.2.1, 4.2.2, 4.2.3 e 4.2.4 respectivamente.

As componentes 5, 6, 7, 8 e 9 testam, respectivamente, a hipótese de independência nas sub-tabelas abaixo:

5

$x^{11}_{1\cdot}$	$x^{12}_{1\cdot}$
$x^{11}_{2\cdot}$	$x^{12}_{2\cdot}$
$\vdots$	$\vdots$
$x^{11}_{r_1\cdot}$	$x^{12}_{r_1\cdot}$

6

$x^{21}_{r_1+1\cdot}$	$x^{22}_{r_1+1\cdot}$
$x^{21}_{r_1+2\cdot}$	$x^{22}_{r_1+2\cdot}$
$\vdots$	$\vdots$
$x^{21}_{r_1+r_2\cdot}$	$x^{22}_{r_1+r_2\cdot}$

7

$x^{11}_{\cdot 1}$	$x^{11}_{\cdot 2}$	...	$x^{11}_{\cdot c}$
$x^{21}_{\cdot 1}$	$x^{21}_{\cdot 2}$	...	$x^{21}_{\cdot c}$

8

$x^{12}_{\cdot c_1+1}$	$x^{12}_{\cdot c_1+2}$	...	$x^{12}_{\cdot c_1+c_2}$
$x^{22}_{\cdot c_1+1}$	$x^{22}_{\cdot c_1+2}$	...	$x^{22}_{\cdot c_1+c_2}$

9

n_{11}	n_{12}
n_{21}	n_{22}

4.3 - TABELAS DE CONTINGÊNCIA DE TRÊS ENTRADAS: TESTE DE INDEPENDÊNCIA

Consideremos três atributos R, C e D com r, c e d categorias respectivamente. Consideremos uma amostra de n observações independentes da população formada por R,C e D. Sejam, x_{ijk} a frequência de ocorrência da i-ésima, j-ésima e k-ésima categorias dos atributos R,C e D, respectivamente e p_{ijk} a probabilidade de ocorrência da casela (i,j,k), como mostra a Tabela 5.3

TABELA 5.3

	C_1				C_2				...	C_c			
	D_1	D_2	...	D_d	D_1	D_2	...	D_d	...	D_1	D_2	...	D_d
R_1	x_{111}	x_{112}	...	x_{11d}	x_{121}	x_{122}	...	x_{12d}	...	x_{1c1}	x_{1c2}	...	x_{1cd}
R_2	x_{211}	x_{212}	...	x_{21d}	x_{221}	x_{222}	...	x_{22d}	...	x_{2c1}	x_{2c2}	...	x_{2cd}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	...	$\vdots$
R_r	x_{r11}	x_{r12}	...	x_{r1d}	x_{r21}	x_{r22}	...	x_{r2d}	...	x_{rc1}	x_{rc2}	...	x_{rcd}

Definamos

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk}, \quad x_{i..} = \sum_j \sum_k x_{ijk}$$

$$x_{.j.} = \sum_i \sum_k x_{ijk} \qquad x_{..k} = \sum_i \sum_j x_{ijk}$$

$$x_{ij\cdot} = \sum_k x_{ijk}, \quad x_{i\cdot k} = \sum_j x_{ijk}, \quad x_{\cdot jk} = \sum_i x_{ijk}.$$

4.3.1 - Hipótese de Independência Total

Para testarmos a independência entre três atributos formulamos a seguinte hipótese:

$$\begin{cases} H_2: p_{ijk} = p^0_{i\cdot\cdot} p^0_{\cdot j\cdot} p^0_{\cdot\cdot k} & \text{para toda casela } (i,j,k) \\ & i=1,2,\ldots,r; j=1,2,\ldots,c; k=1,2,\ldots,d \\ H_1: p_{ijk} \neq p^0_{i\cdot\cdot} p^0_{\cdot j\cdot} p^0_{\cdot\cdot k} & \text{para alguma casela } (i,j,k) \end{cases}$$

onde $\sum_i \sum_j \sum_k p_{ijk} = 1 \quad p_{ijk} > 0; \quad \sum_i p_{i\cdots} = \sum_j p_{\cdot j\cdot} = \sum_k p_{\cdot\cdot k} = 1.$

A informação discriminante mínima é dada por:

$$I(^*:2;O_n) = n \sum_i \sum_j \sum_k p^*_{ijk} \log \frac{p^*_{ijk}}{p^0_{i\cdot\cdot} p^0_{\cdot j\cdot} p^0_{\cdot\cdot k}}.$$

Substituindo p^*_{ijk} pela sua estimativa $\hat{p}_{ijk} = \frac{x_{ijk}}{n}$ obtemos a informação estatística

$$\hat{I}(^*:2;O_n) = \sum_i \sum_j \sum_k x_{ijk} \log \frac{x_{ijk}}{n p^0_{i\cdot\cdot} p^0_{\cdot k\cdot} p^0_{\cdot\cdot k}}$$

a qual sob a hipótese de independência tem distribuição limite de χ^2 com rcd-1 graus de liberdade.

A informação estatística $\hat{I}(^*:2;O_n)$ pode ser decomposta em componentes aditivas que testam sub-hipóteses de in

teresse, como vemos na Tabela 4.4

TABELA 4.4

COMPONENTE	INFORMAÇÃO	G.L.
$H_2(R)$	$2\sum_{i=1}^{r} x_{i..} \log \frac{x_{i..}}{np^0_{i..}}$	r-1
$H_2(C)$	$2\sum_{j=1}^{c} x_{.j.} \log \frac{x_{.j.}}{np^0_{.j.}}$	c-1
$H_2(D)$	$2\sum_{k=1}^{d} x_{..k} \log \frac{x_{..k}}{np^0_{..k}}$	d-1
$2\hat{I}(H_1:H_{24})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{n^2 x_{ijk}}{x_{i..}x_{.j.}x_{..k}}$	rcd-r-c-d+2
TOTAL $2\hat{I}(^*:2;0_n)$	$2\sum_i\sum_j\sum_k x_{ijk} \log \frac{x_{ijk}}{np^0_{i..}p^0_{.j.}p^0_{..k}}$	rcd-1

onde $H_2(R)$ testa a hipótese H_{21}: $p_{i..}=p^0_{i..}$, $H_2(C)$ testa a hipótese H_{22}: $p_{.j.}=p^0_{.j.}$, $H_2(D)$ testa a hipótese H_{23}: $p_{..k}=p^0_{..k}$ e $2\hat{I}(H_1:H_{24})$ testa a hipótese H_{24}: $p_{ijk}=p_{i..}p_{.j.}p_{..k}$.

A componente $2\hat{I}(H_1:H_{24})$ da Tabela 4.4 pode ser decomposta em componentes aditivas, como vemos na Tabela 4.5.

TABELA 4.5

COMPONENTE	INFORMAÇÃO	G.L.
$H_2(C\times D)$	$2\sum_j \sum_k x_{.jk} \log \frac{nx_{.jk}}{x_{.j.}x_{..k}}$	(c-1)(d-1)
$H_2(R\times CD)$	$2\sum_i \sum_j \sum_k x_{ijk} \log \frac{nx_{ijk}}{x_{i..}x_{.jk}}$	(r-1)(cd-1)
TOTAL $2\hat{I}(H_1:H_{24})$	$2\sum_i \sum_j \sum_k x_{ijk} \log \frac{n^2x_{ijk}}{x_{i..}x_{.j.}x_{..k}}$	rcd-r-c-d+2

A construção da tabela 4.5 baseia-se na equivalência entre a hipótese $H_2(R\times C\times D)$ e a interseção das hipóteses $H_2(R\times C)$ e $H_2(R\times CD)$, onde $H_2(R\times C)$ testa a hipótese $H_{241}: p_{ijk} = p_{.j.}p_{..k}$ e $H_2(R\times CD)$ testa a hipótese $H_{242}: p_{ijk} = p_{i..}p_{.jk}$.

É claro que $H_2(R\times C\times D) \iff H_2(C\times RD) \cap H_2(R\times D)$ e $H_2(R\times C\times D) \iff H_2(RC\times D) \cap H_2(R\times C)$.

EXEMPLO 11 - Consideremos a produção defeituosa de 4 tipos de máquinas, A,B,C e D, classificada segundo dois tipos de defeitos, D_1 e D_2, por meio de dois testes, T_1 e T_2, como mostra a Tabela 4.5 (a).

Queremos saber se a produção defeituosa independe do tipo de máquina, tipo do teste e tipo de defeito, isto é,

$$\begin{cases} H_2: p_{ijk} = p_{i..}p_{.j.}p_{..k} \text{ para todo } (i,j,k) \\ H_1: p_{ijk} \neq p_{i..}p_{.j.}p_{..k} \text{ para algum}(i,j,k) \end{cases}$$

$i=1,2,3,4;\ j=1,2$ e $k=1,2$.

TABELA 4.5 (a)

	A		B		C		D		TOTAL
	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2	
T_1	24	8	7	13	7	2	5	2	68
T_2	11	13	2	8	7	7	3	5	56
									124

Na tabela 4.4 vimos que a estatística para testar a hipótese acima é dada por:

$$\hat{I}(H_1:H_2) = 2\sum \sum \sum x_{ijk} \log \frac{n^2 x_{ijk}}{x_{i..}x_{.j.}x_{..k}}$$

que por simplicidade de cálculo pode ser expressa como:

$$\hat{I}(H_1:H_2) = \sum_i \sum_j \sum_k x_{ijk} \log x_{ijk} - \sum_i x_{i..} \log x_{i..} -$$

$$- \sum_j x_{.j.} \log x_{.j.} - \sum_k x_{..k} \log x_{..k} + 2n \log n.$$

Façamos $i = A,B,C$ e D; $j = T_1,T_2$ e $k = D_1,D_2$. De acordo com os valores observados na Tabela 4.5 (a), temos:

$$2\hat{I}(H_1:H_2) = 23,018.$$

Comparando este resultado com o valor crítico de χ^2 com 10 graus de liberdade vemos que a hipótese de independência é rejeitada ao nível de significância $\alpha = 0,05$. Então podemos decompor $\hat{I}(H_1:H_2)$ em componentes, como mostra a ta-

bela 4.5, para analisarmos com mais detalhe a rejeição da hipótese. Dos dados obtemos que

$$H_2 \text{ (Teste} \times \text{Defeito)} = 6{,}100$$

$$H_1 \text{ (Máquinas} \times \text{(Teste,Defeito))} = 16{,}918$$

Estas duas últimas componentes testam as hipóteses

$$H_{21} \quad p_{.jk} = p_{.j}.p_{..k}$$

$$H_{22} \quad p_{ijk} = p_{i..}p_{.jk},$$

respectivamente.

Para tornar mais clara a análise vamos apresentar os resultados obtidos, na tabela 4.5 (b)

TABELA 4.5 (b)

COMPONENTE	INFORMAÇÃO	G.L.
H_2 (Teste × Defeito)	6,100	1
H_2 (Máquinas × (Teste,Defeito))	16,918	9
$2\hat{I}(H_1:H_2)$	23,018	10

Comparando os valores observados, 6,100 e 16,918 com os respectivos valores do χ^2 crítico, 3,84 e 16,92, ao nível de significância $\alpha = 0{,}05$ temos que:

a) a hipótese de independência entre teste e defeito é rejeitada.

TABELA 4.6

COMPONENTE	INFORMAÇÃO	G.L.
$H_2(R\|K)$	$2\sum_{i=1}^{r} x_{i\cdot k} \log \dfrac{x_{i\cdot k}}{x_{\cdot\cdot k}(p^0_{i\cdot k}/p^0_{\cdot\cdot k})}$	r-1
$H_2(C\|K)$	$2\sum_{j=1}^{i} x_{\cdot jk} \log \dfrac{x_{\cdot jk}}{x_{\cdot\cdot k}(p^0_{\cdot jk}/p^0_{\cdot\cdot k})}$	c-1
$H_2((R\|K)\times(C\|K))$	$2\sum_i \sum_j x_{ijk} \log \dfrac{x_{ijk}}{x_{i\cdot k}x_{\cdot jk}/x_{\cdot\cdot k}}$	(r-1)(c-1)
$2\hat{I}(^*:2;O_n)$	$2\sum_i \sum_j x_{ijk} \log \dfrac{x_{ijk}}{x_{\cdot\cdot k}(p^0_{i\cdot k}/p^0_{\cdot\cdot k})(p^0_{\cdot jk}/p^0_{\cdot\cdot k})}$	r_{c-1}

4.4 - HOMOGENEIDADE ENTRE TABELAS DE CONTINGÊNCIA DE DUPLA ENTRADA

4.4.1 - Homogeneidade Total

Suponhamos que seja de interesse o teste de homogeneidade de r tabelas de contingência de ordem c×d, em que o total de cada tabela seja fixado. Esta hipótese é formulada por:

$$\begin{cases} H_2: p_{ijk} = p^0_{jk} & i=1,2,\ldots,r;\ j=1,2,\ldots,c \\ & \quad e\ k=1,2,\ldots,d. \\ H_1: p_{ijk} \neq p^0_{jk} & \text{para algum } (i,j,k). \end{cases}$$

Considerando que a população de uma tabela de contingência pode ser tratada como uma população multinomial

com rc categorias, o problema será resolvido testando a homogeneidade entre r dessas populações, o que é semelhante ao teste de homogeneidade na seção 3.3 do capítulo III. Deste modo a análise da hipótese em estudo está resumida na tabela 4.7.

TABELA 4.7

COMPONENTE	INFORMAÇÃO	G.L.
Entre	$2\sum_{j=1}^{c}\sum_{k=1}^{d} x_{.jk} \log \frac{x_{.jk}}{np_{jk}^{0}}$	cd-1
$2\hat{I}(H_1:H_{22})$	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{nx_{ijk}}{x_{i..}x_{.jk}}$	(r-1)(cd-1)
TOTAL $2\hat{I}(^*:2;O_n)$	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{x_{i..}p_{jk}^{0}}$	r(cd-1)

onde a componente ENTRE testa a hipótese $H_{21}: p_{.jk} = p_{jk}^{0}$ e a componente $2\hat{I}(H_1:H_{22})$ testa a hipótese H_{22}: $p_{ijk} = p_{jk}$.

4.4.2 - Homogeneidade Condicional

Consideremos r tabelas de contingência de dupla entrada, como na seção anterior. O problema agora é testar a homogeneidade entre as r amostras do atributo D para uma dada categoria do atributo C.

Esta hipótese é formulada por:

$$\begin{cases} H_2: \ p_{ijk} = p^0_{jk} & \begin{matrix} i=1,2,\ldots,r \\ k=1,2,\ldots,d \end{matrix} \quad j \text{ fixo.} \\ \\ H_1: \ p_{ijk} \neq p^0_{jk} \end{cases}$$

O problema é o mesmo de testar a homogeneidade entre r populações multinomiais com d categorias.

Portanto, a construção da tabela para análise desta hipótese é análoga à construção da tabela 3.3.

TABELA 4.8

COMPONENTE	INFORMAÇÃO	G.L.
ENTRE	$2\sum_{k=1}^{d} x_{.jk} \log \dfrac{x_{.jk}}{x_{.j.}(p^0_{jk}/p^0_{.j.})}$	d-1
$2\hat{I}(H_1:H_{22})$	$2\sum_{i=1}^{r}\sum_{j=1}^{d} x_{ijk} \log \dfrac{x_{ijk}}{x_{ij.}(x_{.jk})/x_{.j.}}$	(r-1)(d-1)
TOTAL $2\hat{I}(*:2;0_n)$	$2\sum_{i=1}^{r}\sum_{k=1}^{d} x_{ijk} \log \dfrac{x_{ijk}}{x_{ij.}(p^0_{jk}/p^0_{.j.})}$	r(d-1)

A componente $2\hat{I}(H_1:H_{22})$ testa a hipótese $H_{22}: p_{ijk} = \dfrac{p_{ij.}p_{jk}}{p_{.j.}}$ e a componente ENTRE testa a hipótese

$$H_{21}: \frac{p_{jk}}{p^0_{.j.}} = \frac{p^0_{jk}}{p^0_{.j.}}.$$

A componente $2\hat{I}(H_1:H_{22})$ da tabela 4.7 pode ser analisada em componentes aditivas como mostra a Tabela 4.9.

TABELA 4.9

COMPONENTE	INFORMAÇÃO	G.L.
Homogeneidade (C)	$2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij.} \log \frac{nx_{ij.}}{x_{i..}x_{.j.}}$	(r-1)(c-1)
Homogeneidade condicional (D/C)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ijk} \log \frac{x_{ijk}}{(x_{ij.}x_{.jk})/x_{.j.}}$	c(r-1)(d-1)
Homogeneidade (C,D)	$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{d} x_{ij} \log \frac{nx_{ijk}}{x_{i...}x_{.jk}}$	(r-1)(cd-1)

EXEMPLO: Os dados da Tabela 4.10 apresentam a distribuição de 1937 moscas por sexo e mortalidade entre 12 aplicações sucessivas de um inseticida padrão (Kullback, 1968).

Nosso objetivo é testar a hipótese de homogeneidade de sexo e mortalidade entre as 12 aplicações do inseticida, isto é,

$$\begin{cases} H_2: p_{ijk} = p_{jk} \text{ para todo } (ijk) \\ H_1: p_{ijk} \neq p_{jk} \text{ para algum } (ijk) \end{cases}$$

$$i=1,2,\ldots,12;\ j=1,2 \text{ e } k = 1,2.$$

TABELA 4.10

MACHOS				FÊMEAS		
NÍVEL	VIVO	MORTO	TOTAL	VIVA	MORTA	TOTAL
1	17	40	57	46	6	52
2	14	44	58	44	5	49
3	19	42	61	48	5	53
4	21	23	54	41	4	45
5	9	39	48	68	8	76
6	21	38	59	70	5	75
7	19	40	59	56	4	60
8	15	32	47	51	8	59
9	20	35	55	73	9	82
10	15	29	44	78	5	83
11	12	19	31	69	2	71
12	12	29	41	75	3	78
TOTAL	194	420	614	719	64	783

De acordo com a tabela 4.9 a Estatística para testar esta hipótese é dada por:

$$2 \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{d} x_{ijk} \log \frac{n x_{ijk}}{x_{i..} x_{.jk}}$$

O valor desta Estatística para os dados da tabela é 57.044 que comparado com o valor crítico do $\chi^2_{33},(55)$, ao nível de 0,01 nos leva a rejeitar a hipótese de homogeneidade. Então empregaremos a partição da tabela 4.9 para anali-

sarmos com mais detalhes esta hipótese. Os resultados obtidos estão apresentados na tabela 4.11.

TABELA 4.11

COMPONENTE	INFORMAÇÃO	G.L.
Homogeneidade (sexo)	36.874	11
Homogeneidade condicional de mortalidade dado sexo	20.170 Macho 8.906 Fêmea 11.264	22 11 11
Homogeneidade (sexo, mortalidade)	57.044	33

Os valores críticos do χ^2 com 11 e 22 graus de liberdade, ao nível de 0,01, são respectivamente, 24,72 e 40,29. Comprando estes valores com os valores observados das estatísticas concluímos que:

a) a hipótese de homogeneidade de sexo é rejeitada.

b) a hipótese de homogeneidade de mortalidade dado o sexo não é rejeitada.

- ¤ -

CAPÍTULO V

TABELAS DE CONTINGÊNCIA INCOMPLETAS

Neste capítulo faremos uma exposição de vários exemplos encontrados na literatura e dos diferentes métodos de análise indicados para cada caso. Serão ilustrados os três seguintes métodos de análise:

1) um método baseado na partição da Tabela Incompleta em subtabelas apropriadas.
2) um método baseado na estimação das frequências esperadas na Tabela Incompleta calculadas sob a suposição de quase-independência.
3) um método baseado na estimativa de certas interações na Tabela Incompleta.

5.1 - DEFINIÇÕES PRELIMINARES

Consideremos uma Tabela de Contingência de ordem $r \times c$. Seja p_{ij} a probabilidade de ocorrência da casela (i,j), $i=1,2,\ldots,r$ e $j=1,2,\ldots,c$.

DEFINIÇÃO 1 - As classificações linha e coluna de uma Tabe-

la de Contingência de ordem r×c são ditas independentes se

$$p_{ij} = a_i b_j \qquad i=1,2,\ldots,r \text{ e } j=1,2,\ldots,c,$$

onde a_i e b_j são constantes positivas.

Se $p_{ij} = a_i b_j$ para toda casela (i,j) e $\sum a_i = 1$ e $\sum b_j = 1$, então a_i e b_j são as probabilidades de ocorrência da i-ésima linha e j-ésima coluna da Tabela, respectivamente.

A definição 1 aplica-se a todas as caselas da Tabela r×c. Consideremos agora um dado subconjunto S das r×c caselas, por exemplo, todas as caselas exceto as da diagonal principal ou todas as caselas exceto as acima da diagonal principal ou todas as caselas exceto uma particular, etc.

DEFINIÇÃO 2 - As classificações linha e coluna do subconjunto S são ditas quase-independentes (ou S-independentes) se $p_{ij} = a_i b_j$ para toda casela (i,j) em S, onde a_i e b_j, i=1,2,...,r e j=1,2,...,c, são constantes positivas.

Este conceito de quase-independência é usado não só para a análise de Tabelas de Contingência Incompletas como também para a análise de Tabelas de Contigência Completa desde que seja de interesse uma análise mais detalhada da associação entre as classificações linha e coluna.

A hipótese H_2: $p_{ijk} = a_i b_j$ para toda casela (i,j),

$i=1,2,\ldots,r$; $j=1,2,\ldots,c$ é equivalente à interseção das seguintes subhipóteses:

1) H_{21}: $p_{ij} = a_i b_j$ para toda casela (i,j) em um subconjunto S da Tabela Completa $r \times c$,

2) H_{22}: $p_{ij} = a_i b_j$ para toda casela (i,j) pertencente ao complementar de S em relação à tabela $r \times c$, onde a_i e b_j são as mesmas constantes que tornam a hipótese H_{21} verdadeira.

A maioria dos métodos que apresentaremos pode ser aplicado também à análise de quase homogeneidade entre r populações multinomiais. Para entendermos melhor esta extensão considederemos uma Tabela de Contingência $r \times c$, onde p_{ij} representa a probabilidade conjunta da i-ésima linha e j-ésima coluna.

Se imaginarmos esta tabela como sendo um conjunto de r amostras de populações multinomiais com c categorias, a probabilidade de ocorrência da j-ésima categoria, $j=1,2,\ldots,c$ da i-ésima população será dada por:

$$(5.1) \qquad P_{ij} = \frac{p_{ij}}{p_{i.}} \text{ onde } p_{i.} = \sum_j p_{ij} \text{ e } \sum_j P_{ij} = 1, \quad i=1,2,\ldots,r.$$

quando as classificações linha e coluna forem independentes temos:

$$P_{ij} = \frac{b_j}{\sum_{k=1}^{c} b_k}, \qquad i=1,2,\ldots,r.$$

DEFINIÇÃO 3 - Dizemos que r populações multinomiais são homogêneas se

$$P_{ij} = \frac{b_j}{\sum_{k=1}^{c} b_k} \text{ para } i=1,2,\ldots,r.$$

No caso em que $\sum_{k=1}^{c} b_k = 1$, b_j será interpretado como a probabilidade de ocorrência da j-ésima categoria da i-ésima população.

Suponhamos agora que as classificações linha e coluna de uma tabela r×c sejam independentes mas as probabilidades de ocorrência de certas caselas não possam ser averiguadas. Considerando somente o subconjunto S de caselas em que as probabilidades de ocorrência possam ser averiguadas (fazendo vazias as caselas que não estejam em S) temos de (5.1) e da definição 3 que:

$$P_{ij} = \frac{\delta_{ij} b_j}{\sum_{k=1}^{c} \delta_{ik} b_k}, \qquad i=1,2,\ldots,r$$

onde

$$\delta_{ij} = \begin{cases} 1 \text{ para toda casela } (i,j) \text{ em } S \\ 0 \text{ caso contrário.} \end{cases}$$

Observemos que quando as linhas e colunas de S forem quase-independentes e $\sum_{j=1}^{c} b_j = 1$, b_j será a probabilidade de ocorrência da j-ésima categoria da i-ésima população na situação hipotética em que nenhuma das categorias na i-ésima população seja vazia e a probabilidade de ocorrência da j-ésima categoria seja a mesma para cada uma das r populações.

DEFINIÇÃO 4 - Dizemos que r populações multinomiais são quase-homogêneas com relação a S se

$$P_{ij} = \frac{\delta_{ij} b_j}{\sum_{k=1}^{c} \delta_{ik} b_k}, \qquad i=1,2,\ldots,r$$

onde

$$\delta_{ij} = \begin{cases} 1 \text{ para toda casela } (i,j) \in S \\ 0 \text{ caso contrário.} \end{cases}$$

Este procedimento pode ser empregado não só para testar a hipótese de quase-homogeneidade entre as r populações multinomiais como também para estimar as proporções hipotéticas b_j, $j=1,2,\ldots,c$. No caso de quase-homogeneidade, o

valor de b_j, $j=1,2,\ldots,c$ é estimativa da probabilidade de ocorrência da j-ésima categoria sob a suposição de homogeneidade e de que nenhuma das categorias sejam vazias. Este mesmo argumento vale para o caso de independência.

Os métodos de análise que apresentaremos são válidos nos seguintes casos:

1) quando as linhas e colunas não são fixas
2) quando somente as linhas marginais são fixas
3) quando somente as colunas marginais são fixas.

Na próxima seção analisaremos alguns tipos de Tabelas Incompletas por meio da Estatística da Razão de Verossimilhança.

5.2 - ANÁLISE DE ALGUMAS TABELAS DE CONTINGÊNCIA INCOMPLETAS

Vamos considerar dois tipos de Tabelas de Contingência Incompletas a saber:

1) Tabelas que consistem de dois subconjuntos separáveis, S_1 e S_2, em que S_1 e S_2 não tenham linhas e nem colunas em comum.
2) Tabelas que não consistem de subconjuntos separáveis.

5.2.1 - Tabelas Separáveis

Consideremos a Tabela 5.1 em que a metade das case

las são vazias por considerações a priori, desde que para os coeficientes de simetria radial 0,47, 0,82, 1,25, 1,41 e 1,70 a primeira e quarta composição locular não possam ocorrer e para os coeficientes de simetria radial 0,00, 0,94, 1,63 e 1,89 a segunda e terceira composição locular também não possam ocorrer. Esta tabela foi apresentada por Harris (1929).

TABELA 5.1 - RELAÇÃO ENTRE SIMETRIA RADIAL E COMPOSIÇÃO LOCULAR (SÉRIE A)

COMPOSIÇÃO LOCULAR	COEFICIENTE DE SIMETRIA RADIAL								
	0,00	0,47	0,82	0,94	1,25	1,41	1,63	1,70	1,89
3 pares - 0 ímpar	462	-	-	130	-	-	2	-	1
2 pares - 1 ímpar	-	614	138	-	21	14	-	1	-
1 par - 2 ímpares	-	443	95	-	22	8	-	5	-
0 par - 3 ímpares	103	-	-	35	-	-	1	-	0

Pearson (1930) mostrou que a ordem das linhas e/ou colunas é irrelevante e portanto a Tabela 5.1 poderia ser rearranjada como mostra a Tabela 5.2. As colunas correspondentes aos coeficientes de simetria radial 1,63 e 1,89 foram combinados por terem o número de observações muito pequeno.

TABELA 5.2 - RELAÇÃO ENTRE SIMETRIA RADIAL E COMPOSIÇÃO LOCULAR (SÉRIE A)

COMPOSIÇÃO LOCULAR	COEFICIENTE DE SIMETRIA RADIAL							
	0,00	0,94	1,63-1,89	0,47	0,82	1,25	1,41	1,70
3 pares - 0 ímpar	462	130	3	-	-	-	-	-
0 par - 3 ímpares	103	35	1	-	-	-	-	-
2 pares - 1 ímpar	-	-	-	614	138	21	14	1
1 par - 2 ímpares	-	-	-	443	95	22	8	5

A estatística do qui-quadrado de Pearson para o teste de independência é inaplicável, como mostrou Harris (1929), porque algumas caselas são a priori vazias. Contudo, considerando somente o subconjunto S de caselas não vazias, podemos testar a hipótese nula de que as classificações linha e coluna da Tabela 5.2 sejam quase-independentes, isto é,

$$H_2: p_{ij} = a_i b_j \text{ para toda casela } (i,j) \in S$$

$$H_1: p_{ij} \neq a_i b_j \text{ para alguma casela } (i,j) \notin S.$$

Da definição de quase-independência vemos que a hipótese H_2 pode ser decomposta em duas sub-hipóteses:

1) A sub-hipótese H_{21} de que o subconjunto S_1 seja quase-independente, onde S_1 é a sub-tabela de ordem 2×3 formada pelas duas primeiras linhas e as três primeiras colunas da tabela 5.2.

2) A sub-hipótese H_{22} de que o subconjunto S_2 seja qua-

se-independente, onde S_2 é a sub-tabela de ordem 2×5, formada pelas duas últimas linhas e pelas cinco últimas colunas da tabela 5.2.

No caso particular das sub-tabelas S_1 e S_2 a hipótese de quase-independência, sobre a população correspondente a cada uma delas é equivalente à hipótese de independência, isto é,

$$\begin{cases} H_{21}: p_{ij} = a_i b_j \text{ para toda casela } (i,j) \in S_1 \\ H_{11}: p_{ij} \neq a_i b_j \text{ para alguma casela } (i,j) \notin S_1 \end{cases}$$

e

$$\begin{cases} H_{22}: p_{ij} = a_i b_j \text{ para toda casela } (i,j) \in S_2 \\ H_{12}: p_{ij} \neq a_i b_j \text{ para alguma casela } (i,j) \notin S_2. \end{cases}$$

Sob a hipótese de independência a Estatística da Razão de Verossimilhança, χ^2_{Ld}, pode ser escrita por:

$$\chi^2_{Ld} = 2(\sum_i \sum_j x_{ij} \log x_{ij} - \sum_i x_{i.} \log x_{i.} - \sum_j x_{.j} \log x_{.j} + n \log n).$$

A estatística χ^2_{Ld}, obtida para cada sub-tabela é

$$\chi^2_{L2} = 0,8146 \quad \text{e} \quad \chi^2_{L4} = 6,253.$$

Estas estatísticas são assintoticamente independentes, então a estatística, $\chi^2_{L6} = \chi^2_{L2} + \chi^2_{L4} = 7,0676$, é utilizada para

testar a hipótese de quase-independência entre as classificações linha e coluna da tabela 5.2. Sob a hipótese de quase- independência, as estatísticas, χ^2_{L2}, χ^2_{L4} e χ^2_{L6} tem distribuição limite de χ^2 com 2, 4 e 6 graus de liberdade respectivamente. Portanto a hipótese de quase-independência é não rejeitada.

5.2.2 - Tabelas com uma Casela Vazia

Para a análise de tabelas de contingência com uma casela vazia empregaremos dois métodos propostos por Goodman (1968). Ilustraremos estes métodos analisando a Tabela 5.3 abaixo.

TABELA 5.3 - CLASSIFICAÇÃO CRUZADA DE UMA AMOSTRA DE HOMENS DE ACORDO COM O "STATUS" DO FILHO E DO SEU PAI

"STATUS" DO PAI	"STATUS" DO FILHO		
	ELEVADO	MÉDIO	BAIXO
ELEVADO	588	395	159
MÉDIO	349	714	447
BAIXO	114	320	411

Inicialmente vamos testar a hipótese de independência na tabela completa 5.3. Neste caso a Estatística da Razão de Verossimilhança é $\chi^2_{L4} = 499,6$, o que mostra que as classificações linha e coluna são fortemente dependentes.

Ref: Jesse, Nunesmaia, Maria das Mercês Santana

Análise de dados categóricos usando Teoria da Informação

Teoria da Informação
Teoria da Comunicação

(?) $P[H_i/x] = \dfrac{P[H_i]\, f_i(x)}{P[H_1] f_1(x) + P(H_2) f_2(x)}$, $i = 1, 2$ (Rever T. Bayes)

$P(H_i)$: probabilidade a priori;
$P[H_i/x]$: probabilidade a posteriori;
(?) rever propriedades de log;
(?) rever multipl. Lagrange;

O significado das categorias envolvidas nos critérios de classificação da Tabela 5.3 sugere que,a primeira casela seja a responsável por tão alto grau de dependência entre as duas classificações. A fim de verificar esta suspeita seria interessante testar a hipótese de quase-independência na tabela imcompleta 5.4 obtida da tabela 5.3, excluindo a primeira casela.

TABELA 5.4

"STATUS" DO PAI	"STATUS" DO FILHO		
	ELEVADO	MÉDIO	BAIXO
ELEVADO	-	395	159
MÉDIO	349	714	447
BAIXO	114	320	411

O primeiro método de análise consiste em decompor a tabela 5.3 em subtabelas convinientesde acordo com o método descrito na seção 3.3 do capítulo 3 (Tabela 3.4), e excluir de considerações a tabela que contiver a casela vazia, no caso a tabela 5.3.1.

5.3.1

	ELEVADO	NÃO ELEVADO
ELEVADO	588	554
NÃO ELEVADO	463	1.892

5.3.2

	MÉDIO	BAIXO
MÉDIO	714	447
BAIXO	320	411

5.3.3

	ELEVADO	NÃO ELEVADO
MÉDIO	349	1.161
BAIXO	114	731

5.3.4

	MÉDIO	BAIXO
ELEVADO	395	159
NÃO ELEVADO	1.034	858

A hipótese de quase-independência na tabela 5.4 é equivalente à interseção das sub-hipóteses H_{2i}, i=2,3,4, onde H_{2i} é a hipótese de independência correspondente à tabela 5.3.i, i=2,3,4.

As estatísticas χ^2_{L1}, utilizadas para testar a hipótese de independência nas subtabelas 5.3.2, 5.3.3 e 5.3.4 têm distribuição limite de χ^2_1 e são assintoticamente independentes. Assim a Estatística χ^2_{L3}, obtida da soma das três anteriores tem distribuição limite de χ^2_3 e testa a hipótese de quase-independência na tabela 5.4.

Notemos que o número de graus de liberdade da estatística para testar a hipótese de quase-independência na tabela 5.4 é igual à diferença entre o número de graus de liberdade para testar a independência na tabela completa e o número de caselas vazias.

Os valores observados das estatísticas χ^2_{L1} correspondentes às tabelas 5.3.2, 5.3.3 e 5.3.4 são 56,9, 33,2 e 50,5, respectivamente. Portanto a soma $\chi^2_{L3} = 140,6$, o que nos

leva a rejeitar a hipótese de quase independência ao nível de 0,01%.

O segundo método para testar a hipótese de quase-independência na tabela 5.4 é baseado na partição desta tabela da seguinte forma:

-	X
Y	Z

onde: X é o vetor linha $[x_1, x_2]$

Y é o vetor coluna $[y_1, y_2]$

Z é a tabela $[z_{ij}]$ de ordem 2×2

Sejam

$$x_. = \sum_{j=1}^{2} x_j, \quad y_. = \sum_{i=1}^{2} y_i, \quad z_{i.} = \sum_{j=1}^{2} z_{ij}$$

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

Neste caso os estimadores das frequências esperadas, calculadas sob à hipótese de quase-independência, são dadas por:

$$(5.2.1) \quad \hat{x}_j = \frac{x_.(x_j+z_{.j})}{x_.+z_{..}}, \quad \hat{y}_i = \frac{y_.(y_i+z_{i.})}{y_.+z_{..}} \quad \text{e } \hat{z}_{ij} = \frac{(y_i+z_{i.})(x_j+z_{.j})}{(y_.+z_{..})(x_.+z_{..})}.$$

As expressões (5.2.1) são estimadores de máxima verossimilhança das frequências esperadas, sob à hipótese de quase-independência (Goodman, 1968). Com estes estimadores obtemos a Estatística

$$\chi^2_{Ld} = 2\sum_i \sum_j x_{ij} \log \frac{x_{ij}}{\hat{x}_{ij}}$$

que, sob a hipótese de quase-independência tem distribuição limite de χ^2 com 4 - 1 = 3 graus de liberdade. O valor observado desta Estatística, para testar a hipótese de independência na Tabela 5.3 é $\chi^2_{L3} = 140,6$, a qual coincide com o χ^2_{L3} calculado pelo primeiro método.

5.2.3 - Tabelas com um Bloco de Caselas Vazias

Consideremos agora o caso em que temos um bloco retangular de caselas vazias. Neste caso podemos apresentar a Tabela da seguinte forma:

TABELA 5.3

-	X
Y	Z

onde $X = \{x_{ij}\}$, $Y = \{y_{ij}\}$ e $Z = \{z_{ij}\}$ são Tabelas de ordem $r' \times c$, $r \times c'$ e $r \times c$, respectivamente.

Sejam,

$$x_{i.} = \sum_i x_{ij}, \quad x_{.j} = \sum_j x_{ij}, \quad x_{..} = \sum_i \sum_j x_{ij}$$

$$z_{*.} = (z_{1.}, z_{2.}, \ldots, z_{r.})' \text{ vetor coluna}$$

$$z_{.*} = (z_{.1}, z_{.2}, \ldots, z_{.r}) \text{ vetor linha.}$$

A hipótese de quase-independência sobre a população da tabela 5.5 pode ser decomposta em sub-hipóteses de independência sobre as populações correspondentes às sub-tabelas 5.5.1, 5.5.2 e 5.5.3.

TABELA 5.5.1

z_{11}	z_{12}	...	z_{1c}
z_{21}	z_{22}	...	z_{2c}
$\vdots$	$\vdots$	...	$\vdots$
z_{r1}	z_{r2}	...	z_{rc}

TABELA 5.5.2

Y	$Z_{*.}$

TABELA 5.5.3

X
$Z_{.*}$

Sob a hipótese de independência as estatísticas χ^2_{Ld}, correspondentes às tabelas acima tem distribuição limite de χ^2_d com d igual a

$$(r-1)(c-1), \quad (r-1)c' \quad \text{e} \quad r'(c-1)$$

respectivamente e são assintoticamente independentes. A soma dessas estatísticas tem distribuição limite de $\chi^2_{d'}$, onde $d' = (r-1)(c-1) + (r-1)c' + r'(c-1)$ e testa a hipótese de quase-independência na Tabela 5.5.

O segundo método para testar a hipótese de quase-independência na tabela 5.5 consiste em determinar o valor da estatística $\chi^2_{L_d}$, obtida ao compararmos as frequências esperadas e observadas. Neste caso precisamos conhecer as estimativas das frequências esperadas, as quais são obtidas através das relações:

$$\hat{x}_{ij} = \frac{x_{i.}(x_{.j}+z_{.j})}{x_{..}+z_{..}}, \quad \hat{y}_{ij} = \frac{y_{.j}(y_{i.}+z_{i.})}{y_{..}+z_{..}}$$

$$\text{e} \quad \hat{z}_{ij} = \frac{(y_{i.}+z_{i.})(x_{.j}+z_{.j})z_{..}}{(y_{..}+z_{..})(x_{..}+z_{..})}.$$

No caso em que várias entradas da tabela (não necessariamente um bloco delas) em uma única linha (ou única coluna) são vazias, os métodos apresentados acima podem ser aplicados desde que as colunas (ou linhas) possam ser rearranjadas de tal modo que as caselas vazias na linha ou coluna formem um bloco retangular.

5.2.4 - Tabelas com Caselas Vazias na Mesma Linha e/ou Mesma Coluna

Consideremos o caso em que uma particular casela da tabela seja excluída e assim várias outras caselas na mesma linha (e/ou mesma coluna). Neste caso a tabela pode ser escrita da seguinte forma:

TABELA 5.6

-	-	U
-	V	X
W	Y	Z

onde $U = \{u_{ij}\}$, $X = \{x_{ij}\}$, $Z = \{z_{ij}\}$, $V = \{v_{ij}\}$, $Y = \{y_{ij}\}$ e $W = \{w_{ij}\}$ são matrizes de ordem $r''\times c$, $r'\times c$, $r\times c$, $r'\times c'$, $r\times c'$ e $r\times c''$ respectivamente.

Sejam, $x_{i.}$, $x_{.j}$, $x_{..}$, $x_{.*}$, $x_{*.}$, etc. definidas como antes.

A hipótese de quase-independência na tabela 5.6 pode ser particionada em sub-hipóteses de independência nas subtabelas.

TABELA 5.6.1

V	X
Y	Z

TABELA 5.6.2

W	$y_{*.}+z_{*.}$

TABELA 5.6.3

U
$X_{.*}+Z_{.*}$

Os métodos para testar a hipótese de quase-independência na tabela 5.6 são os mesmos descritos nas seções anteriores.

Até aqui tratamos de dois métodos para resolver o problema de testar a hipótese de quase-independência em al-

guns tipos de tabelas. Na próxima seção abordaremos o problema de estimar interações em tabelas incompletas.

5.2.5 - Tabela Triangular

Consideremos a Tabela 5.7 em que as caselas acima da diagonal principal são vazias. Esta tabela apresenta a distribuição de uma amostra de impressão digital da mão direita, classificada segundo o número de verticilos e de pequenos arcos.

TABELA 5.7

VERTICILOS	PEQUENOS ARCOS					
	0	1	2	3	4	5
5	50	-	-	-	-	-
4	104	26	-	-	-	-
3	125	38	7	-	-	-
2	130	92	55	15	-	-
1	106	153	126	80	32	-
0	78	144	204	211	179	45

Para testar a hipótese de quase-independência na tabela 5.7, consideremos primeiro a partição desta tabela nas subtabelas, 5.7.1, 5.7.2, 5.7.3 e 5.7.4.

TABELA 5.7.1

VERTICILOS	PEQUENOS ARCOS				
	0	1	2	3	4
1	106	153	126	80	32
0	78	144	204	211	179

TABELA 5.7.2

VERTICILOS	PEQUENOS ARCOS			
	0	1	2	3
2	130	92	55	15
Menos de 2	184	297	330	291

TABELA 5.7.3

VERTICILOS	PEQUENOS ARCOS		
	0	1	2
3	125	38	7
Menos de 3	314	389	385

TABELA 5.7.4

VERTICILOS	PEQUENOS ARCOS	
	0	1
4	104	26
Menos de 4	439	427

Os valores das estatísticas, χ^2_{Ld}, para testar a hipótese de independência nas sub-tabelas 5.7.1, 5.7.2, 5.7.3 e 5.7.4 são:

$\chi^2_{L4} = 118,9376$; $\chi^2_{L3} = 143,8936$; $\chi^2_{L2} = 145,3245$ e $\chi^2_{L1} = 42,1378$, respectivamente. A adição dessas estatísticas produz o resultado $\chi^2_{L10} = 450,2332$. Neste caso a hipótese de quase-independência é rejeitada ao nível de 0,01%-

A hipótese de quase-independência na tabela 5.7 pode, também, ser testada utilizando a Estatística da Razão da Verossimilhança obtida quando comparamos frequências esperadas e observadas. Neste caso as frequências esperadas são estimadas através das relações:

$$\hat{x}_{ij} = \frac{x_{i.}}{y_i} \cdot x_{j-1} y_{j-1}, \qquad i=1,2,\ldots,5 \text{ e } j=2,3,\ldots,5 \text{ e } \hat{x}_{i1} = \frac{x_{i.}}{y_i}$$

onde

$$x_{i.} = \sum_j x_{ij}, \quad x_j = \frac{x_{.j+1}}{\sum_{j=1}^{c} (x_{.j} - x_{j.})}, \quad y_{i+1} = \prod_{j=1}^{i} (1+x_j) \text{ e } y_1 = 1, \; x_0 y_0 = 1.$$

para $i < j$ $\hat{x}_{ij} = 0$.

As estatísticas obtidas pelos, primeiro e segundo métodos são assintoticamente equivalentes.

5.3 - INTERAÇÃO

A hipótese de quase-independência em uma tabela incompleta pode ser testada, também, através do conceito de interação que será desenvolvido nesta seção. Este método tem vantagens sobre os métodos anteriores no sentido de que ele resolve o problema de testar a hipótese de quase-independência em tipos de tabelas para as quais os processos anteriores não se aplicam. Além disso as interações fornecem medidas de dependência entre as categorias envolvidas. Portanto quando rejeitamos uma hipótese de quase-independência, estimativas das interações nos dão idéia do grau de dependência entre as categorias.

Para ilustrarmos o conceito de interação considere

mos p_{ij} como a probabilidade de ocorrência da i-ésima linha e j-ésima coluna em uma tabela completa 3×3. Sob a hipótese de independência nesta tabela temos que:

$$\frac{P_{22}P_{33}}{P_{32}P_{23}} = 1, \tag{5.3.1}$$

tomando o logaritmo natural de ambos os membros de (5.3.1) temos:

$$\log P_{22} + \log P_{33} - \log P_{32} - \log P_{23} = 0. \tag{5.3.2}$$

Representemos por L o primeiro membro de (5.3.2) ou seja

$$L = \log P_{22} + \log P_{33} - \log P_{32} - \log P_{23}. \tag{5.3.3}$$

Neste caso L é uma interação em uma tabela 3×3. Este conceito pode ser generalizado para tabelas r×c através da definição abaixo.

<u>DEFINIÇÃO 1</u> - A função $L(\alpha) = \sum_{i=1}^{r} \sum_{j=1}^{c} \alpha_{ij} \log P_{ij}$ é definida como uma interação de primeira ordem numa tabela de ordem r×c, onde $\sum_i \alpha_{ij} = \sum_j \alpha_{ij} = 0$ e $\alpha_{ij} \neq 0$ para alguma casela (i,j), i=1,2,...,r; j=1,2,...,c.

O estimador de Máxima Verossimilhança de $L(\alpha)$ é dado por:

$$\hat{L}(\alpha) = \sum_i \sum_j \alpha_{ij} \log x_{ij} \tag{5.3.4}$$

onde x_{ij} é a frequência observada da casela (i,j).

A variância de $\hat{L}(\alpha)$ é estimada por:

$$S^2(\alpha) = \sum_i \sum_j \alpha_{ij}^2 g_{ij} \tag{5.3.5}$$

onde $g_{ij} = \frac{1}{x_{ij}}$.

Sob à hipótese de independência em uma tabela completa r×c a estatística $\frac{\hat{L}^2(\alpha)}{S^2(\alpha)}$ tem distribuição de χ_d^2 onde $d = (r-1)(c-1)$. Podemos então testar hipóteses do tipo H_2: $L(\alpha) = 0$ ou construir intervalos de confiança para as interações os quais nos dão uma medida de associação entre as classificações.

Para tabelas incompletas o procedimento é semelhante, com as seguintes observações:

1) Para as caselas vazias, $\alpha_{ij} = 0$.

2) Os graus de liberdade da estatística $\frac{\hat{L}(\alpha)}{S(\alpha)}$ é igual ao anterior diminuído do número de caselas vazias na tabela.

Com o exposto acima podemos tratar agora do método de testar quase-independência por meio de interação. Para isto consideremos a Tabela 5.3, excluindo as caselas da diagonal, como mostra a Tabela 5.8.

TABELA 5.8

"STATUS" DO PAI	"STATUS" DO FILHO		
	ELEVADO	MÉDIO	BAIXO
ELEVADO	-	395	159
MÉDIO	349	-	447
BAIXO	114	320	-

Seja p_{ij} a probabilidade de ocorrência da casela (i,j), $i \neq j$. A hipótese de quase-independência na tabela 5.8 pode ser expressa pela equação

$$\frac{P_{12}P_{32}P_{31}}{P_{13}P_{32}P_{21}} = 1$$

a qual pode também ser escrita por $L(\alpha) = 0$, onde

$$L(\alpha) = \log P_{23} + \log P_{31} - \log P_{13} - \log P_{32} \log P_{21}.$$

De (5.3.4) obtemos $\hat{L}(\alpha) = 0,125$ e de (5.3.5)

$$S^2(\alpha) = 0,0258.$$

Neste caso a hipótese de independência é testada por $\chi_1^2 = \frac{\hat{L}^2(\alpha)}{S^2(\alpha)} = 0,6$, a qual não é rejeitada ao nível de 0,05.

Os métodos desenvolvidos na seção anterior não se aplicam para tabelas do tipo 5.7.

Goodman (1968) desenvolveu um método geral para es

timar as frequências esperadas em uma Tabela Incompleta, envolvendo processos iterativos, muitas vezes, complicados. Considerando que, os valores obtidos pelo processo iterativo são os mesmos obtidos pelos métodos descritos neste capítulo, não trataremos do método iterativo aqui.

- ¤ -

BIBLIOGRAFIA

Bishop, I.M.M. and Fienberg, S.E. - Incomplete Two-Dimensional contingency tables, *Biometrics*, *25* (1969), 118-128.

Chen, T. and Fienberg, S.E. - Two-Dimensional contingency tables with both completely and partially cross - classified data, *Biometrics*, *30* (1974), 629-642.

———————————— - The analysis of contingency tables with incomplete classifieds data, *Biometrics*, *32* (1976), 133-144.

Fienberg, S.E. - Quasi-independency and maximum Likelihood Estimation in Incomplete contingency tables, *Journal of The American Statistical Association*, *65* (1970), 1.610-1.616.

Goodman, L.A. - Simultaneous confidence limits for cross-Product Ratios in contingency tables, *Journal fo the Royal Statistical Society*, Série B, 26 (1963), 86-102.

———————— - The analysis of cross-classified data: Independence, Quasi-indenpendence, and Interations in contingency tables with or without missing entries, *Journal of the American Statistical Association* *63*, (1968), 1.091-1.131.

Grizzle, J.E. and Williams, O.D. - Log Linear Models and tests of independence for contingency tables, *Biometrics*, *66* (1972), 137-156.

Harris, J.A. and Tu, C. - A second category of limitations in the applicability of the Contingency coefficient, *Jornal of the American Statistical Association*, *24* (1929), 367-615.

Hocking, R.R. and Oxspring, H.H. - Maximum Likelihood Estimation with Incomplete Multinomial data, *Journal of the American Statistical Association, 66* (1971), 65-70.

Kullback, S. - Information Theory and Statistics, Dover Publications, Inc. New York, 1968.

Lacaster, H.D. - The Derivation and Partition of χ^2 in certain Discrete Distribution, *Bicmetrika, 36* (1949), 117-129.

Fraser, D.H.S., Nonparametric Methods in Statistics, John Wiley & Sons, New York, (1951).

- ¤ -