TÍTULO DA TESE:
"ANÁLISE DE TABELAS DE CONTINGÊNCIA 2x2. APLICAÇÕES À EPIDEMIOLOGIA"

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Sra. **JULIA MARIA PAVAN** e aprovada pela Comissão Julgadora.

Campinas, 04 de dezembro de 1987

Euclydes Custódio de Lima Filho
Prof. Dr. EUCLYDES CUSTÓDIO DE LIMA FILHO
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de **Mestre em ESTATÍSTICA**

"ANÁLISE DE TABELAS DE CONTINGÊNCIA 2x2 APLICAÇÕES À EPIDEMIOLOGIA"

JÚLIA MARIA PAVAN

UNIVERSIDADE ESTADUAL DE CAMPINAS

UNICAMP

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA DA COMPUTAÇÃO

CAMPINAS

1 9 8 7

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação da Universidade Estadual de Campinas - UNICAMP - como requisito parcial para obtenção do título de Mestre em Estatística.

NOVEMBRO - 1987

Orientador:

Prof. Dr. EUCLYDES CUSTÓDIO DE LIMA FILHO

Aos meus pais
À Dita,
pela vida,
por tudo!

Ao meu Alonso
pelo amor,
pelo presente
e futuro!

Aos meus irmãos,
pelo apoio
que sempre tive!

AGRADEÇO

Ao Professor Euclydes a quem eu muito admiro, pela amizade e orientação brilhante que sempre me dedicou.

Ao Professor Paulo Roberto Curi a quem eu sou eternamente grata pela iniciação nesta carreira, pelo apoio e amizade.

Ao Professor Dalton Francisco de Andrade, pelo entusiasmo e incentivo que me foram dados.

A todos os meus mestres que de uma forma ou de outra são responsáveis pela minha formação.

A tia Rosa e a Cila pelo carinho e atenção com que me acolheram.

As queridas amigas Lú e Ruth, pela experiência de vida, paciência e presença no meu dia a dia.

A todos os amigos do IMECC, em especial à nossa turma e a Iara, que foram responsáveis pela gratificante convivência durante o curso de mestrado.

A todos os amigos do IPEA - Presidente Prudente pelo apoio recebido.

As Instituições, FAPESP e CNPq, pelo apoio financeiro recebido na realização de meu programa de mestrado.

RESUMO

É realizada uma revisão da teoria estatística apropriada para análise de tabelas de contingência 2x2, dando ênfase à aplicação destes resultados a dados gerados de estudos epidemiológicos. O desenvolvimento teórico explora a noção de condicionar a distribuição da estatística do teste em uma estatística ancilar, o que implica na tão polêmica análise de tabelas de contingência usando marginais fixadas. São apresentadas também outras formulações teóricas como a revisão de vários procedimentos de estimação pontual ou por intervalo do parâmetro razão de produtos cruzados através de expressões exatas empregando métodos numéricos (Cornfield, 1956) ou expressões aproximadas (Cox, 1958; Gart, 1962; Wolf, 1954), considerando métodos condicionais ou não condicionais. A quantidade razão de produtos cruzados se destaca como uma medida de associação superando outras quantidades devido a suas propriedades estatísticas e de aplicações. Três exemplos da área aplicada à Epidemiologia são desenvolvidos caracterizando cada tipo de estudo observacional prospectivo, retrospectivo e transversal e analisando criticamente os intervalos de confiança calculados.

ÍNDICE

pág.

CAPÍTULO I

INTRODUÇÃO

Epidemiologia é o estudo da distribuição e dos fatores determinantes da ocorrência de doenças em grupos populacionais. Nesta definição, estão indicadas as duas principais áreas de investigação: o estudo da distribuição de doenças e a pesquisa dos fatores determinantes da distribuição observada.

A primeira área de investigação, descreve a distribuição do estado de saúde da população em termos de idade, sexo, raça, geografia e outros. Desta forma, alterações na ocorrência da doença definem estratos nos grupos populacionais. A segunda área, envolve a explanação do modelo de distribuição de uma doença em termos de fatores causais.

O conceito de causa é bastante polêmico. Muitos pesquisadores reconhecem a dificuldade e, em alguns casos, a incapacidade de estabelecer um fator como ativo na etiologia de doenças. Certos eventos ou circunstâncias tendem a seguir outros no tempo, mas, é questionável até que ponto estas associações temporais representam associações entre causa e efeito.

Mediante estas dificuldades, em Epidemiologia é apresentada uma proposta prática que define **associação causal** como uma associação entre categorias de eventos em que uma alteração na freqüência ou qualidade de uma categoria é seguida por uma mudança na outra (MacMahon e Pugh, 1970).

Neste sentido, métodos quantitativos em Bioestatística oferecem grande contribuição nas pesquisas epidemiológicas, indicando como um primeiro passo, a verificação de **associação estatística** entre as categorias de eventos. Realmente, quanto maior o grau de associação, mais evidências temos a favor de um possível relacionamento causal. Contudo, muitos fatores podem estar atuando na etiologia da doença, além daquele considerado, ou mesmo, uma associação estatística forte pode resultar do efeito de um terceiro fator. Deste modo, conclusões epidemiológicas sobre a causa de doenças devem ser confirmadas por estudos de natureza clínica, patológica e laboratorial.

Em Epidemiologia, o conhecimento de qualquer tipo de associação entre a doença e fatores relacionados é importante no estabelecimento de medidas preventivas.

Assim, uma das primeiras preocupações dos epidemiologistas no planejamento de uma pesquisa é a identificação de fatores que supostamente alteram a distribuição da doença na população. Estes são definidos como **fatores de risco** para o desenvolvimento da doença. O grupo exposto a tais fatores constitui a **população de risco.** Geralmente, o termo fator de risco ao invés de causa é usado para indicar uma variável que possivelmente esteja relacionada com a probabilidade de um indivíduo desenvolver a doença (Kleinbaum, D.G.; Kupper, L.L.; Morgenstern, H. (1982)).

Nestas investigações, os delineamentos básicos uti

lizados com bastante freqüência são os estudos observacionais, **prospectivo ("follow-up"), retrospectivo (caso-controle)** e **transversal ("cross-sectional")**, os quais se caracterizam pelo tipo de amostragem adotada. A diferença essencial entre estes três métodos está na maneira com que os grupos em estudo são avaliados quanto a ocorrência da doença ou da exposição ao fator de risco. Apesar disto, a forma de apresentação dos dados obtidos através destes procedimentos é a **mesma**, uma **tabela de contingência.**

Desta forma, quando um conjunto de dados está disposto em uma tabela deste tipo, é necessário identificar sob qual delineamento estes dados foram gerados, reconhecendo quais os atributos classificadores controlados **(fixados)** e quais os não controlados **(aleatórios).** Somente depois deste julgamento sobre a origem dos dados podemos prosseguir com uma análise.

Quando epidemiologistas realizam estudos observacionais, em geral, o interesse reside na determinação do **tipo de associação** entre **fatores de risco** e uma certa **doença.** Assim, por exemplo, será que o hábito de fumar está associado com o desenvolvimento de câncer no pulmão? Se estes fatores forem identificados como ativos na distribuição da doença, isto é, se sua ocorrência altera a probabilidade do desenvolvimento da doença, uma intervenção pode ser feita para verificar se realmente a modificação destas condições em pacientes é seguida por uma consequente alteração no estado doente. Conclusões deste tipo podem orientar medidas de prevenção à doença.

O problema epidemiológico estando bem esclarecido, conduz o estatístico, quando da análise de dados, a algumas propostas de **medidas de associação** que têm interesse clínico, como **risco relativo, risco adicional, risco adicional relativo.** A dificuldade se impõe na obtenção de uma medida que seja "apropriada", simultaneamente, sob

o ponto de vista médico e estatístico.

Muitos autores têm sugerido numerosas medidas quantitativas de associação, como Goodman e Kruskal (1954 e 1959). Edwards (1963), estabelece propriedades que estas quantidades devem satisfazer. O problema se concentra em encontrar medidas que suportem a aplicação de métodos clássicos de Inferência.

Nossa proposta de estudo, considera o caso de grupos **doente** ou **não doente** e a **presença** ou **ausência** do fator de interesse. Esta, é uma possível situação que leva à construção de **tabelas de contingência 2x2.** Frequentemente, é desejável analisar um fator em termos de vários níveis de exposição ou também, é desejável classificar a doença em muitas subcategorias, o que gera tabelas de múltiplas entradas. A análise destas últimas não é objetivo de nosso trabalho. Algumas bibliografias podem ser consultadas sobre o assunto: Darroch, J.N. (1962); Plackett, R.L. (1962); Birch, M.W. (1963, 1964 e 1965); Bhapker, V.P. e Koch, G.G. (1968); Goodman, A. (1964 e 1969).

O presente trabalho tem por finalidade essencial, revisar na literatura o desenvolvimento teórico da metodologia estatística utilizada na análise de tabelas de contingência 2x2, aplicando tais resultados a dados gerados de estudos epidemiológicos. Sempre que possível, tivemos o cuidado de empregar uma linguagem ascessível a pesquisadores da área médica, exceto nos capítulos que tratam com mais detalhe da teoria estatística envolvida no problema.

Os estudos observacionais obedecem a procedimentos de amostragem específicos. O capítulo II, concentra-se na descrição destes métodos que são denominados **modelos epidemiológicos.** De acordo com isto, podemos definir a cada tipo de estudo um correspondente **modelo probabilístico,** como conseqüência do que está fixado **(fator)** e o que é aleatório **(resposta)**

no conjunto de dados. Ainda neste capítulo, a quantidade **razão de produtos cruzados ("odds-ratio")** é introduzida como uma medida de associação apropriada aos três tipos de estudos observacionais, estando em analogia com o **risco relativo.**

No capítulo III, é mostrada a **equivalência condicional** entre os três estudos epidemiológicos quanto à distribuição de probabilidades imposta aos dados. Condicionalmente, ou seja, fixando as marginais linha de uma **Multinomial** (estudo transversal), o mesmo modelo probabilístico pode ser utilizado, tanto para a tabela originada a partir deste delineamento, como para aquela originada a partir de duas **Binomiais** (estudos prospectivo e retrospectivo). Por conseqüência, inferências feitas em um ou outro caso são baseadas na mesma distribuição.

Além disso, na seqüência do capítulo III, considerando o **modelo Produto de Binomiais** que é condicionalmente comum aos três estudos, novamente, empregamos o procedimento condicional, fixando as marginais coluna deste modelo. Desta forma, chegamos à **distribuição Hipergeométrica Generalizada**, parametrizada pela quantidade razão de produtos cruzados ("odds-ratio"), que é a base dos procedimentos de inferência utilizados na análise de tabelas de contingência 2x2. Impondo a restrição de igualar o parâmetro à unidade, esta distribuição se reduz à Hipergeométrica. Cox (1958, 1970), Hannan e Harkness (1963), Harkness (1964), consideram propriedades desta distribuição como momentos, estimação de parâmetros e distribuições assintóticas.

O uso de distribuições de probabilidades condicionais em situações multiparamétricas é, sob certas condições, preferível ao modelo não condicional na realização de Inferências específicas sobre um subconjunto dos parâmetros (Lehmann, 1959). As razões para sua derivação na análise de tabelas de contingência são fortalecidas pelos trabalhos de Fisher (1935) que considera

os totais marginais como "não informativos" sobre a proporcionalidade das freqüências no interior das tabelas. Alguns autores questionam esta conduta, propondo alternativas Bayesianas (Basu, 1977 e 1979; Irony, 1984).

No capítulo IV, apresentamos uma descrição do **Método da Redução**, uma técnica de Inferência Parcial utilizada na redução de dimensionalidade em situações com parâmetros "nuisanse" (Basu, 1977; Cox, 1975). No caso da análise de tabelas de contingência 2x2, este método, juntamente com o conceito de G-ancilaridade (Godambe, 1980), fundamentam a metodologia aplicada.

No capítulo V, com base na teoria de inferência apresentada, são propostas algumas medidas de associação que têm interesse para os epidemiologistas e, verifica-se em cada caso, a possibilidade de utilização do Método da Redução. Ressaltamos que o problema é considerado sugerindo transformações biunívocas no espaço original dos parâmetros, isto é, cada medida de associação conduz a uma reparametrização, sob a qual, desejamos fazer inferências sobre um parâmetro de interesse e eliminar parâmetros "nuisance".

Toda a argumentação desenvolvida até aqui, nos leva a justificar a escolha da razão de produtos cruzados ("odds-ratio") como a medida de associação consagrada na literatura, superando outras quantidades. Adequa-se sob o ponto de vista estatístico, sustentando a aplicação de procedimentos clássicos e, também sob o ponto de vista clínico, dada a sua analogia com o risco relativo.

Nesta altura do trabalho, o problema do epidemiologista na determinação da etiologia de doenças, foi abordado analiticamente. Apresentamos todas as ferramentas que são necessárias para a identificação do tipo de relacionamento entre

fatores de risco e doença. Resta considerarmos possíveis formas de utilização prática destas ferramentas.

Com esta finalidade, no capítulo VI, apresentamos uma revisão de **procedimentos de teste** para o parâmetro razão de produtos cruzados, de acordo com métodos exatos e assintóticos. Considerando o testo do parâmetro igual a **unidade**, isto é, investigar a hipótese de que o risco da doença é o mesmo na presença ou ausência do fator, um teste exato pode ser construído baseado na distribuição Hipergeométrica, o qual corresponde ao Teste Exato de Fisher (Fisher, 1935). Um método assintótico para testar esta mesma hipótese é derivado propriedades de convergência desta distribuição para a Normal.

Considerando hipóteses gerais, sem restrição para o parâmetro, o que corresponde a investigar a proporcionalidade entre o risco da doença na presença do fator e o risco da doença na ausência do fator, muitas alternativas de testes baseadas em limites de confiança assintóticos para a razão de produtos cruzados, têm sido propostas por vários autores. Cornfield (1956) fornece expressões para o cálculo exato do intervalo de confiança; a determinação destes limites sem o auxílio de recursos computacionais, bem como de métodos iterativos é impraticável devido à complexidade numérica envolvida. Thomas (1971), Mantel e Hankey (1971), desenvolveram programas específicos para a determinação destes limites.

Cox (1958), obtém limites de confiança assintóticos recorrendo a resultados da teoria de amostragem sem reposição de população finita e à função geratriz dos cumulantes da distribuição Hipergeométrica Generalizada.

Em Cornfield (1956) é proposto limites de confiança aproximados para a moda da distribuição Hipergeométrica Generalizada e, considerando que a razão de produtos cruzados po

de ser aproximada por uma função monotônica da moda, estabelece-se o correspondente intervalo. Os cálculos requerem métodos iterativos para obtenção das estimativas.

Gart (1962), propõe dois métodos de obtenção do intervalo de confiança, um apropriado para amostras grandes, com base em procedimentos **não condicionais** e outro, para amostras pequenas, derivado do modelo **condicional**. Ainda, citado por Gart, temos o intervalo obtido por Wolf (1954) baseado na transformação logito e em estimativas de máxima verossimilhança **não condicionais**. Este artigo nos parece bastante interessante, pois compara os métodos de Cornfield, Cox, Gart e Wolf, considerando a precisão, o comprimento dos intervalos e a complexidade dos cálculos envolvidos. Questionamos, do ponto de vista teórico, até que ponto os limites de intervalos baseados em procedimentos condicionais e não condicionais, podem ser comparados.

Com a descrição destes métodos, na seção VI.3 são desenvolvidas algumas aplicações. Na análise das tabelas é dado destaque especial para o tipo de estudo epidemiológico utilizado.

Finalmente, o leitor que esteja interessado, somente no sentido prático do trabalho, pode consultar apenas os capítulos II, V e os exemplos aplicados do capítulo VI juntamente com as discussões.

CAPÍTULO II

MODELO EPIDEMIOLÓGICO - MODELO PROBABILÍSTICO

Os modelos em Epidemiologia são caracterizados pelo tipo de delineamento que o pesquisador utiliza na coleta dos dados. Como conseqüência do planejamento fica definido para o estatístico o que está fixado (fator) e o que é aleatório (resposta) no modelo. O conhecimento destes fatos, como veremos a seguir, determina a análise e interpretação dos resultados.

II.1. Estudo Prospectivo ("Follow-Up")

O estudo prospectivo considera um grupo de pessoas (uma coorte), todos identificados no início do experimento como estando **livres** de uma certa característica, mas que **variam na exposição** a um fator de interesse, possivelmente relacionado com o desenvolvimento de tal característica. O grupo é seguido durante certo tempo com a finalidade de determinar diferenças na taxa de desenvolvimento da resposta em estudo de acordo com o fator de exposição. Assim, grupos de mulheres com e sem um certo fator de risco (tabagismo) são seguidas durante a gestação e o peso do recém-nascido é anotado (baixo peso ou não).

No contexto do presente trabalho com a finalidade de utilizar uma linguagem mais clara ao leitor, a característica ou resposta de interesse será indicada como a presença ou ausência de uma dada doença.

Neste método de estudo é essencial que os indivíduos sejam corretamente classificados quanto à exposição aos possíveis fatores de risco. Suposições básicas são que, indivíduos expostos no estudo sejam **representativos** de todas as pessoas expostas com respeito ao risco da doença, por exemplo, e, que indivíduos não expostos, sejam igualmente representativos de todas as pessoas não expostas na população.

Nos estudos prospectivos muitas são as maneiras de selecionar uma coorte. Um grupo particular pode ser escolhido devido a sua disponibilidade (voluntários), devido a seus antecedentes, ou devido a terem experimentado algum tipo particular de exposição. Em qualquer situação, é importante que uma coorte consista de indivíduos que compartilham de uma experiência comum durante um período de tempo definido.

O termo **direcionalidade** se referé ao **relacionamento temporal** entre as observações do fator em estudo e as observações da condição da doença. Os estudos prospectivos destacam a habilidade do pesquisador em distinguir condições **antecedentes** de **consequentes**, critério este, importante para a identificação de que o fator é de risco para a doença (Kleinbaum, Kupper e Morgenstern, 1982).

A maior vantagem destes estudos é que a coorte é classificada em relação à exposição ao fator **antes** do desenvolvimento da doença, o que elimina certos vícios de seleção que limitam seriamente outros delineamentos básicos. O fato de ter conhecimento da exposição ao fator, por sua vez, pode interferir na susceptibilidade à doença, mas isto não é difícil de ser

controlado. A conseqüência mais séria, é que estudos prospectivos permitem o cálculo direto de **taxas de incidência** entre grupos expostos e não expostos; este fato será analisado mais adiante.

A principal desvantagem de tal estudo é que geralmente é requerido um longo tempo para a coleta de dados, além de ser dispendioso. O grupo de indivíduos seguidos durante o estudo é grande, particularmente se a doença tem baixa incidência. O fato de ser necessário seguir a coorte durante longo período de tempo (estudo longitudinal) resulta em obstáculos especiais, tal como perda de pacientes pela migração, morte por outras causas, mudanças no critério de diagnóstico. Ainda, estes estudos não são eficientes em gerar novas hipóteses de fatores que supostamente possam participar na etiologia da doença.

Desta forma, os dados gerados sob este planejamento, podem ser dispostos no formato apresentado na Tabela II.1.

TABELA II.1. Notação e formato para a distribuição dos dados de um estudo prospectivo

RESPOSTA / FATOR	D	D'	TOTAL
A	x	m-x	m
A'	y	n-y	n
TOTAL	t	N-t	N

onde:

A e A' são indicadores dos grupos exposto e não exposto ao fator A, respectivamente;

D e D' são indicadores dos indivíduos que desenvolveram ou

não a resposta durante um certo tempo, respectivamente. A resposta de interesse pode ser, por exemplo, a ocorrência de uma doença.

Em correspondência ao experimento que é realizado nestes estudos, podemos estabelecer que existem **m** indivíduos aleatoriamente selecionados (com reposição), expostos ao fator A e **n** indivíduos também aleatoriamente selecionados (com reposição) não expostos ao fator A (grupo com a característica A'). Todos, indivíduos **sadios**, livres da doença. Para cada uma destas categorias observa-se durante um certo tempo o número de indivíduos que desenvolveram a doença, x e y, respectivamente.

Vale destacarmos, que na prática, mesmo utilizando processos de amostragem **sem reposição**, podemos conduzir os resultados teóricos adotando um esquema com reposição, pois as populações biológicas são infinitas internamente, o que torna equivalente os dois processos.

Sejam as variáveis aleatórias X_i $(i=1,2,\ldots,m)$ e Y_j $(j=1,2,\ldots,n)$, definidas como:

$$X_i = \begin{cases} 1 & \text{se o i-ésimo indivíduo do grupo A desenvolve a doença;} \\ 0 & \text{caso contrário.} \end{cases}$$

$$Y_j = \begin{cases} 1 & \text{se o j-ésimo indivíduo do grupo A' desenvolve a doença;} \\ 0 & \text{caso contrário.} \end{cases}$$

Com estas definições podemos esquematizar a Tabela II.2.

TABELA II.2. Distribuição das probabilidades de ocorrência da doença (D) para as populações expostas e não expostas ao fator A

RESPOSTA / FATOR	D	D'	TOTAL
A	p_1	$1-p_1$	1
A'	p_2	$1-p_2$	1

onde:

p_1 = probabilidade de um indivíduo pertencente à **população** com o fator A presente desenvolver a doença após um certo tempo;

p_2 = probabilidade de um indivíduo pertencente à **população** com o fator A ausente (A') desenvolver a doença após um certo tempo.

Desta forma, as variáveis X_i e Y_j seguirão distribuições de Bernoulli, isto é:

$$X_i \sim Ber\ (1; p_1) \qquad i = 1,2,\ldots,m$$

$$Y_j \sim Ber\ (1; p_2) \qquad j = 1,2,\ldots,n$$

Como:

$$X = \sum_{i=1}^{m} X_i \qquad e \qquad Y = \sum_{j=1}^{n} Y_j$$

estes números seguirão distribuições Binomiais, isto é:

$$X \sim Bin\ (m;p_1) \qquad x = 0,1,\ldots,m$$

$$Y \sim Bin\ (n;p_2) \qquad y = 0,1,\ldots,n$$

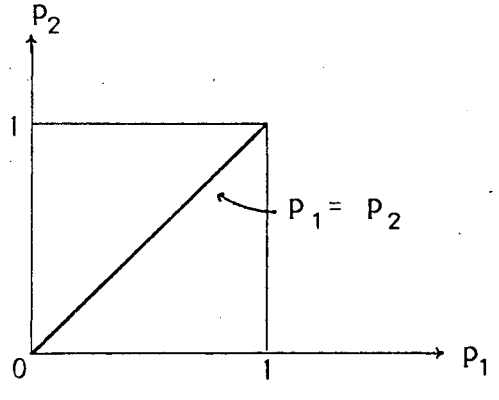
A distribuição conjunta das variáveis aleatórias (X,Y), é dada por:

$$P(x,y/N,m,p_1,p_2) = \binom{m}{x} p_1^x (1-p_1)^{(m-x)} \binom{n}{y} p_2^y (1-p_2)^{(n-y)} \qquad (1)$$

$$n = N-m$$
$$x = 0,1,\ldots,m$$
$$y = 0,1,\ldots,n$$

O **espaço dos parâmetros** para este modelo probabilístico, anotado por $\Theta = (p_1,p_2)$, é o conjunto dos pontos pertencentes ao seguinte quadrado:

FIGURA II.1. Espaço dos parâmetros (p_1,p_2) da distribuição conjunta de (X,Y). Indicação do segmento de reta $p_1=p_2$



O interesse do pesquisador está em avaliar se o fato do indivíduo ter o fator A presente ou ausente (A'), não influencia a probabilidade de que ele desenvolva a doença D, durante um certo tempo. Portanto, a hipótese a ser testada é:

$$H_0 : Pr(D/A) \quad Pr(D/A')$$

e pode ser escrita como

$$H_0 : p_1 = p_2$$

A hipótese de nulidade (H_0) é a **hipótese de homogeneidade** entre as distribuições das variáveis X e Y, isto é, $p_1 = p_2$. Os pontos no espaço dos parâmetros que satisfazem H_0 são todos aqueles pertencentes ao segmento de reta $p_1 = p_2$ indicado na Figura II.1.

As definições dadas aos parâmetros p_1 e p_2 merecem algumas considerações. Como já foi visto, nos estudos prospectivos populações classificadas em dois subgrupos são seguidas durante certo período, após o qual é contado o número de **novos casos** de uma certa doença. Epidemiologistas definem p_1 como a **taxa de incidência** da doença para a população exposta ao fator A e p_2 como a **taxa de incidência** da doença para a população não exposta ao fator A (Mausner e Kramer, 1985).

Como destacamos, uma das vantagens dos estudos prospectivos é que eles permitem o cálculo direto das taxas de incidência da doença. Isto porque, os dois grupos, A e A', representam **populações de risco bem definidas**, no início do experimento livres da doença e, que são seguidas durante certo tempo para anotar-se o desenvolvimento da doença.

Algumas estatísticas para testar a hipótese de homogeneidade são de particular interesse aos epidemiologistas e serão propostas no capítulo V. Contudo, vamos fazer alguns comentários sobre uma delas, o **risco relativo (RR)**. Este é definido como a razão das taxas de incidência para pessoas expostas e não expostas a um fator de risco. Em nossa notação:

$$RR \quad \frac{\text{taxa de incidência para expostos}}{\text{taxa de incidência para não expostos}}$$

$$RR = \frac{Pr(D/A)}{Pr(D/A')} = \frac{p_1}{p_2} \qquad (2)$$

Considerando a hipótese H_0, o risco relativo se igua la à unidade. Um resultado bastante interessante é que, ao de finirmos uma outra quantidade,

$$\Theta = \frac{p_1 \ / \ (1-p_1)}{p_2 \ / \ (1-p_2)} = \frac{p_1 \ (1-p_2)}{p_2 \ (1-p_2)} \qquad (3)$$

toda vez que igualarmos o risco relativo à unidade, o valor de Θ também se iguala à unidade (Mantel e Haenszel, 1959). Isto é uma indicação da relevância destas duas medidas em testar a hipótese $p_1 = p_2$.

A quantidade Θ é chamada **razão de produtos cruzados ("odds-ratio")** e é neste sentido, que se estabelece a analo gia entre esta e o risco relativo como medida de associação.

II.2. Estudos Retrospectivos (Caso-Controle)

No estudo retrospectivo, pessoas diagnosticadas no início do experimento como tendo uma certa característica **(caso)** são comparadas com pessoas que não têm a característica **(controles)**, de acor do com a presença de algum fator nas suas experiências passa das. A finalidade é determinar se os dois grupos diferem na proporção de pessoas com o fator presente. Assim, uma popu lação pode ser classificada em pessoas com e sem câncer pulmo nar. Utilizando uma amostra dela verifica-se o número de pes soas em cada grupo que tem o hábito de fumar. Também neste ca

so, definiremos os grupos caso e controle como constituídos por indivíduos com e sem uma dada doença, respectivamente.

O fato de que nestes estudos não é feito o seguimento dos grupos, o que proporcionaria ao pesquisador acompanhar o desenvolvimento da doença, caracteriza este delineamento como **não direcional**, isto é, nenhuma condição, fator ou doença, pode ser unicamente identificada como tendo ocorrido primeiro. Deste modo não é possível obter informação sobre a seqüência temporal dos eventos.

Nestes estudos, o estabelecimento dos grupos caso e controle, bem como os estágios da doença que serão incluídos no estudo, deve ser precisamente especificado no início da pesquisa. Uma seleção ótima, consiste de todos os casos recentemente diagnosticados **(incidentes)**, com certas características especificadas durante um período de tempo também especificado. Casos incidentes são preferíveis a casos **prevalentes**, devido ao fato de estes últimos representarem um subgrupo selecionado de todos os casos incidentes mais aqueles diagnosticados num período antecedente, que realmente não devem fazer parte do estudo. Casos prevalentes na amostra conduzem a uma maior sobrevida de pacientes, o que pode ser interpretado erroneamente como estando associado com uma predisposição excessiva à doença.

É interessante salientar, que no método retrospectivo o grupo de casos pode ser gerado por uma variedade de fontes, tais como, prontuários de hospital, certidão de óbito, receitas médicas. Da mesma forma, o grupo controle também pode ser obtido de várias fontes, incluindo a população geral ou mesmo hospitalar.

Entre as mais importantes considerações práticas que afetam estes estudos está a obtenção de casos e controles. A suposição fundamental na análise de dados retrospectivos é que os casos e controles sejam **representativos** do universo definido sob investigação. A escolha do controle, por exemplo, deve ser

tal, que os indivíduos que constituirão este grupo devem ter tido a mesma oportunidade de serem expostos aos fatores de risco que os indivíduos do grupo de casos.

Quando uma doença é endêmica ou epidêmica, a escolha de um controle é bastante evidente, como por exemplo, pessoas livres da doença na área afetada. No caso de doenças com alta taxa de fatalidade, cuja a ocorrência não atinge um grupo bem definido, a escolha do controle é mais difícil.

Por outro lado, tendo sido identificados os indivíduos que constituirão um e outro grupo, os estudos retrospectivos apresentam a vantagem de que, geralmente, a informação requerida sobre eventos passados está disponível em relatos de rotina.

Desta forma, os dados gerados por um estudo deste tipo podem ser dispostos no formato apresentado na Tabela II.3.

TABELA II.3. Notação e formato para a distribuição dos dados de um estudo retrospectivo

RESPOSTA / GRUPO	A	A'	TOTAL
CASO (D)	x	m-x	m
CONTROLE (D')	y	n-y	n
TOTAL	t	N-t	N

Considerando o experimento que é realizado nos estudos retrospectivos, existem **m** indivíduos diagnosticados no início do estudo como pertencentes ao grupo doente (D), e **n** indivíduos como pertencentes ao grupo não doente (D'). Para cada uma destas duas categorias de pacientes é verificado, co

mo resposta, o número de indivíduos expostos ao fator A, x e y, respectivamente.

Sejam as variáveis aleatórias X_i (i=1,2,...,m) e Y_j (j 1,2,...,n), definidas como:

$$X_i = \begin{cases} 1 & \text{se o i-ésimo paciente do grupo doente (caso) tem o fator A presente;} \\ 0 & \text{caso contrário} \end{cases}$$

$$Y_j = \begin{cases} 1 & \text{se o j-ésimo paciente do grupo controle tem o fator A presente;} \\ 0 & \text{caso contrário} \end{cases}$$

A seguinte tabela pode ser esquematizada:

TABELA II.4. Distribuição das probabilidades de ocorrência do fator de risco (A) para os grupos populacionais caso (D) e controle (D')

RESPOSTA / GRUPO	A	A'	TOTAL
CASO (D)	p_1	$1-p_1$	1
CONTROLE (D')	p_2	$1-p_2$	1

onde:

p_1 = probabilidade de um paciente pertencente à população doente ter o fator A presente;

p_2 = probabilidade de um paciente pertencente à população controle ter o fator A presente.

Como no caso anterior para os estudos prospectivos, temos:

$$X = \sum_{i=1}^{m} X_i \qquad \text{e} \qquad Y = \sum_{j=1}^{n} Y_j$$

com distribuições de probabilidades dadas por:

$$X \sim Bin\,(m\,;\,p_1) \qquad x = 0,1,\ldots,m$$

$$Y \sim Bin\,(n\,;\,p_2) \qquad y = 0,1,\ldots,n$$

A distribuição conjunta de (X,Y) é obtida pelo produto das duas Binomiais independentes, indicada na expressão (1).

Comparando as Tabelas II.1 e II.3 pode ser observado que os modelos epidemiológicos prospectivo e retrospectivo, apesar de consistirem de delineamentos diferentes, geram dados que podem ser esquematizados pelo mesmo formato, uma tabela de contingência 2x2 e, ainda mais, as variáveis respostas, X e Y, obedecem à mesma distribuição de probabilidades, um modelo Binomial. A diferença entre os dois estudos, está na identificação de quais variáveis foram fixadas e quais foram aleatorizadas.

Nos estudos retrospectivos, em uma população classificada em grupos que têm e não têm a doença, determina-se o número de pessoas expostas ao fator. Desta forma, p_1 e p_2 não representam taxas de incidência da doença, como nos estudos prospectivos, mas sim, **proporções de presença do fator dentro de cada grupo populacional.** Consequentemente, a hipótese de homogeneidade entre os dois grupos (H_0: $p_1=p_2$) não tem a mes

ma interpretação dada anteriormente.

Deve ficar claro que nos estudos retrospectivos, taxas de incidência não podem ser calculadas diretamente, pois a população de risco (com o fator de risco presente) não está definida (fixada). Assim, apesar de p_1 e p_2 nos dois modelos epidemiológicos serem estimados pela mesma expressão matemática (ver seção V.2) a interpretação destas quantidades em cada caso deve ser feita com cuidado.

No entanto, um resultado importante é mostrado em Cornfield (1956) e Cornfield e Haenszel (1960). Mesmo sob as considerações feitas, o **risco relativo pode ser estimado em um estudo retrospectivo,** se algumas suposições estiverem satisfeitas:

(i) o grupo controle deve ser representativo da população geral;

(ii) o grupo de casos deve ser representativo da população com a doença presente;

(iii) a freqüência da doença na população deve ser suficientemente pequena.

Vamos apresentar a demonstração deste fato.

Seja P a proporção da população total que apresenta a doença. Assim definido, P é a **taxa de prevalência** da doença na população.

Sejam p_1 e p_2 definidos como na Tabela 4. A taxa de prevalência da doença restrita a pessoas expostas ao fator A, o que corresponde a **taxa de incidência da doença para pessoas expostas** é, pelo Teorema de Bayes:

$$Pr\ (D/A) \quad Pr\ (D,A)/Pr\ (A)$$

$$\frac{Pr(A/D)\ Pr(D)}{Pr(A/D)\ Pr(D) + Pr(A/D')\ Pr(D')}$$

$$= \frac{p_1\ P}{p_1\ P + p_2\ (1-P)}$$

Do mesmo modo, a taxa de prevalência da doença para pessoas não expostas ao fator A, isto é, a **taxa de incidência da doença para pessoas não expostas** é:

$$Pr(D/A') = Pr(D,A')/Pr(A')$$

$$= \frac{Pr(A'/D)\ Pr(D)}{Pr(A'/D)\ Pr(D) + Pr(A'/D')\ Pr(D')}$$

$$= \frac{(1-p_1)\ P}{(1-p_1)\ P + (1-p_2)\ (1-P)}$$

A prevalência da doença para expostos (grupo A) relativo a não expostos ao fator (grupo A') define o **risco relativo** de ocorrência da doença, o que é obtido pela razão:

$$\frac{Pr(D/A)}{Pr(D/A')} = \frac{p_1\ [(1-p_1)\ P + (1-p_2)\ (1-P)]}{(1-p_1)\ [p_1 P + p_2\ (1-P)]}$$

Para P suficientemente pequeno podemos escrever:

$$RR = \frac{Pr(D/A)}{Pr(D/A')} \cong \frac{p_1\ (1-p_2)}{p_2\ (1-p_1)} = \Theta \qquad (4)$$

A expressão (4) nos permite fazer estimativas do risco relativo usando a quantidade razão de produtos cruzados ("odds-ratio"), estando as condições (i), (ii), (iii) satis

feitas. Em **doenças raras**, onde estudos prospectivos são difíceis de conduzir, é assim que o risco relativo pode ser estimado através de estudos retrospectivos. Novamente, foi possível estabelecer uma analogia entre estas duas medidas.

II.3. Estudos Transversais ("Cross-Sectional")

Os estudos transversais representam outra forma dos epidemiologistas realizarem observações nas populações. Neste tipo de delineamento, uma amostra representativa da população em estudo deve compor a pesquisa, em que, ambos, fator de risco e doença, por exemplo, são verificados ao **mesmo tempo** em cada unidade experimental. Por esta razão, o termo "através de uma secção" é usado como tradução a "cross-sectional".

Um modelo epidemiológico deste tipo, apesar de ser fácil e rápido de realizar, não estabelece a seqüência temporal dos eventos, que é necessária para o pesquisador extrair inferências sobre causas, por exemplo, entre a doença e fator de risco (Mausner e Kramer, 1985). Assim, este é um delineamento **não direcional**, pois o pesquisador não consegue distinguir condições antecedentes de consequentes, o que somente é conferido aos estudos prospectivos.

Os dados gerados por um estudo transversal podem ser dispostos como na Tabela II.5, a seguir.

TABELA II.5.Notação e formato para a distribuição dos dados de um estudo transversal

VARIÁVEL D / VARIÁVEL A	D	D'	TOTAL
A	x	m-x	m
A'	y	n-y	n
TOTAL	t	N-t	N

As variáveis A(A') e D(D') podem ser, por exemplo, presença (ausência) do fator de risco e da doença, respectivamente. Observe na Tabela II.5 que a distribuição de linhas e colunas para estas variáveis é aleatória, além do que, N é o único total marginal que está fixado.

São determinadas as seguintes categorias de respostas:

AD : evento que indica presença simultânea do fator de risco (A) e da doença (D);

AD' : evento que indica presença do fator de risco (A) e ausência da doença (D');

A'D : evento que indica ausência do fator de risco (A') e presença da doença (D);

A'D' : evento que indica ausência simultânea do fator de risco (A') e da doença (D').

Considerando o experimento que é realizado nos estu

dos transversais, existem N indivíduos aleatoriamente selecionados (com reposição) de uma população, os quais são classificados nas categorias AD, AD', A'D e A'D'. Desta forma, x indivíduos pertencem à classe AD, (m-x) indivíduos pertencem à classe AD' e, assim por diante.

Como anteriormente, vale destacarmos, que para populações biológicas que são infinitas internamente, processos de amostragem com ou sem reposição podem receber o mesmo tratamento. Na população em estudo, esquematiza-se a seguinte tabela:

TABELA II.6.Distribuição das probabilidades de ocorrência das quatro categorias de respostas

VARIÁVEL D / VARIÁVEL A	D	D'	TOTAL
A	p	r	p+r
A'	q	s	q+s
TOTAL	p+q	r+s	1

onde:

p = probabilidade de um indivíduo da população pertencer à categoria de resposta AD ;

r = probabilidade de um indivíduo da população pertencer à categoria de resposta AD';

q = probabilidade de um indivíduo da população pertencer à categoria de resposta A'D;

s probabilidade de um indivíduo da população pertencer à categoria de resposta A'D'.

A disposição dos elementos da amostra nas quatro categorias de respostas segue uma **distribuição Multinomial**, que é dada por:

$$P(x,y,m/N,p,q,r) = \frac{N!}{x!\ (m-x)!\ y!\ (n-y)!}\ p^x\ r^{(m-x)}\ q^y\ s^{(n-y)} \qquad (5)$$

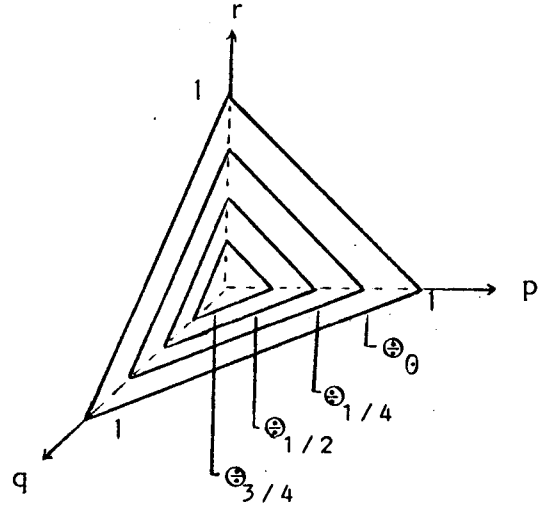
para n = N-m

O espaço dos parâmetros para este modelo probabilístico, denotado por $\Theta = (p,q,r,s)$, assumindo o valor de N conhecido, é o conjunto dos pontos pertencentes à seguinte superfície tridimensional no espaço $\mathbb{R}^4$:

Θ: $p+q+r+s = 1$; $0 <= p,q,r,s <= 1$

As curvas de nível atribuindo alguns valores a s para esta superfície estão apresentadas na Figura II.2.

FIGURA II.2. Representação das curvas de nível para o espaço dos parâmetros do modelo Multinomial



onde: Θ_0 : $p+q+r = 1$ para $s = 0$

$\Theta_{1/4}$: $p+q+r = 3/4$ para $s = 1/4$

$\Theta_{1/2}$: $p+q+r = 1/2$ para $s = 1/2$

$\Theta_{3/4}$: $p+q+r = 1/4$ para $s = 3/4$

Em um modelo Multinomial deste tipo, é possível ao estatístico testar a **hipótese de independência** entre as duas variáveis classificatórias, fator de risco e doença. Observe, que não se trata de um teste de homogeneidade entre duas distribuições, o qual é apropriado para os modelos Binomiais já descritos. Neste caso, pretende-se verificar se a ocorrência da doença, em relação à ocorrência do fator, é casual ou sistemática.

Para testarmos esta hipótese, é suficiente investigar se um dos componentes interiores da tabela pode ser fatorado como produto das marginais de linha e colunas correspondentes. O teste Qui-quadrado, bem como o coeficiente de contingência de Pearson, são procedimentos bastante utilizados. Edwards (1963), sugere que o uso destas estatísticas deve ser evitado, principalmente no caso de tabelas 2x2; no capítulo V são apresentadas as razões que justificam tais argumentações.

O epidemiologista, analisa o problema questionando se a ocorrência da doença não é afetada pela presença (ou ausência) do fator. Neste sentido com a notação apresentada na Tabela II.6, podemos estabelecer a seguinte hipótese:

$$H_0 : p/(p+r) = q/(q+s)$$

isto é, a ocorrência da doença é a mesma para indivíduos expostos e não expostos ao fator de risco.

Destacamos, que a hipótese H_0 é equivalente à hipótese de independência, ou seja, H_0 é verdadeira se, e somente se, uma das proporções fatorar como produto das marginais.

Em estudos transversais, o risco relativo, definido anteriormente, pode ser obtido por:

$$RR = \frac{Pr(D/A)}{Pr(D/D')} = \frac{p/(p+r)}{q/(q+s)} \qquad (6)$$

Podemos observar, que a hipótese H_0 corresponde a testar que o risco relativo é igual à unidade.

Por outro lado, recorremoss a um resultado interessante. A quantidade razão de produtos cruzados em um estudo transversal é obtida por:

$$\Theta = \frac{p/r}{q/s} = \frac{p\ s}{q\ r} \qquad (7)$$

Se igualarmos o risco relativo à unidade, Θ assim definido, também se iguala à unidade (Mantel e Haenszel, 1959). Desta forma, dispomos de várias estatísticas que são relevantes para a hipótese de independência. Além disso, está correspondência estabelece a analogia entre o risco relativo e a razão de produtos cruzados em estudos transversais.

CAPÍTULO III

EQUIVALÊNCIA ENTRE OS ESTUDOS OBSERVACIONAIS - DISTRIBUIÇÕES CONDICIONAIS -

No capítulo anterior descrevemos as características dos três delineamentos epidemiológicos básicos, adotando em cada caso um modelo probabilístico adequado. Utilizando resultados da Teoria de Probabilidade, juntamente com alguns conceitos dados por Fisher (1935) sobre Quantidade de Informação, podemos estabelecer equivalência entre estes estudos. Um modelo de distribuição condicional é obtido para os dados no interior de uma tabela 2x2. Especificamente, considerando a notação utilizada até aqui, o modelo se reduz à distribuição de X fixadas as marginais m e t, independente do estudo epidemiológico que gerou os dados. Para tanto, são analisadas algumas propriedades das distribuições Multinomial e Binomial.

III.1. Redução do Modelo Multinomial para o Produto de Binomiais Através do Condicionamento em Relação a uma Marginal

Seja $X = (X_1, X_2, \ldots, X_k)$ uma variável aleatória Multinomial com vetor de probabilidade $p = (p_1, p_2, \ldots, p_k)$. A função de probabilidades de X é dada por:

$$P(x/N,p) = N! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}$$

onde os x_i's de $x = (x_1, x_2, \ldots, x_k)$ tomam os valores 0,1,... ..., N, sujeitos à restrição

$$\sum_{i=1}^{k} x_i = N$$

supondo o valor de N conhecido. Os valores dos p_i's pertencem ao intervalo [0,1], sujeitos à restrição

$$\sum_{i=1}^{k} p_i = 1 .$$

A distribuição Multinomial pode ser gerada por uma série de experimentos, que têm as seguintes propriedades:

- o resultado de cada experimento pode ser classificado dentro de k categorias de respostas;

- as probabilidades de ocorrência das categorias, $p=(p_1,p_2,\ldots, p_k)$, são as mesmas para cada experimento;

- o resultado de cada experimento é independente de todos os outros;

- a série de experimentos é realizada um número fixado de vezes, N.

Estas condições se enquadram na descrição do modelo epidemiológico transversal.

A seguir, é enunciado o teorema que permite a **fatoração da distribuição Multinomial.** Sua prova encontra-se em Guenther (1968) ou Kshirsagar (1972), por exemplo.

TEOREMA III.1. Seja $X = (X_1, X_2, \ldots, X_k)$ uma variável aleatória Multinomial. Então:

(i) a distribuição marginal da soma de componentes Muntinomiais é também Multinomial;

(ii) a distribuição condicional de um subconjunto de componentes Multinomiais, dado o valor observado da soma destes componentes, é também Multinomial.

O resultado apresentado em (i) é de grande importância quando temos interesse na **distribuição dos totais marginais** de uma tabela de contingência, sendo a distribuição dos componentes internos Multinomial (como a Tabela II.5). Isto implica, que os totais de linhas e, similarmente, os totais de colunas têm distribuição Multinomial com o mesmo tamanho amostral N e vetor de probabilidades marginais de linha e de coluna, respectivamente.

Do mesmo modo, (ii) pode ser aplicado a uma tabela de contingência como a Tabela II.5, para a **distribuição condicional** dos componentes de cada linha (coluna), estando fixados os totais observados das linhas (colunas).

Assim, estamos em condições de fatorar a distribuição Multinomial em partições que também representam funções Multinomiais. Para ilustração, seja (x,y,m) um conjunto de dados gerados através de um estudo transversal. Suponha o mesmo formato e notação das Tabelas II.5 e II.6.

O modelo Multinomial (5) pode ser fatorado como:

$$P(x,y,m/N,p,q,r) = P(m/N,p,q,r)\ P(x,y/N,m,p,q,r) \qquad (8)$$

isto é, a **distribuição conjunta** de (x,y,m) se decompõe na **dis**

tribuição marginal de (m) multiplicada pela **distribuição condicional** de (x,y/m).

Utilizando os resultados do Teorema III.1:

$$(m/N,p,q,r) \sim Bin\ (N\ ;\ p+r) \qquad (9)$$

que é a distribuição marginal linha de uma Multinomial (tridimensional).

Vejamos a distribuição condicional dos componentes (x,y) **fixado** o valor observado das marginais linha:

$$P(x,y/N,m,p,q,r) = P(x,y,m/N,p,q,r) \;/\; P(m/N,p,q,r)$$

$$= \frac{\dfrac{N!}{x!\ y!\ (m-x)!\ (n-y)!}\ p^x\ q^y\ r^{(m-x)}\ s^{(n-y)}}{\dfrac{N!}{m!\ (N-m)!}\ (p+r)^m\ (q+s)^{N-m}} =$$

$$= \binom{m}{x}\left(\frac{p}{p+r}\right)^{m}\left(\frac{r}{p+r}\right)^{(m-x)} \binom{n}{y}\left(\frac{q}{q+s}\right)^{y}\left(\frac{s}{q+s}\right)^{(n-y)}$$

Fazendo:

$$p_1 = p/(p+r) \qquad \text{e} \qquad p_2 = q/(q+s)$$

$$P(x,y/N,m,p,q,r) = P(x,y/N,m,p_1,p_2)$$

$$= \underbrace{\binom{m}{x}\ p_1^x\ (1-p_1)^{(m-x)}}_{Bin\ (m\ ;\ p_1)} \underbrace{\binom{n}{y}\ p_2^y\ (1-p_2)^{(n-y)}}_{Bin\ (n\ ;\ p_2)} \qquad (10)$$

Portanto, o termo condicional (x,y/m) do modelo Multinomial (estudo transversal) corresponde ao produto de duas Binomiais, o que equivale ao modelo Produto de Binomiais expresso em (1) para os estudos prospectivo e retrospectivo. Observe, ainda mais, que o tipo de fatoração indica que se trata de duas **variáveis independentes.**

Vamos examinar com mais detalhe o que o fato de condicionar (fixar) marginais aleatórias está causando na distribuição dos dados, sujeitos a um modelo Multinomial.

Pode ser observado, que a distribuição marginal só depende dos valores (p,q,r,s) através de (p+r), ou seja, devido a (9):

$$P(m/N,p,q,r) = P(m/N,\ p+r)$$

De acordo com Fisher (1935), isto é uma indicação de que a distribuição marginal não fornece informação sobre (p,q,r,s), que são os parâmetros em que temos interesse. Realmente, conhecer o valor (p+r) não quantifica biunivocamente os valores individuais dos parâmetros. Por esta razão, este termo do modelo Multinomial pode ser abandonado, utilizando se apenas a parte condicional na realização de inferências. Isto significa, que a **informação contida em P(x,y/N,m,p,q,r) é a mesma informação contida em P(x,y,m/N,p,q,r)**; verifique a fatoração obtida em (8).

Este resultado é a origem de divergências entre vários grupos de estatísticos. Levantaremos alguns pontos de discussão no próximo capítulo.

Assim, estabelece-se a **equivalência condicional** entre os modelos Multinomial (estudo transversal) e o Produto de Binomiais (estudos prospectivo e retrospectivo), no sentido de ser pos

sível utilizar o mesmo teste para os três estudos, com base na mesma distribuição para os dados (expressões 1 ou 10). Apesar disso, o estatístico, inevitavelmente, deve conhecer a natureza dos dados, principalmente, na formulação de hipóteses e interpretação de resultados.

III.2. Redução do Modelo Produto de Binomiais para o Modelo Hipergeométrico Generalizado Através do Condicionamento em Relação a uma Marginal

Até agora, caminhamos passo a passo, no sentido de adotarmos o modelo Produto de Binomiais comum para os três estudos epidemiológicos. No prosseguimento do assunto são investigadas propriedades que permitem reduzir ainda mais este modelo. Vamos elucidar dois resultados. Quanto ao modelo Multinomial (Tabela II.5), as conclusões estabelecidas para a marginal linha observada **m**, podem ser verificadas também para a marginal coluna observada **t**, isto é, pelo Teorema III.1:

$$(t/N,p,q,r) \sim \text{Bin}\ (N\ ;\ p+q) \qquad (11)$$

e, portanto, pela mesma razão descrita anteriormente, a distribuição marginal de t não é informativa sobre os parâmetros (p,q,r,s).

Do mesmo modo, para o modelo Binomial (Tabelas II.1 e II.3), a marginal observada **t=x+y** (soma de duas variáveis Binomiais independentes) apresenta distribuição de probabilidades dadas por:

$$(t/N,m,p_1,p_2) \sim \text{Bin}\ (N\ ;\ p_1+p_2) \qquad (12)$$

e, portanto, só depende de (p_1,p_2) através de (p_1+p_2), o que a torna não informativa, no sentido Fisheriano, sobre o valor individual dos parâmetros em que temos interesse.

Verificando a Tabela II.5, bem como as Tabelas II.1 e II.3 temos $y=t-x$ e $N=m+n$. Podemos então reescrever o modelo Produto de Binomiais obtido nas expressões (1) e (10):

$$P(x,y/N,m,p_1,p_2) \quad P(x;t/m,n,p_1,p_2)$$

Fazendo a decomposição desta distribuição temos:

$$P(x,t/m,n,p_1,p_2) = P(t/m,n,p_1,p_2)\ P(x/m,n,t,p_1,p_2) \qquad (13)$$

Desta forma, podemos obter a distribuição marginal de t por:

$$P(t/m,n,p_1,p_2) = \sum_{u=k}^{l} P(u,t/m,n,p_1,p_2) =$$

$$= (1-p_1)^m\, p_2^t\, (1-p_2)^{(n-t)} \sum_{u=k}^{l} \binom{m}{u} \binom{n}{t-u} \left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^u \qquad (14)$$

onde: $k = \max(0, t-n)$

$l = \min(t,m)$

e portanto,

$$P(x/m,n,t,p_1,p_2) = P(x,t/m,n,p_1,p_2) \;/\; P(t/m,n,p_1,p_2)$$

$$= \frac{\binom{m}{x}\binom{n}{t-x}\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^x}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^u}$$

Recordando a expressão (3):

$$\theta \quad \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

temos:

$$P(x/m,n,t,p_1,p_2) = P(x/m,n,t,\theta)$$

$$= \frac{\binom{m}{x}\cdot\binom{n}{t-x}\theta^x}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\theta^u} \qquad (15)$$

Essa distribuição de probabilidades condicional **depende apenas do parâmetro θ**, definido no capítulo II como **razão de produtos cruzados ("odds-ratio")**. Já enfatizamos o interesse na hipótese de igualar θ à unidade (θ=1), o que corresponde a supor que o risco da doença é o mesmo na presença ou ausência do fator (RR=1, ou ainda, $p_1=p_2$). Sob esta hipótese obtemos:

$$P(x/m,n,t,\theta=1) = \frac{\binom{m}{x}\binom{n}{t-x}}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}} = \frac{\binom{m}{x}\binom{n}{y}}{\binom{N}{t}} \qquad (16)$$

onde:

$$N = n+m \qquad k = \max(0, t-n)$$
$$t = x+y \qquad l = \min(t, m)$$

que é o modelo da **distribuição Hipergeométrica.**

Utilizando os resultados apresentados em (11) e (12), o modelo Produto de Binomiais pode ser reduzido apenas ao termo condicional, ou seja, podemos tomar a observação t como **fixada** e aban

nar o termo marginal. A **informação contida em $P(x/m,n,t,p_1,p_2)$ é a mesma contida em $P(x,t/m,n,p_1,p_2)$**; verifique a fatoração obtida em (13).

Observe porém, que apesar de Fisher considerar a distribuição das marginais não informativas, a expressão (14) indica que o termo marginal depende do parâmetro Θ, isto é, carrega alguma informação sobre Θ que é desprezada quando se considera apenas o modelo condicional. No capítulo IV a validade deste procedimento é considerada.

Concluindo, é possível verificar que testar $H_0: p_1=p_2$ no modelo Produto de Binomiais é equivalente a testar $H_0: p/(p+r)=q/(q+s)$ no modelo Multinomial, sendo que ambas hipóteses se reduzem ao teste do parâmetro $\Theta=1$. Deste modo, podemos caracterizar Θ como uma medida de homogeneidade em estudos prospectivos e retrospectivos e uma medida de associação em estudos transversais. A distribuição de probabilidades utilizada no teste é a Hipergeométrica (16). Tal modelo reduzido, foi obtido considerando-se que as distribuições das marginais (m,t) não contêm informação sobre as hipóteses em questão e, portanto, podem ser consideradas como fixadas, independente do tipo de estudo observacional que está sendo realizado.

Em algumas situações, o pesquisador tem interesse em propor valores mais gerais ao parâmetro Θ, **diferentes da unidade**, como na hipótese de o risco da doença na presença do fator ser proporcional ao risco na ausência do fator (RR = k, ou ainda, $p_1=kp_2$). Nestes casos, inferências sobre Θ são baseadas na expressão (15), definida como **distribuição Hipergeométrica Generalizada.**

Cox (1958), apresenta formas de se obter os momentos desta distribuição. Os cálculos exatos tornam-se inconvenientes devido a complexidade das expressões matemáticas en

volvidas, por conseqüência, expressões aproximadas são desenvolvidas através das propriedades assintóticas da distribuição Hipergeométrica Generalizada. Harkness (1965) analisa três formas de convergência dependendo de como as marginais m, t e N tendem para o infinito. Cornfield (1956) Hannan e Harkness(1963) mostram em detalhe a convergência em distribuição para a Normal, que é de maior interesse em procedimentos de teste para o parâmetro θ.

III.3. Considerações Gerais sobre Modelos Condicionais e Não Condicionais

Antes de concluirmos o assunto tratado neste capítulo, uma observação merece ser destacada para validar, de certa forma, a aplicação dos resultados estabelecidos na seção anterior. Esta, diz respeito aos métodos pelos quais são feitas inferências sobre o parâmetro θ. É útil chamarmos a atenção para o fato de que existem duas distribuições de probabilidades envolvendo θ. A primeira é a **distribuição não condicional** indicada em (1) para o modelo Produto de Binomiais e em (5) para o modelo Multinomial, expressões estas que podem ser reescritas em termos de θ, por exemplo, da seguinte forma:

$$P(x,y/N,m,p_1,p_2) = \binom{m}{x} p_1^x (1-p_1)^{(m-x)} \binom{n}{y} p_2^y (1-p_2)^{(n-y)}$$

$$= \binom{m}{x} \binom{n}{y} \theta^x \left(\frac{1-p_1}{1-p_2}\right) p_2^t (1-p_2)^{(N-t)}$$

ou

$$P(x\ y,m/N,p,q,r) = \frac{N!}{x!\ y!\ (m-x)!\ (n-y)!} p^x q^y r^{(m-x)} s^{(n-y)} =$$

$$= \frac{N!}{x!\ y!\ (m-x)!\ (n-y)!}\ \theta^{x}\ (r/s)^{m}\ (p/s)^{t}\ s^{N}$$

de acordo com o tipo de estudo observacional. A outra é a distribuição **condicional** obtida, $P(x/N,m,t,\theta)$, comum aos três estudos.

Existem razões que justificam o emprego da distribuição condicional em preferência à não condicional na realização de inferências sobre θ. Utilizando resultados da Estatística Matemática, verificamos que os modelos não condicionais mostrados acima pertencem à **família de distribuições exponenciais multiparamétrica** e., de acordo com Lehmann (1959), testes UMPU sobre θ (uniformemente mais poderosos não viciados) são derivados da distribuição condicional.

Uma outra propriedade, de natureza diferente desta, conduz à mesma alternativa. Se recorrermos aos conceitos estabelecidos por Fisher (1935) de que os totais marginais da tabela são não informativos sobre a proporcionalidade das freqüências no interior da mesma, mas sim, transportam **informação ancilar**, isto é, sobre parâmetros em que não temos interesse, cabe verificarmos, que nos modelos não condicionais, a não ser θ, os demais parâmetros estão ligados a totais marginais. Devemos portanto, não considerar estes termos, o que corresponde a adotar o modelo condicional.

No capítulo IV estes fatos são tratados com mais detalhe, desenvolvendo uma base metodológica que permite aplicar e analisar estes resultados com mais clareza.

CAPÍTULO IV

CONSIDERAÇÕES SOBRE INFERÊNCIA PARCIAL - MÉTODO DA REDUÇÃO -

A exposição que se segue, consiste de um dos assuntos mais polêmicos da Inferência Estatística. Entretanto, salientamos que não é nosso objetivo discutir as justificativas ou possíveis implicações da metodologia em questão, mas especificamente, descrevê-la como uma técnica útil na análise de problemas que envolvem outros parâmetros, além daqueles em que estamos interessados.

Chamamos a atenção para o fato de que no capítulo anterior alguns procedimentos de redução de modelos completos a termos condicionais foram introduzidos. Passaremos a analisar o suporte teórico que garante a validade de tais resultados.

IV.1. Função de Verossimilhança e Quantidade de Informação

Para a apresentação da metodologia em destaque neste capítulo, será adotada a seguinte notação:

Seja X uma variável aleatória definida no espaço de probabilidade (Ω, Λ, P), onde Ω é o espaço amostral de um certo experimento, Λ é a sigma-álgebra de eventos em Ω, e P é a função de probabilidade definida em Λ e parametrizada por $\Theta = (\theta, \Phi)$, tal que, para um ponto X x pertencente à sigma álgebra Λ, $P(x) = P(x/\theta, \Phi)$.

Duas considerações preliminares são necessárias. A primeira, diz respeito à **quantidade de informação** contida em uma amostra. Entre os estatísticos existe bastante polêmica em torno deste tema.

Suponha, inicialmente o caso em que o espaço paramétrico é unidimensional, $\Theta = \theta$, por exemplo. Ainda, seja P uma função de probabilidade definida para a variável aleatória X e parametrizada por θ. O valor:

$$E\left[\frac{\partial}{\partial\theta} \ln P(x/\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2} \ln P(x/\theta)\right]$$

foi definido por Fisher como a **quantidade de informação** sobre θ contida na observação X=x. Observe que esta medida se refere a uma média para todas as possíveis observações de X.

Uma generalização pode ser feita para o caso multiparamétrico, por exemplo, $\Theta = (\theta, \Phi)$. O análogo da informação de Fisher é uma **matriz de informação**, dada por:

$$E\begin{bmatrix} \left(\frac{\partial}{\partial\theta} \ln P(x/\theta,\Phi)\right)^2 & \frac{\partial}{\partial\Phi} \ln P(x/\theta,\Phi)\ \frac{\partial}{\partial\Phi} \ln P(x/\theta,\Phi) \\ \frac{\partial}{\partial\Phi} \ln P(x/\theta,\Phi)\ \frac{\partial}{\partial\theta} \ln P(x/\theta,\phi) & \left(\frac{\partial}{\partial\Phi} \ln P(x/\theta,\phi)\right)^2 \end{bmatrix}$$

$$= -E \begin{bmatrix} \frac{\partial^2}{\partial\theta^2} \ln P(x/\theta,\phi) & \frac{\partial}{\partial\theta\partial\phi} \ln P(x/\theta,\phi) \\ \frac{\partial}{\partial\phi\partial\theta} \ln P(x/\theta,\phi) & \frac{\partial^2}{\partial\phi^2} \ln P(x/\theta,\phi) \end{bmatrix}$$

Neste sentido, a dificuldade que frequentemente surge é que as situações experimentais sugerem modelos deste tipo, envolvendo muitos parâmetros, sendo que, o interesse reside em obter informação apenas sobre um subconjunto deste espaço. Considere, por exemplo, que $X = (X_1, X_2, \ldots, X_n)$ representa uma seqüência de variáveis aleatórias Normais, independentes e identicamente distribuídas, com média μ e variância σ^2, isto é, $\Theta = (\mu, \sigma^2)$. Suponha, que temos interesse em fazer inferência apenas sobre μ, independente do parâmetro σ^2. Como obter **informação específica** sobre μ?

O problema é abordado com a seguinte questão: Como eliminar **parâmetros "nuisance"** do modelo, isto é, parâmetros que não são de interesse relevante? Este fato é analisado por muitos autores sob diferentes pontos de vista. Na seção IV.2 este assunto é discutido com mais detalhe.

Uma segunda consideração importante a ser feita se refere à **função de verossimilhança.** De acordo com a notação anterior, esta pode ser definida como:

$$L(\theta,\phi/x) = P(x/\theta,\phi)$$

isto é, tomando-se P como função apenas dos parâmetros do modelo, onde a amostra x não é considerada estocástica, mas efetivamente observada. Note, que apesar de existir uma equivalência entre as formas funcionais de L e P, no caso da função

de verossimilhança L, a amostra x já foi observada e para a função de probabilidades P, a amostra x é um elemento aleatório.

Quando um experimento é realizado o pesquisador, objetivando adquirir conhecimento sobre parâmetros, em geral, dispõe somente de um conjunto de dados observados. Para Fisher, a função de verossimilhança é a única entidade que carrega toda a informação sobre os parâmetros contida em um conjunto de dados.

Desta forma, em situações multiparamétricas, em que desejamos obter informação objetiva sobre um subconjunto dos parâmetros, parece natural, que o problema de eliminar parâmetros "nuisance", seja investigado a partir da função de verossimilhança. Será que existe uma estatística que esteja associada especificamente com a parte do espaço paramétrico que estamos interessados? Procedimentos deste tipo serão considerados na próxima seção.

Retomando o capítulo I, supondo que os dados apresentados na Tabela II.1 (Tabela II.3) representam o resultado que foi efetivamente observado na realização de um estudo prospectivo (retrospectivo), a função de verossimilhança é definida como:

$$L(m,n,p_1,p_2/x,y) = \binom{m}{x} p_1^x (1-p_1)^{(m-x)} \binom{n}{y} p_2^y (1-p_2)^{(n-y)} \qquad (16)$$

para $0 <= p_1, p_2 <= 1$

Do mesmo modo, considerando os dados da Tabela II.5 para o modelo Multinomial, como resultado de um experimento transversal, a função de verossimilhança é definida como:

$$L(N,p,q,r/x,y,m) \quad \frac{N!}{x!\ y!\ (m-x)!\ (n-y)!}\ p^x\ q^y\ r^{(m-x)}\ s^{(n-y)} \qquad (17)$$

para $n = N-m$

$0 <= p,q,r,s <= 1 \quad ; \quad p+q+r+s = 1$

IV.2. Método da Redução

Como já foi mencionado no capítulo anterior, um teste da hipótese de homogeneidade no modelo Produto de Binomiais ou independência no modelo Multinomial, pode ser baseado na X **distribuição condicional** de X, fixadas as marginais (m,t), que é a distribuição Hipergeométrica.

Na prática, métodos de análise baseados em distribuições condicionais são bastante aplicados. Andersen (1967), através de propriedades derivadas de distribuições condicionais, obtém estatísticas equivalentes à "t de Student" para testar o valor da média em um modelo Normal. O mesmo autor (1970), obtém estimativas consistentes para parâmetros estruturais, maximizando a função de verossimilhança condicional. Cox (1975), utiliza métodos condicionais para analisar tabelas de vida.

Todas estas situações, apesar de apresentarem natureza diferente, sob o escopo da Inferência, podem ser reduzidas a um único caso. O espaço paramétrico envolve **outros parâmetros** além daqueles sobre os quais desejamos obter informação. Desta forma, como já citamos na seção anterior, o problema de inferência se baseia na **eliminação de parâmetros "nuisance"** do modelo.

Vários métodos de eliminação têm sido propostos na literatura. Uma alternativa bastante utilizada para grandes

amostras é substituir o parâmetro "nuisance" por sua estimativa de máxima verossimilhança. Outros métodos utilizam argumentos Bayesianos. Em particular, destacamos o **Método da Redução,** um procedimento da Inferência cuja origem pode ser traçada nos trabalhos de Fisher.

Muitos autores discutem e reexaminam os argumentos utilizados no Método da Redução (Gart, 1971; Kalbfleisch e Sprott, 1973; Cox, 1975; Basu, 1975, 1977 e 1979). A técnica consiste precisamente na **fatoração da função de verossimilhança** em termos que dependem apenas do parâmetro de interesse e termos que dependem do parâmetro "nuisance". Em um certo sentido, a característica principal deste método de análise consiste na obtenção de distribuições livres do parâmetro "nuisance" e, com base nestes modelos, são realizadas inferências específicas relativas ao parâmetro de interesse.

Em muitas situações multiparamétricas, nem sempre é possível obter a fatoração desejada. O caso mais favorável seria aquele em que os parâmetros de interesse e "nuisance" estão definidos de tal forma, que seja possível obter estatísticas, que forneçam informação sobre cada subconjunto do espaço paramétrico independentemente (Andersen, 1967).

Em geral, como veremos na próxima seção, temos interesse em uma certa função dos parâmetros do modelo em estudo, o que corresponde a introduzir uma reparametrização que defina o novo parâmetro de interesse. As seguintes condições devem estar satisfeitas:

(i) a reparametrização proposta pelo pesquisador deve consistir de uma **transformação biunívoca** no espaço dos parâmetros, o que é facilmente obtido fazendo uma escolha adequada do parâmetro "nuisance". Além disso, os novos parâmetros devem ser **identificáveis,** isto é, para cada variação dos novos parâmetros a função de verossimilhança tam

bém deve variar;

(ii) a reparametrização deve ser tal, que a função de verossimilhança possa ser fatorada apropriadamente de maneira a tornar possível retirar dos dados **informação específica** a respeito do parâmetro (ou função dele) de interesse, e abandonar os termos ligados a parâmetros "nuisance".

A obtenção de distribuições livres do parâmetro "nuisance" pode ser conduzida de duas maneiras essencialmente diferentes. Por esta razão, dois princípios são formulados originando as versões do Método da Redução. Para enunciá-los, considere na notação anterior que no espaço $\Theta = (\theta, \phi)$, θ é o parâmetro de interesse e ϕ o parâmetro "nuisance". De acordo com Basu (1977) considere ainda mais, que o parâmetro ϕ é de **variação independente** complementar a θ, ou seja, o produto cartesiano $\theta \times \phi$ define o espaço dos parâmetros Θ para o modelo P.

Observe ainda que correspondendo a qualquer estatística $T=T(x)$ é possível obter a fatoração de P da forma:

$$P(x/\theta,\phi) = \sum_{T(x)} P(x,T(x)/\theta,\phi)$$

$$= P_1(T/\theta,\phi)\ P_2(x/T,\theta,\phi)$$

onde, P_1 é a distribuição marginal de T e P_2 a distribuição condicional de X dado T. Recorde que, como T é função de X, para cada valor $X=x$ a soma

$$\sum_{T(x)} P(x,T(x)/\theta,\phi)$$

se reduz a um único termo, que pode ser indicado como acima.

Com isto exposto, estamos em condições de estabele

cer os seguintes princípios:

1. **Princípio da Marginalização.** Se a fatoração do modelo original satisfaz

$$P(x/\theta,\phi) = P_1(T_1/\theta)\ P_2(x/T_1,\phi) \qquad (18)$$

então,

(i) o termo marginal P_1 depende apenas do parâmetro de interesse θ. T_1 é **parcialmente ancilar para ϕ.**
Basu (1977) define T_1 como S-ancilar para ϕ.

(ii) o termo condicional P_2 depende apenas de ϕ. T_1 é **parcialmente suficiente para θ.**
Basu (1977) define T_1 como p-suficiente para θ.

2. **Princípio do Condicionamento.** Se a fatoração do modelo original satisfaz:

$$P(x/\theta,\phi) = P_1(T_2/\phi)\ P_2(x/T_2,\theta) \qquad (19)$$

então,

(i) o termo marginal P_1 depende apenas de ϕ.
T_2 é **parcialmente ancilar** para θ.
Segundo Basu, T_2 é S-ancilar para θ.

(ii) o termo condicional P_2 depende apenas de θ.
T_2 é **parcialmente suficiente** para ϕ.
Segundo Basu, T_2 é p-suficiente para ϕ.

As denominações "parcialmente suficiente" e "parcialmente ancilar" são introduzidas para indicarem que as propriedades de suficiência e ancilaridade não se aplicam ao espaço

paramétrico total, mas a um subconjunto deste. Podemos qualificá-las como generalizações dos conceitos usuais de suficiência e ancilaridade, que são utilizados para os casos em que o parâmetro de interesse coincide totalmente com o parâmetro do modelo.

Um outro ponto que precisa ser esclarecido é que, considerando a definição de função de verossimilhança, os critérios de fatoração nos dois Princípios, podem ser reescritos ou mesmo analisados, para a amostra x efetivamente observada, isto é, em termos da $L(\theta,\Phi/x)$. Deste modo, em cada caso, temos definida a função de verossimilhança marginal e condicional. Neste sentido, Cox (1975) define **função de verossimilhança parcial**, a partir da qual pode ser realizada inferência parcial.

Vejamos como o conceito de Inferência Parcial se aplica ao Princípio da Marginalização. Neste caso, a obtenção de distribuições livres do parâmetro "nuisance" Φ, é garantida devido à existência de estatísticas T_1, cujas distribuições dependem apenas de θ. Fisher (1935), introduziu a propriedade de ancilaridade para tais estatísticas. Assim, inferências sobre o parâmetro de interesse θ podem ser feitas a partir do **modelo marginal** e, de acordo com a fatoração obtida na expressão (18), T_1 é parcialmente suficiente para θ e, portanto, o termo condicional pode ser abandonado sem ocorrer **perda de informação**, pois este último termo não depende de θ.

Por outro lado, o Princípio do Condicionamento é um recurso conhecido na teoria estatística como método de testes condicionais. Inferências parciais sobre θ são realizadas com base na distribuição condicional dado uma estatística T_2, parcialmente suficiente para o parâmetro "nuisance" Φ. Segue da definição de suficiência que este modelo condicional não depende de Φ. Em adição, T_2 é parcialmente ancilar para θ (observe a expressão 19) e, portanto, o termo marginal pode ser des

prezado pois é não informativo.

Estamos usando os conceitos de "informação" e "não informação" em correspondência aos conceitos de suficiência parcial e ancilaridade parcial. Destacamos porém, que existem muitas discussões sobre estas afirmações. Não é nosso objetivo entretanto levantar tais controvérsias.

Basicamente, os Princípios da Marginalização e Condicionamento consideram uma simplificação do modelo original para um modelo dependendo apenas do parâmetro θ, em que temos interesse. Uma justificativa para estes métodos de eliminação é que a substituição do modelo original por um reduzido não leva à perda de informação sobre θ, isto é, devido ao tipo de fatoração obtida nos dois casos, as funções de verossimilhança original e reduzida são **proporcionais**, ou seja, carregam a mesma quantidade de informação sobre θ.

É fácil verificar que a quantidade de informação, segundo Fisher, sobre θ no termo condicional da expressão (18) é nula:

$$E\left[\frac{\partial}{\partial\theta} \ln P_2(x/T_1, \phi)\right]^2 = 0$$

Do mesmo modo, o termo marginal da expressão (19):

$$E\left[\frac{\partial}{\partial\theta} \ln P_1(T_2/\phi)\right]^2 = 0$$

Ainda, em relação ao tipo de fatoração obtida nos dois Princípios, observamos que a estatística T_1 (T_2) conduz a uma partição muito especial do modelo original. Quando isto é possível, define-se que esta estatística estabelece um corte de **Barndorff-Nielsen** na função (Barndorff-Nielsen, 1978),

ou seja, existem dois subparâmetros de variação independente e complementar (θ, Φ), tal que, a distribuição marginal de T_1 (T_2) depende apenas de θ (Φ) e a distribuição dos dados condicionada nesta estatística depende apenas de Φ (θ). Estas situações apresentam a vantagem de que inferências específicas sobre θ e Φ podem ser feitas independentemente.

Passaremos a descrever um resultado bastante interessante apresentado no artigo de Andersen (1967). Suponha que para um certo problema de natureza multiparamétrica existem estatísticas T_1 e T_2 tal como definidas anteriormente, isto é:

$$P(x/\theta,\Phi) = P_1(T_1/\theta)\, P_2(x/T_1,\Phi)$$
$$= P_1(T_2/\Phi)\, P_2(x/T_2,\theta)$$

Com o tipo de fatoração indicado acima, o autor demonstra que as estatísticas T_1 e T_2 são **independentes** se, e somente se, a família de distribuições marginais da estatística parcialmente suficiente minimal para o parâmetro "nuisance" é **completa.** De acordo com nossa notação, se, e somente se, a família $(P_1(T_2/\Phi);\ \Phi \in \Theta)$ for completa.

Note, que se esta condição estiver satisfeita os modelos marginal e condicional que dependem apenas do parâmetro de interesse θ, conduzem à mesma distribuição, isto é:

$$P_2(x/T_2,\theta) = P(T_1/T_2,\theta) = P_1(T_1/\theta)$$

A primeira igualdade segue do fato de que a propriedade de suficiência minimal de T_1 (com respeito a θ) também se aplica à família de distribuições condicionais dado T_2. Logo, se desejamos realizar inferência objetiva sobre o parâmetro θ podemos restringir nossa atenção à distribuição condicional de T_1/T_2 e, como estas estatísticas são independentes, segue a segunda igualdade. Portanto, os dois métodos de eliminar parâ

metros "nuisance", a partir de modelos condicionais e marginais, utilizam a mesma distribuição.

Por outro lado, este resultado pode ser utilizado para caracterizar uma outra situação: suponha que para um certo problema de natureza multiparamétrica existam estatísticas G_1 e G_2, tais que

(i) $P(x/\theta,\Phi) = P_1(G_1/\theta)\ P_2(x/G_1,\theta,\Phi)$
isto é, G_1 é parcialmente ancilar para o parâmetro "nuisance" Φ, mas não é parcialmente suficiente para o parâmetro de interesse θ;

(ii) $P(x/\theta,\Phi) = P_1(G_2/\theta,\Phi)\ P_2(x/G_2,\theta)$
isto é, G_2 é parcialmente suficiente para Φ, mas não é parcialmente ancilar para θ.

Realmente, as estatísticas G_1 e G_2 não estabelecem um corte de Barndorff-Nielsen no modelo original. Contudo, o termo marginal $P_1(G_1/\theta)$ e o termo condicional $P_2(x/G_2,\theta)$, dependem apenas do parâmetro de interesse θ, ou seja, eliminam o parâmetro "nuisance" Φ.

Neste sentido, as estatísticas G_1 e G_2 são independentes se, e somente se, para $\theta \in \Theta$ fixado e arbitrário, a família de distribuições marginais da estatística parcialmente suficiente minimal para o parâmetro "nuisance" $(P_1(G_2/\theta,\Phi);\ \Phi \in \Theta)$ é **completa**. Estando esta condição satisfeita, os modelos marginal e condicional podem ser reduzidos à mesma distribuição na realização de inferências objetivas sobre θ.

Um exemplo simples ilustra a aplicação deste resultado.

Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias Normais

independentes e identicamente distribuídas, com média μ e variância σ^2. Suponha que temos interesse em testar apenas o valor de μ. Um teste comumente usado pode ser construído, com base na quantidade:

$$t_\mu = \sqrt{n-1}\sqrt{n}\ (\bar{x} - \mu)/s$$

onde $\bar{x} = \sum_{i=1}^{n} x_i/n$; $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$

que tem distribuição "t" com (n-1) graus de liberdade. A vantagem real desta distribuição é naturalmente, que ela independe de σ^2, isto é, t_μ é, parcialmente ancilar para σ^2. Contudo, t_μ não é parcialmente suficiente para μ. Por outro lado, não é difícil verificarmos que o modelo condicional também conduz à estatística "t" de Student.

Sabemos que $(\bar{x}, s^2)$ são conjuntamente suficientes minimais para o vetor de parâmetros (μ, σ^2). Além disso,

$$s_\mu^2 = \sum_{i=1}^{n} (x_i - \mu)^2$$

é parcialmente suficiente minimal para σ^2. No entanto, não é parcialmente ancilar para μ.

Sem perda de generalidade, satisfazendo à propriedade de suficiência global, isto é, para o espaço paramétrico total, a distribuição da amostra $(X_1, X_2, \ldots, X_n)$ condicionada na estatística s_μ^2 pode ser definida como a distribuição condicional de $(\bar{x}, s^2)$ dado s_μ^2, que independe de σ^2.

Por outro lado, existe uma correspondência biunívoca entre (t_μ, s_μ^2) e $(\bar{x}, s^2)$ e, portanto, a distribuição condicional pode ser definida por t_μ dado s_μ^2. Segue que t_μ e s_μ^2 são independentes, o que além de ser uma propriedade pecu

liar à distribuição Normal, pode ser verificada através da completividade da estatística suficiente minimal s^2_μ com respeito ao parâmetro σ^2, para $\mu \in R$ fixado e arbitrário.

Desta forma, a distribuição condicional de t_μ dado s^2_μ coincide com a distribuição marginal de t_μ. Assim, inferências parciais sobre μ podem ser feitas a partir da estatística "t" de Student. Entretanto, neste exemplo não foi levado em conta o tipo de fatoração estabelecido no Princípio do Condicionamento ou da Marginalização para que seja possível avaliar a precisão destas estimativas. Realmente, o problema de eliminar o parâmetro de locação μ da distribuição Normal não é simples (Basu, 1977).

A presença de parâmetros "nuisance" em vários tipos diferentes de modelos que surgem na prática, nem sempre admitem um corte de Barndorff-Nielsen. Em função disso, torna-se necessário revermos outros casos que consideram argumentos menos rigorosos.

Muitas definições de suficiência e ancilaridade na presença de parâmetros "nuisance" são encontradas na literatura (Fraser, 1956; Cox, 1958 e 1975; Andersen, 1967 e 1970 ; Sprott, 1975; Basu, 1977 e 1979; Godambe, 1980). No entanto, muitas das propostas apresentadas são de aplicação limitada, como a definição de Fraser que considera espaços paramétricos com subconjuntos de parâmetros de variação independente e complementar; ou de difícil verificação como a condição de pivotalidade de Cox. As definições de Andersen e Godambe são justificadas adotando o conceito de informação de Fisher, o que é bastante criticado por Basu.

O artigo de Andersen (1970), analisa as propriedades assintóticas dos **estimadores de máxima verossimilhança condicional.** O autor demonstra que sob certas condições de regu

laridade, estes estimadores são sempre **consistentes**, com **distribuição assintótica Normal**. Por outro lado, utilizando o limite mínimo para variância assintótica de estimadores consistentes (limite mínimo de Cramér-Rao), não é possível garantir a propriedade de **eficiência**, exceto nos casos denominados de **ancilaridade forte ou fraca** das estatísticas suficientes mínimais para os parâmetros "nuisance".

A contribuição de Andersen neste artigo é fundamental em estabelecer critérios para a validade e precisão da Inferência Parcial. Relativamente à definição de **ancilaridade forte**, esta se refere à ancilaridade parcial descrita no princípio do Condicionamento. Deste modo, se existir uma estatística T que promove um corte de Barndorff-Nielsen no modelo, os estimadores derivados da função de verossimilhança parcial, **são eficientes**. Este resultado confirma a menção que fizemos anteriormente considerando como proporcionais as funções de verossimilhança original e reduzida, nos dois Princípios.

Já em relação à condição de **ancilaridade fraca**, esta engloba uma classe mais ampla de aplicações, conservando ainda, a propriedade de **eficiência** dos estimadores de máxima verossimilhança condicional. Considere as seguintes condições:

(i) no espaço paramétrico Θ, os parâmetros θ e Φ não são de variação independente e complementar;

(ii) existe uma estatística T tal que:

$$P(x/\theta,\Phi) = P_1(T/\theta,\Phi)\ P_2(x/T,\theta) \qquad (20)$$

Não é difícil verificar, que com esta fatoração a estatística T não estabelece um corte de Barndorff-Nielsen no modelo original.

Andersen (1970) discute que mesmo considerando a fa

toração apresentada em (20) é possível obter estimadores de máxima verossimilhança condicional eficientes, utilizando apenas o termo P_2, se, a família de distribuições marginais de T tiver a seguinte propriedade: suponha que para qualquer conjunto de valores (θ_0, Φ_0) pertencentes a Θ e para qualquer outro valor de θ existe um ponto $\Phi = \Phi(\theta)$ tal que:

$$P_1(T/\theta, \Phi(\theta)) = P_1(T/\theta_0, \Phi_0)$$

Para Sverdrup (1966), citado por Andersen (1970), esta é uma forma natural de definir ancilaridade de T com respeito a θ na presença do parâmetro Φ. A distribuição de T depende de ambos, θ e Φ, mas a família $(P_1(T/\theta, \Phi)$; $\Phi = \Phi(\theta))$ é a mesma qualquer que seja o valor de θ. Isto significa, que observações da variável aleatória T não fornecem nenhuma informação sobre θ que não dependa completamente de uma especificação do valor de Φ.

O autor obtém critérios que facilitam a verificação desta propriedade, denominada de ancilaridade fraca. Porém, não é nosso objetivo abordar este assunto com mais detalhe pois, como veremos no próximo capítulo, trataremos de problemas envolvendo parâmetros de variação independente e complementar.

Por outro lado, uma outra definição de ancilaridade na presença de parâmetros "nuisance" merece ser destacada, principalmente devido à sua aplicação em muitos casos de importância prática, particularmente, na análise de tabelas de contingência 2x2.

Vamos retomar a fatoração apresentada em (20). Considerando que no espaço dos parâmetros **θ e Φ são de variação independente e complementar**, Basu (1977) define a estatística T como **especificamente suficiente para Φ.**

Neste sentido, Godambe (1980) desenvolveu o conceito de **G-ancilaridade**, que se refere a uma generalização do Princípio do Condicionamento, garantindo que, de acordo com a fatoração obtida em (20), o modelo condicional $P_2(x/T,\theta)$ e o modelo original $P(x/\theta,\Phi)$, contêm a mesma informação sobre o parâmetro θ se, e somente se, para cada θ pertencente a Θ, a classe de distribuições marginais de T, $\{P_1(T/\theta,\Phi), \Phi \in \Theta\}$, é completa para cada valor fixado de θ.

Alguns pontos importantes precisam ser tratados com cuidado. Em primeiro lugar, a propriedade de G-ancilaridade é, sem dúvida, uma alternativa que procura justificar que inferências possam ser realizadas com base apenas no modelo condicional, abandonando um termo que depende do valor do parâmetro de interesse. Resta saber, se esta propriedade garante que o modelo marginal $P_1(T/\theta,\Phi)$ é não informativo com respeito a θ.

A questão que se coloca é, portanto, a respeito da eficiência dos estimadores de máxima verossimilhança parcial obtidos a partir do termo condicional $P_2(x/T,\theta)$, eliminando da análise o termo $P_1(T/\theta,\Phi)$.

Como mencionamos anteriormente, devido a Andersen, os únicos casos que fornecem estimadores de máxima verossimilhança condicional eficientes consistem de ancilaridade forte (que equivale à obtenção de um corte de Barndorff-Nielsen) ou ancilaridade fraca, que não se restringe a θ e Φ de variação independente.

Deste modo, torna-se interessante a seguinte consideração. Supondo a estatística T G-ancilar com respeito a θ, decorre que:

- T é especificamente suficiente para Φ;

- Para cada valor fixado de $\theta \in \Theta$ a classe de distribuições marginais de T é completa.

Generalizando o Teorema de Lehmann-Scheffé, estas condições estabelecem que para $\theta \in \Theta$, fixado e arbitrário, T é **estatística especificamente suficiente minimal** para a família $\{P(x/\theta,\Phi) ; \Phi \in \Theta\}$ e, é única.

Assim, T é função de todas as outras estatísticas especificamente suficientes para inferir sobre Φ. Consequentemente, é a estatística mais resumida que mantém a propriedade de suficiência. Logo, se T contiver alguma informação sobre θ, esta será mínima se comparada à informação contida nas demais estatísticas especificamente suficientes com respeito a Φ. Realmente se concordarmos em adotar um modelo reduzido do tipo $P_2(x/T,\Phi)$, é melhor que seja utilizada a estatística G-ancilar T. Neste caso, se houver perda de informação sobre θ ao ser abandonado o fator $P_1(T/\theta,\Phi)$, tal perda será a menor possível.

Com as definições dadas até aqui, no próximo capítulo são propostas algumas medidas de interesse em epidemiologia, o que determina o tipo de reparametrização mais conveniente e, em cada caso, é verificado como ocorre a fatoração da função de verossimilhança. Ressaltamos contudo, que o procedimento de escolha do parâmetro "nuisance" fica a critério do pesquisador e não da natureza do problema. Toda escolha de parâmetro que conduza a uma transformação biunívoca poderá ser adotada. Logicamente, a cada tipo de reparametrização corresponde uma específica fatoração da função de verossimilhança . Este, é um dos principais geradores de críticas ao Método da Redução.

CAPÍTULO V

ALGUMAS PROPOSTAS DE MEDIDAS DE ASSOCIAÇÃO ENTRE A DOENÇA E O FATOR DE RISCO

Quando um estatístico se propõe a interpretar uma tabela de contingência 2x2 gerada de estudos observacionais, a primeira preocupação que surge é sobre qual medida de associação será adotada. Na verdade, esta decisão deve conter a participação de dois especialistas, do estatístico que analisará os dados e do epidemiologista que os coletou com uma específica finalidade. Por este motivo, qualquer medida deste tipo deve preenchèr a dois requisitos: ser "tratável" do ponto de vista analítico, isto é, poder ser estimada e testada e, ser também de interesse clínico.

Goodman e Kruskal (1954 e 1959), discutem em detalhe várias medidas e índices de associação usados principalmente na interpretação de tabelas rxs, geradas de estudos em Ciências Sociais. Uma vasta referência bibliográfica é citada, sendo indicadas também medidas tradicionais baseadas na estatística Qui-Quadrado.

Edwards (1963), descreve propriedades convenientes a uma medida de associação em estudos transversais. O autor

apresenta duas proposições que são essenciais para definir uma medida de associação em uma tabela 2x2. Para exemplificar, suponha a notação e formato utilizados na Tabela I.6:

Proposição 1. A medida de associação entre as variáveis A e D deve ser uma função das proporções p/(p+r) e q/(q+s) ou, alternativamente, das proporções p/(p+q) e r/(r+s);

Proposição 2. Estas duas medidas alternativas devem ser iguais.

Com base nestas proposições, três corolários bastante interessantes podem ser deduzidos:

Corolário 1. A medida de associação deve ser alguma função da razão de produtos cruzados (ps/qr).

Corolário 2. A medida de associação não é influenciada pelos tamanhos relativos dos totais marginais.

Corolário 3. A medida de associação deve ser simétrica, isto é, a associação entre as variáveis A e D quando presentes é a mesma que aquela quando ausentes.

No artigo algumas quantidades são examinadas como a estatística Qui-Quadrado (χ^2) e o coeficiente de contingência de Pearson, onde verifica-se que tais medidas não satisfazem o Corolário 1, sendo que, a estatística χ^2 sofre ainda influência do tamanho relativo dos totais marginais.

Outra dificuldade que surge nestes estudos, é a estimação e teste das medidas de associação propostas. Neste sentido, os procedimentos de Inferência Parcial descritos no capítulo anterior são oportunos, considerando que trata-se do problema de eliminação de parâmetros indesejáveis ("nuisance").

As próximas seções, se referem a propostas de medidas de associação entre duas categorias de eventos, por exemplo, doença e fator de risco, que aparecem com freqüência em textos de Epidemiologia. Cada caso é analisado mediante os critérios do Método da Redução. Não podemos deixar de ressaltar, que outros procedimentos estatísticos poderiam ser considerados, como métodos Bayesianos que não pertencem ao escopo da metodologia clássica de Inferência que estamos adotando, ou mesmo, recursos assintóticos que não utilizam a distribuição exata.

V.1. Risco Relativo

O risco relativo foi definido no capítulo II, expressão (2), como a razão das taxas de incidência da doença para pessoas expostas e não expostas a um fator de risco. De acordo com nossa notação e considerando um estudo prospectivo:

$$RR = \frac{p_1}{p_2}$$

Podemos estabelecer algumas propriedades desta medida:

(i) RR é indeterminada se, e somente se, $p_2 = 0$, isto é, se a probabilidade de um indivíduo que não apresenta o fator desenvolver a doença for nula;

(ii) caso contrário, o valor de RR varia entre $(0,\infty)$ supondo $0 < p_1(p_2) <= 1$, isto é, a probabilidade de um indivíduo que apresenta (não apresenta) o fator desenvolver a doença é sempre positiva não nula;

(iii) RR é 1 se, e somente se, $p_1 = p_2$, o que ocorre quando a probabilidade do indivíduo desenvolver a doença é a mesma na presença ou ausência do fator. Observe, que este caso corresponde à hipótese de homogeneidade no modelo Produto de Binomiais;

(iv) na maioria das situações experimentais temos interesse em hipóteses alternativas (H_1) para as quais $p_1 > p_2$, ou seja, onde o risco da doença para indivíduos com o fator é maior que para aqueles sem o fator. Esta suposição é bastante razoável quando temos evidências que confirmam o fator como sendo de risco para o desenvolvimento da doença. Deste modo, RR > 1.

Já destacamos, que os estudos prospectivos são os únicos que permitem o cálculo direto do RR, devido às populações de risco estarem bem definidas (fixadas). Além disso, a razão de produtos cruzados ("odds-ratio") é uma quantidade que pode ser utilizada em analogia com o risco relativo, independente do modelo epidemiológico. Para análise do RR nos deteremos ao caso prospectivo.

A função de verossimilhança, $L(m,n,p_1,p_2/x,y)$, para os dados gerados por estes estudos, está indicada em (16). Vamos definir a seguinte reparametrização:

$$\begin{cases} \psi = p_1/p_2 \quad RR \\ \Phi = p_2 \end{cases} \Rightarrow \begin{cases} p_1 = \Phi\ \psi \\ p_2 = \Phi \end{cases}$$

onde ψ (risco relativo) é o parâmetro de interesse e Φ é o parâmetro "nuisance", adotado por conveniência.

Podemos reescrever a expressão (16) em termos destes novos parâmetros:

$$L(m,n,p_1,p_2/x,y) = L(m,n,\psi,\Phi/x,y)$$

$$= \binom{m}{x}\binom{n}{y}\ \psi^x\ (1-\psi\Phi)^{(m-x)}\ \Phi^{(x+y)}\ (1-\Phi)^{(n-y)} \qquad (21)$$

para $1 <= \psi < \infty$ (desde que $p_1 >= p_2$)

$0 < \Phi < 1$

Pode ser verificado facilmente que a transformação proposta para o espaço original dos parâmetros do modelo prospectivo, é **biunívoca.** O determinante do jacobiano para esta transformação é diferente de zero para $p_2 > 0$.

$$\det J(p_1,p_2;\ \psi,\Phi) = \begin{vmatrix} \frac{\partial p_1}{\partial \psi} & \frac{\partial p_1}{\partial \Phi} \\ \frac{\partial p_2}{\partial \psi} & \frac{\partial p_2}{\partial \Phi} \end{vmatrix} = \Phi \neq 0;\ \text{desde que } p_2 > 0$$

Também, os novos parâmetros são **identificáveis,** pois cada par (ψ,Φ) determina um único valor da função de verossimilhança em (21). Além disso, são de **variação independente** com $1 <= \psi < \infty$ e $0 < \Phi < 1$.

No modelo Produto de Binomiais, parametrizado por (p_1,p_2) o interesse é testar:

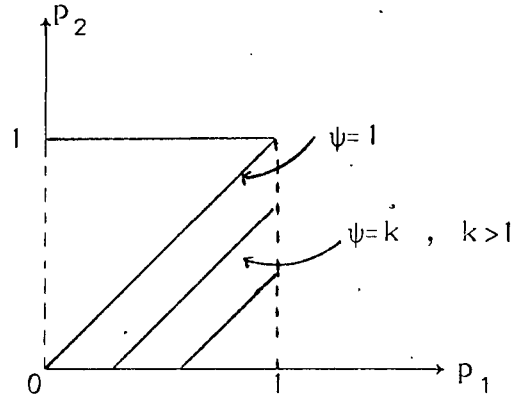
$H_0: p_1 = p_2$ x $H_1: p_1 > p_2$

Através da reparametrização (ψ,Φ), as hipóteses em teste tomam a seguinte forma:

$H_0: \psi = 1$ x $H_1: \psi > 1$ (Φ não especificado)

É interessante ilustrarmos como os valores de ψ se comportam no espaço dos parâmetros (p_1,p_2).

FIGURA V.1. Indicação de valores do parâmetro ψ no espaço(p_1,p_2)



Quanto maior o valor de ψ, mais evidências temos contra a hipótese de homogeneidade ($\psi = 1$), ou seja, quanto maior a razão entre as taxas de incidência, mais pronunciado é o efeito do fator de risco no desenvolvimento da doença.

Temos interesse em fazer inferências apenas sobre o parâmetro ψ. Considerando que o experimento em estudo está indexado por (ψ,Φ), o problema é encontrar um **modelo reduzido** que seja função somente do risco relativo (ψ), eliminando-se o parâmetro "nuisance" Φ.

A fatoração da função de verossimilhança em (21) in

dica que, para cada valor de $\psi > 1$, a estatística parcialmente suficiente minimal para Φ são os dados (x,t), onde $t=x+y$. Portanto, a reparametrização do modelo em termos de (ψ,Φ) não é apropriada para a utilização do Método da Redução, pois não é possível eliminar Φ através do Princípio do Condicionamento.

Por outro lado, para cada valor de Φ no intervalo $(0,1)$, X é estatística parcialmente suficiente minimal para ψ mas não é parcialmente ancilar para Φ pois, da expressão (21) temos que a distribuição marginal de X depende de Φ:

$$X \sim Bin(m; \Phi\psi)$$

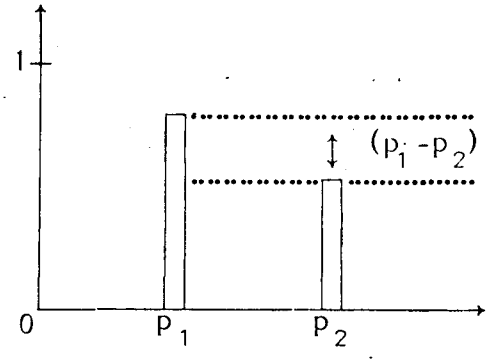
Deste modo, também o critério da marginalização não pode ser aplicado.

Assim, apesar de o risco relativo ser muito informativo ao pesquisador interessado em estudar como o desenvolvimento da doença é afetado pela presença do fator, a análise estatística baseada nesta quantidade não pode ser realizada com sucesso através dos testes exatos.

V.2. Risco Adicional: $p_1 - p_2$

Em um estudo prospectivo, a medida (p_1-p_2) representa a diferença nas taxas de incidência da doença para os grupos exposto e não exposto ao fator e é de bastante interesse para os epidemiologistas. Contudo, para estudos retrospectivos, trata-se simplesmente da diferença entre as proporções de indivíduos que apresentam o fator nos grupos caso e controle, o que não traz muita informação sobre a etiologia da doença. A Figura V.2 ilustra a medida representada pelo risco adicional.

FIGURA V.2. Indicação da medida risco adicional



Algumas propriedades desta medida são:

(i) é linear nos parâmetros, o que facilita cálculos de distribuições e outros;

(ii) varia no intervalo [-1,1];

(iii) é nula se, e somente se, $p_1 = p_2$;

(iv) em geral, temos interesse em hipóteses definidas como $p_1 > p_2$, isto é, a ocorrência da doença é maior no grupo com o fator e, nessas condições, esta medida assume somente valores positivos em (0,1].

Para análise desta quantidade, vamos definir a seguinte reparametrização:

$$\begin{cases} \xi = p_1 - p_2 \\ \Phi = p_2 \end{cases} \Rightarrow \begin{cases} p_1 = \xi + \Phi \\ p_2 = \Phi \end{cases}$$

onde ξ é o parâmetro de interesse e Φ o parâmetro "nuisance".

A expressão (16) pode ser reescrita em termos destes

novos parâmetros:

$$L(m,n,p_1,p_2/x,y) = L(m,n,\xi,\Phi/x,y)$$

$$= \binom{m}{x}\binom{n}{y}(\xi+\Phi)^x(1-\xi-\Phi)^{(m-x)}\Phi^y(1-\Phi)^{(n-y)} \qquad (22)$$

para $0 <= \xi <= 1$ (desde que $p_1 >= p_2$)

$0 < \Phi < 1$

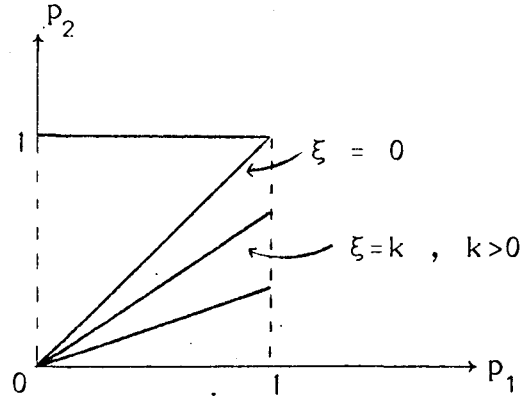
Pode ser verificado que a transformação proposta no espaço dos parâmetros é **biunívoca.** O determinante do jacobiano é diferente de zero para todo valor de (p_1,p_2). Além disso, (ξ,Φ) é identificável e de variação independente.

Através desta reparametrização, a hipótese em teste (H_0: $p_1=p_2$ x H_1: $p_1 > p_2$) toma a seguinte forma:

H_0: $\xi = 0$ x H_1: $\xi > 0$ (Φ não especificado)

A figura a seguir, mostra como os valores de ξ se comportam no espaço dos parâmetros (p_1,p_2).

FIGURA V.3. Indicação de valores do parâmetro ξ no espaço(p_1,p_2)



Note, que quanto menor o coeficiente angular da reta $\xi = k$ ($p_1 = k+p_2$) maior é a relação entre a exposição ao

fator e a chance de desenvolver a doença. Deste modo, em termos dos novos parâmetros (ξ,Φ), temos interesse em realizar inferências apenas sobre ξ.

Como pode ser verificado na expressão da $L(\xi,\Phi)$, para cada valor fixado de $\xi > 0$, a estatística parcialmente suficiente minimal para Φ são os dados $(x;y)$ e, portanto, não há como eliminar Φ da função de verossimilhança. Não existe uma estatística, função dos dados, suficiente para Φ, o que nos impede de obter os termos da $L(\xi,\Phi)$ que dependem apenas de ξ, derivados da distribuição condicional dos dados nesta estatística.

Além disso, não é possível obter uma estatística ao mesmo tempo parcialmente suficiente para ξ e parcialmente ancilar para Φ, no sentido do Princípio da Marginalização. Observe, que para cada valor de $0 < \Phi < 1$, X é estatística parcialmente suficiente minimal para ξ, mas não é parcialmente ancilar para Φ; a distribuição marginal de X depende de Φ:

$$X \sim Bin(m;\ \xi+\Phi)$$

Assim, também para a reparametrização proposta o método condicional de Fisher, isto é, o critério da marginalização, não pode ser aplicado. Portanto, apesar de o risco adicional $(p_1 - p_2)$ ser de importância nos estudos prospectivos, a análise estatística baseada na metodologia em questão não é apropriada.

Neste caso, em particular, é comum a utilização de um método alternativo baseado na **distribuição assintótica** da estimativa de máxima verossimilhança de $(p_1 - p_2)$.

De acordo com o modelo Produto de Binomiais já descrito para os dados gerados de um estudo prospectivo (ou mesmo retrospectivo), temos que:

$$X \sim Bin(m;p_1) \qquad e \qquad Y \sim Bin(n;p_2)$$

Sejam $\hat{p}_1$ e $\hat{p}_2$ as estimativas de máxima verossimilhança não condicional dos parâmetros p_1 e p_2, respectivamente. Então:

$$\hat{p}_1 = x/m \qquad e \qquad \hat{p}_2 = y/n$$

onde x e y são os valores amostrais observados na realização do experimento.

Pode ser mostrado que a distribuição assintótica dos estimadores $\hat{p}_1$ e $\hat{p}_2$ é Normal. Assim,

$$\hat{p}_1 \sim N(p_1 \; ; \; p_1(1-p_1)/m) \qquad e \qquad \hat{p}_2 \sim N(p_2 \; ; \; p_2(1-p_2)/n)$$

A estimativa de máxima verossimilhança de $\xi = p_1 - p_2$ é dada por:

$$\hat{\xi} = \hat{p}_1 - \hat{p}_2$$

com a seguinte distribuição assintótica:

$$\hat{\xi} \sim N(p_1-p_2 \; ; \; p_1(1-p_1)/m + p_2(1-p_2)/n)$$

Desta forma, um teste de $H_0: \xi = 0$ x $H_1: \xi > 0$ pode ser construído a partir da **distribuição Normal Padrão;** isto é, da estatística z definida, sob H_0, como:

$$z = \frac{\hat{\xi} - E(\hat{\xi})}{\hat{dp}(\hat{\xi})} \sim N(0;1)$$

onde E(.) e $\hat{dp}(.)$ são, respectivamente, a esperança e a estimativa do desvio padrão de $\hat{\xi}$. Sob H_0, a estimativa da variância de $\hat{\xi}$ é:

$$\hat{V}(\hat{\xi}) = \hat{p}(1-\hat{p}) \; (1/m + 1/n)$$

$$\hat{p} = (m\hat{p}_1 + n\hat{p}_2)/(m + n)$$

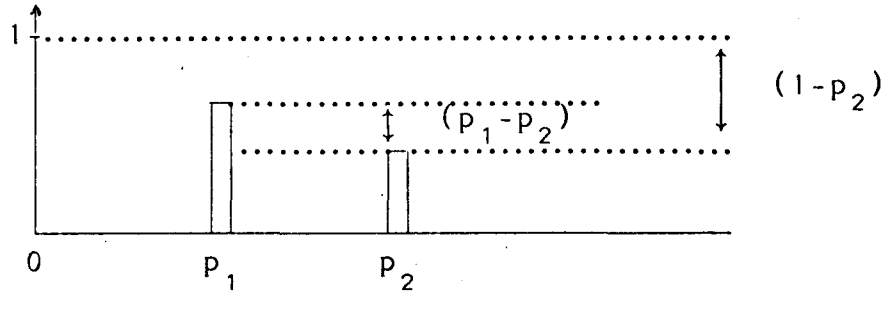
A utilização deste procedimento assintótico apresenta vantagens estatísticas, pois a obtenção da distribuição exa

ta para o parâmetro $\xi = p_1 - p_2$ não foi possível através da metodologia em questão. Além disso a estimativa $\hat{\xi} = \hat{p}_1 - \hat{p}_2$ define uma combinação linear de variáveis aleatórias Normais o que facilita o cálculo de distribuições de probabilidades. Contudo, deve ser chamada a atenção para o fato de que a expressão dada acima para $\hat{V}(\hat{\xi})$ é obtida sob a hipótese H_0 e, portanto, não é útil para a construção de intervalos de confiança; neste caso, deve ser utilizada as estimativas das variâncias parciais de $\hat{p}_1$ e $\hat{p}_2$.

V.3. Risco Adicional Relativo: $(p_1-p_2)/(1-p_2)$

Como nos casos apresentados anteriormente, existem situações em que o pesquisador ao realizar um **estudo prospectivo**, se propõe a investigar se a diferença entre as probabilidades do indivíduo desenvolver a doença quando o fator de risco está presente e ausente (p_1-p_2), tem a mesma ordem se comparada com a probabilidade do indivíduo não desenvolver a doença quando o fator de risco está ausente $(1-p_2)$. Esta quantidade é denominada risco adicional relativo e está ilustrada na Figura V.4.

FIGURA V.4. Indicação da medida risco adicional relativo $(p_1-p_2) / (1-p_2)$



Vamos analisar o que acontece no modelo quando escolhemos a seguinte reparametrização:

$$\begin{cases} \pi = (p_1-p_2)/(1-p_2) \\ \Phi = p_2 \end{cases} \Rightarrow \begin{cases} p_1 = \pi + \Phi(1-\pi) \\ p_2 = \Phi \end{cases}$$

onde π é o parâmetro de interesse e Φ o parâmetro "nuisance".

Pode ser verificado que a transformação proposta é **biunívoca**, pois o determinante do jacobiano é diferente de zero para $p_2 < 1$, restrição esta, que elimina o ponto de indeterminação do parâmetro π. Além disso, $0 <= \pi <= 1$ (desde que $p_1 >= p_2$) e $0 <= \Phi < 1$ e, portanto, são de **variação independente** no espaço (π,Φ).

Assim, cada par (π,Φ) determina um único valor da função de verossimilhança, que pode ser escrita como:

$$L(m,n,p_1,p_2/x,y) = L(m,n,\pi,\Phi/x,y)$$

$$= \binom{m}{x}\binom{n}{y} (\pi+\Phi-\pi\Phi)^x [(1-\pi)(1-\Phi)]^{(m-x)} \Phi^y (1-\Phi)^{(n-y)} \quad (23)$$

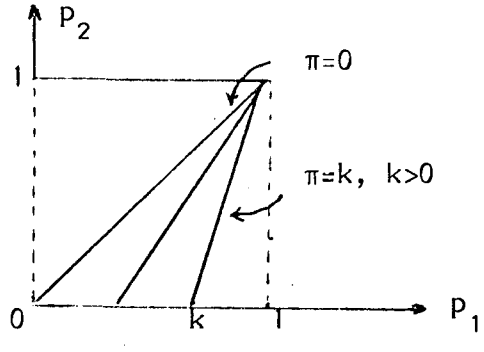
para $0 <= \pi < 1$ e $0 < \Phi < 1$

Neste caso, a hipótese H_0: $p_1=p_2$ x H_1: $p_1>p_2$, pode ser escrita em termos de (π,Φ):

H_0: $\pi = 0$ x H_1: $\pi > 0$ (Φ não especificado)

A Figura V.5 mostra como os valores de π se distribuem no espaço dos parâmetros (p_1,p_2).

FIGURA V.5. Indicação de valores do parâmetro π no espaço (p_1,p_2)



O próximo passo é observar como ocorre a faforação da função de verossimilhança. A expressão da $L(\pi,\Phi)$ indica que para cada valor fixado de $\pi > 0$; a estatística parcialmente suficiente minimal para Φ é (x,y) e, portanto, não temos como eliminar o parâmetro Φ através do critério condicional.

Por outro lado, para cada valor fixado de Φ sobre o intervalo $(0,1)$, X é estatística parcialmente suficiente minimal para π, mas não é parcialmente ancilar para Φ. A distribuição marginal de X depende de Φ:

$$X \sim Bin(m;\ \pi+\Phi-\pi\Phi)$$

Novamente, também a reparametrização (π,Φ) não permite a realização de inferência parcial sobre π.

Note, que a quantidade $\pi = (p_1-p_2)/(1-p_2)$ não tem nenhuma importância clínica relevante quando se trata de um estudo retrospectivo, pois p_1 e p_2, neste caso, não definem taxas de incidência da doença, mas simplesmente, proporções da presença do fator nos dois grupos, com a doença e controle.

V.4. Razão de Produtos Cruzados ("odds-ratio")

A razão de produtos cruzados é uma medida de associação bastante discutida por vários autores. Atualmente, sua aplicação na análise de tabelas de contingência tem sido requerida não apenas pelos estatísticos, mas pelos próprios epidemiologistas, que a reconhecem como efetiva na comparação dos riscos de desenvolvimento da doença nas populações exposta e não exposta ao fator.

Como foi visto em capítulos anteriores, além de ocorrer em analogia com o risco relativo, a razão de produtos cruzados tem ainda a vantagem de parametrizar o modelo condicional (15), utilizado na realização de inferências em procedimentos de teste, independente da distribuição de probabilidade proposta para os dados.

No prosseguimento, os resultados obtidos no capítulo III sobre a redução de modelos completos a termos condicionais, são novamente colocados, concluindo o problema de encontrar uma medida que permita uma fatoração apropriada da função de verossimilhança.

Inicialmente, considerando um **estudo transversal,** vamos definir a seguinte reparametrização no espaço (p,q,r), a qual conduz a uma fatoração bastante interessante da função de verossimilhança:

$$\begin{cases} \Theta = \dfrac{p/r}{q/s} \\ \psi = \dfrac{q}{q+s} \\ \Phi = p+r \end{cases} \qquad \begin{cases} p = \dfrac{\psi\,\Theta}{1-\psi+\Theta\psi}\,\Phi \\ q = \psi\,(1-\Phi) \\ r = \dfrac{(1-\psi)\,\Phi}{1-\psi+\Phi\psi} \end{cases}$$

onde, Θ é o parâmetro de interesse, a razão de produtos cruzados e (ψ,Φ) são os parâmetros "nuisance".

Não é difícil constatar que a condição de **bijeção** entre os espaços (p,q,r) e (Θ,ψ,Φ) está satisfeita. Ainda, o novo espaço paramétrico é tal, que $\Theta > 0$, $0 < \psi, \Phi < 1$ e, neste caso, temos que Θ, ψ e Φ são de **variação independente.**

Algumas propriedades de Θ já foram destacadas no capítulo I:

(i) Θ é sempre positivo. Somente é nulo se $p = 0$. É indeterminado para $r = 0$ ou $s = 0$ ou $q = 0$;

(ii) $\Theta = 1$ se, e somente se, $p/r = q/s$, o que equivale a

$p/(p+r) = q/(q+s)$, ou seja o risco de desenvolver a doença é o mesmo nos dois grupos, exposto e não exposto ao fator;

(iii) Em geral, existe razão para assumir $p/(p+r) > q/(q+s)$ isto é, o risco da doença é maior no grupo exposto ao fator. Portanto, adotamos hipóteses alternativas (H_1) definidas como $\Theta > 1$.

Desta forma, quando o pesquisador tem interesse em estudar se o fator é de risco para a doença, o que equivale, neste caso, ao teste de independência, as hipóteses podem ser escritas em termos de (Θ, ψ, Φ) como:

$H_0: \Theta = 1 \quad x \quad H_1: \Theta > 1 \qquad (\psi \text{ e } \Phi \text{ não especificados})$

Neste sentido, nosso objetivo é encontrar um modelo reduzido que dependa apenas de Θ e, portanto, que elimine ψ e Φ.

A expressão (17) da função de verossimilhança para os dados de um estudo transversal, pode ser reescrita em termos destes novos parâmetros, além do que, recordando que $y=t-x$:

$$L(N,p,q,r/x,y,m) = L(N,p,q,r/x,m,t) = L(N,\Theta,\psi,\Phi/x,m,t) =$$

$$= \frac{N!}{x!\,(t-x)!\,(m-x)!\,(n-t+x)!}\,\Theta^x\left(\frac{\psi}{1-\psi}\right)^t\left(\frac{\Phi}{1-\Phi}\right)^m\frac{[(1-\psi)(1-\Phi)]^N}{(1-\psi+\Theta\psi)^m} \qquad (24)$$

$\Theta >= 1$ (desde que $p/(p+r) >= q/(q+s)$

$0 < \psi, \Phi < 1$

A fatoração obtida para $L(\Theta,\psi,\Phi)$ indica que para cada valor fixado de $\Theta >= 1$, (m,t) é estatística conjuntamente suficiente para o subparâmetro (ψ,Φ). Conclue-se, que a dis

tribuição de X condicionada nas marginais (m,t) depende apenas de θ.

Vamos analisar este resultado com mais detalhe. Podemos estabelecer a seguinte fatoração da distribuição conjunta (x,m,t):

$$P(x,m,t/N,\theta,\psi,\Phi) = P(m,t/N,\theta,\psi,\Phi)\ P(x/N,m,t,\theta,\psi,\Phi)$$

$$= P(m,t/N,\theta,\psi,\Phi)\ P(x/N,m,t,\theta) \qquad (23)$$

A segunda igualdade decorre da propriedade de suficiência da estatística (m,t). Em particular, trata-se de estatística **especificamente suficiente** para (ψ,Φ). Por outro lado:

$$P(m,t/N,\theta,\psi,\Phi) = \sum_{x} P(x,m,t/N,\theta,\psi,\Phi)$$

$$= \binom{N}{m}\left(\frac{\psi}{1-\psi}\right)^{t}\left(\frac{\Phi}{1-\Phi}\right)^{m}\frac{[(1-\psi)(1-\Phi)]^{N}}{(1-\psi+\theta\psi)^{m}}\sum_{u=k}^{1}\binom{m}{u}\binom{n}{t-u}\theta^{u}$$

$$= \underbrace{\binom{N}{m}\Phi^{m}(1-\Phi)^{(N-m)}}_{P(m/\Phi)}\ \underbrace{\frac{\psi^{t}(1-\psi)^{(N-t)}}{(1-\psi+\theta\psi)^{m}}\sum_{u=k}^{1}\binom{m}{u}\binom{n}{t-u}\theta^{u}}_{P(t/N,m,\psi,\theta)} \qquad (24)$$

Em resumo, podemos considerar a seguinte fatoração da distribuição Multinomial:

$$P(x,m,t/N,\theta,\psi,\Phi) = P(m/N,\Phi)\ P(t/N,m,\theta,\psi)\ P(x/N,m,t,\theta)$$

Observe, que a marginal m é parcialmente suficiente para Φ e parcialmente ancilar para (θ,ψ). Assim, a estatísti

ca m define um corte de Barndorff-Nielsen para o modelo em termos do parâmetro Φ. Se desejássemos fazer inferências apenas sobre este parâmetro, poderíamos usar somente o modelo parcial $P(m/N,\Phi)$, sem ocorrer perda de informação ao abandonarmos os demais termos, isto é, as estimativas baseadas na verossimilhança marginal seriam eficientes.

Contudo, nosso interesse está em obter informação apenas sobre θ. Foi possível derivar o modelo reduzido - $P(x/N,m,t,\theta)$ que depende somente deste parâmetro. Neste sentido, no capítulo III a mesma redução foi obtida, porém, **a partir do modelo Produto de Binomiais**, ou seja:

$$P(x,y/N,m,p_1,p_2) = P(x,t/N,m,\theta,\psi) = P(t/N,m,\theta,\psi)\ P(x/N,m,t,\theta)$$

onde verificamos ainda, que o modelo condicional $P(x/N,m,t,\theta)$ representa a distribuição Hipergeométrica Generalizada (expressão 15).

Deste modo, surge a seguinte questão: podemos fazer inferência objetiva sobre θ utilizando apenas o modelo - $P(x/N,m,t,\theta)$? ou, equivalentemente, podemos questionar: quando o modelo reduzido $P(t/N,m,\theta,\psi)$ embora dependendo de θ pode ser considerado como não informativo com respeito a θ e, assim, ser eliminado da análise?

A questão que se coloca diz respeito à **ancilaridade** das marginais (m,t) em relação ao parâmetro θ. A expressão (23) mostra que a estatística (m,t) não define um corte de Barndorff-Nielsen para o modelo e, além disso, (θ,ψ,Φ) são de variação independente e complementar e, portanto, não podemos introduzir o conceito de ancilaridade fraca proposto por Andersen (1967). Contudo, esta condição satisfaz em parte a definição de **G-ancilaridade** apresentada por Godambe (1980), resta saber se para θ fixado a família de distribuições marginais

$\{ P(m,t/N,\theta,\psi,\phi), \quad 0 < \psi,\phi < 1 \}$ é completa.

Portanto, é necessário mostrar que se:

$E[f(m,t)] = 0 \quad \Rightarrow \quad f(m,t) = 0$ para todo (m,t)

onde f denota uma função qualquer diferente da função nula.

Utilizando a propriedade de completividade da distribuição Binomial, não é difícil verificar que esta família de distribuições marginais é completa. No Apêndice 1 apresentamos a demonstração deste resultado.

A importância destas conclusões reside no fato de que a propriedade de G-ancilaridade das marginais (m,t) para o parâmetro razão de produtos cruzados (θ), confere validade para obtenção de inferências parciais sobre θ a partir do modelo condicional $P(x/N,m,t,\theta)$. Desta forma, caracteriza-se uma das vantagens, sob o ponto de vista estatístico, que a quantidade razão de produtos cruzados oferece como medida de associação em uma tabela de contingência 2x2. Ainda mais, este é o suporte lógico utilizado para validar o Teste Exato de Fisher.

Salientamos, entretanto, que como mencionado no capítulo anterior, a condição de G-ancilaridade não garante a propriedade de eficiência dos estimadores obtidos a partir do modelo reduzido.

CAPÍTULO VI

MÉTODOS EXATO E ASSINTÓTICO PARA ANÁLISE DE TABELAS DE CONTINGÊNCIA 2x2

Já enfatizamos que nos estudos prospectivos, retrospectivos e transversais, envolvendo a presença ou ausência da doença e fator de risco, os dados obtidos por amostragem podem ser apresentados em uma tabela de contingência 2x2.

Como foi visto em capítulos anteriores, na investigação de uma proposta de medida de associação em tabelas deste tipo, a quantidade **razão de produtos cruzados** (θ) se destaca por sua adequação estatística e clínica.

Independente da situação experimental utilizada, inferências sobre o parâmetro θ podem ser baseadas na distribuição condicional (expressão 15), definida como **Hipergeométrica Generalizada.**

A obtenção do **modelo condicional** como função apenas do parâmetro θ, é garantida pela aplicação do **Método da Redução** juntamente com o conceito de **G-ancilaridade** de Godambe (1980). Contudo, como já salientamos, os estimadores obtidos a partir desta distribuição **não são eficientes**, ou seja, a variância assintótica destes estimadores não atinge o limite mínimo de Cramér-Rao. Devido a isto, muitos estudos têm sido dedicados na busca de estimativas mais precisas. Ainda, para estes casos, Andersen (1970) apresenta uma fórmula de como calcular a variância assintótica de estimadores que não atingem o limite mínimo de Cramér-Rao.

Além disso, uma dificuldade que se apresenta é que os **momentos** da distribuição Hipergeométrica Generalizada envolvem razões de polinômios em θ, o que torna inconveniente o uso de **expressões exatas**. Assim, **expressões aproximadas** são necessárias, as quais são derivadas das **propriedades assintóticas** desta distribuição. Harkness (1965), analisa propriedades da distribuição Hipergeométrica Generalizada como os momentos, critérios de estimação de θ e distribuições limites, Poisson, Binomial e Normal. Cornfield (1956) e Hannan e Harkness (1963), apresentam uma demonstração detalhada da convergência para a Normal.

Mediante estas colocações, o problema da tão polêmica análise de tabelas de contingência 2x2 baseada em **marginais fixadas**, se restringe à estimação do parâmetro θ a partir do modelo condicional (15). Muitas propostas são apresentadas por vários autores, indicando estimativas pontuais ou por intervalos, expressões exatas empregando métodos numéricos ou expressões aproximadas. Outros, consideram métodos **não condicionais**. Em qualquer caso, a preocupação é obter **estimativas mais precisas**. Passamos, portanto, a revisar algumas destas propostas.

VI.1. Teste Exato de Fisher

O Teste Exato de Fisher é um tratamento clássico em Estatística, voltado aos testes de hipóteses unilaterais na análise de tabelas de contingência 2x2. De um modo geral, por ser um teste condicional, é utilizado tanto para estudos de independência como de homogeneidade.

Podemos esquematizar a seguinte tabela:

TABELA VI.1. Indicação das freqüências amostrais de acordo com a presença ou ausência das características A e D.

VARIÁVEL D / VARIÁVEL A	D	D'	TOTAL
A	x	m-x	m
A'	y	n-y	n
TOTAL	t	N-t	N

Observe que a tabela acima pode ser gerada considerando um experimento Multinomial (onde N é o único total fixo), bem como, um experimento descrito pelo produto de duas Binomiais (onde m e n são os totais fixos). No capítulo III, mostramos que estes dois modelos podem ser reduzidos à mesma distribuição, a Hipergeométrica Generalizada, usando o argumento de que as marginais (m,t) são "não informativas".

O objetivo é testar:

H_0: $Pr(D/A) = Pr(D/A')$ x H_1: $Pr(D/A) > Pr(D/A')$

o que corresponde a testar:

$H_0: \theta = 1$ x $H_1: \theta > 1$.

No capítulo II, indicamos a equivalência entre as hipóteses acima definidas.

Deste modo, a probabilidade, sob H_0, de se obter a tabela descrita é dada pela distribuição Hipergeométrica, isto é:

$$P(x/N,m,t,\theta=1) = \frac{\binom{m}{x}\binom{n}{y}}{\binom{N}{t}}$$

onde: $\max(0, t-n) < x < \min(m, t)$

$n = N - m$

$y = t - x$

Neste sentido, a característica básica do Teste Exato de Fisher é a utilização de um **modelo de probabilidade condicional**, cuja validade está suportada pela teoria de Inferência Parcial mencionada nos capítulos anteriores. Por isso, o teste é o mesmo qualquer que seja o delineamento experimental, Binomial ou Multinomial.

Por outro lado, quanto ao **critério de decisão** adotado, este decorre da determinação da **probabilidade exata** (condicionada nas marginais) de obtenção de cada uma das possíveis tabelas de freqüências que favoreceriam a rejeição de H_0 com mais ênfase do que aquele obtido experimentalmente.

Assim, o nível descritivo (α_0) deste teste, considerando H_0 verdadeira e os totais marginais (m,t) fixos, é a soma das probabilidades de ocorrências ainda mais extremas, favoráveis a H_1, do que os dados originais, ou seja:

$$\alpha_0 = \sum_{u=x}^{\min(m,t)} \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{N}{t}}$$

Se esta soma for maior que um **nível de significância (α)** pré-especificado, os dados estarão fornecendo evidências de que H_0 é verdadeira e, portanto, não será possível rejeitar H_0.

Se a soma for menor (ou igual) a α, a hipótese H_0 será rejeitada.

Entretanto, na prática a aplicação da prova de Fisher é restrita a amostras pequenas, do contrário, mesmo com a utilização de computadores é difícil a obtenção de cálculos precisos devido a problemas de aproximação. No caso de grandes amostras e considerando um esquema Multinominal, um recurso é o Teste Qui-Quadrado (χ^2), baseado na estatística

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

onde O_{ij} = freqüência observada na casela ij

E_{ij} = freqüência esperada, sob H_0, na casela ij

Os valores X^2 têm distribuição assintótica Qui-Quadrado com 1 grau de liberdade. Como já foi enfatizado, Edwards (1963) não aconselha a utilização deste teste para tabelas 2x2.

No caso de grandes amostras e considerando um esquema de duas variáveis Binomiais independentes, um recurso assintótico é o teste para comparação entre proporções descrito na seção V.2.

VI.2. Intervalos de Confiança para o Parâmetro θ

VI.2.1. Intervalos de Confiança Exatos Baseados no Modelo Condicional

Considerando que o modelo condicional $P(x/N,m,t;\theta)$ depende somente de θ, o parâmetro que desejamos estimar, Cornfield (1956) e independentemente Fisher (1962), apresentam limites de confiança exatos para θ utilizando o método padrão para famílias de distribuições uniparamétricas.

Para uma probabilidade bicaudal pré especificada α, seja θ_2 a solução para θ da equação:

$$\sum_{v=k}^{x} \frac{\binom{m}{v}\binom{n}{t-v}\theta^v}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\theta^u} = \frac{\alpha}{2} \qquad (25)$$

e θ_1 a solução para θ de:

$$\sum_{v=x}^{l} \frac{\binom{m}{v}\binom{n}{t-v}\theta^v}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\theta^u} = \frac{\alpha}{2} \qquad (26)$$

onde

$$k = \max(0, t-n)$$
$$l = \min(m, t)$$

Então:

$$Pr(\theta_1 <= \theta <= \theta_2) = 1 - \alpha$$

Este resultado está naturalmente relacionado com o Teste Exato de Fisher. O teste rejeitará a hipótese de nulidade para um dado conjunto de observações quando, e somente quan

do, os limites de confiança para o mesmo conjunto de observações não incluirem a unidade; ambos adotando o mesmo nível de confiança.

A dificuldade que se apresenta é que estas equações não podem ser resolvidas explicitamente pois envolvem razões de polinômios em θ e, portanto, o uso de técnicas iterativas se faz necessário. Thomas (1971) apresenta um programa computacional para obtenção dos limites de confiança exatos através de um método numérico. Existem atualmente, muitos "softwares" que podem ser utilizados na resolução de equações não lineares deste tipo, citamos por exemplo, o SAS, SOC (EMBRAPA)

VI.2.2. Intervalos de Confiança Assintóticos Baseados no Modelo Condicional

Como a obtenção de expressões exatas envolve dificuldades analíticas são propostas **aproximações assintóticas** para a distribuição condicional $P(x/N,m,t,\theta)$ que utilizam cálculos mais simples.

VI.2.2.1. Método proposto por Cornfield (1956)

Para evitar o problema de obtenção de expressões exatas para os momentos da distribuição Hipergeométrica Generalizada, o autor, se utiliza da distribuição assintótica da razão:

$$\frac{P(x)}{P(\tilde{x})}$$

onde $\tilde{x}$ é a **moda** da distribuição Hipergeométrica Generalizada e é definida pela desigualdade:

$$\frac{\tilde{x}\ (n-t+\tilde{x})}{(m-\tilde{x}+1)\ (t-\tilde{x}+1)} <= \Theta <= \frac{(\tilde{x}+1)\ (n-t+\tilde{x}+1)}{(m-\tilde{x})\ (t-\tilde{x})}$$

Para amostras grandes é suficiente escrever:

$$\frac{\tilde{x}\ (n-t+\tilde{x})}{(m-\tilde{x})\ (t-\tilde{x})} = \Theta \qquad (27)$$

É possível concluir que a distribuição limitante de $P(x/N,m,t,\Theta)$ é Normal com média $\tilde{x}$ e variância:

$$\frac{1}{[1/\tilde{x} + 1/(m-\tilde{x}) + 1/(t-\tilde{x}) + 1/(n-t+\tilde{x})]}$$

Deste modo, seja $\tilde{x}_2$ a **maior raiz real positiva** da seguinte equação quártica em $\tilde{x}$:

$$(\tilde{x}-x-1/2)^2\ [1/\tilde{x} + 1/(m-\tilde{x}) + 1/(t-\tilde{x}) + 1/(n-t+\tilde{x})] = X^2_\alpha$$

e, $\tilde{x}_1$ a **menor raiz real positiva** de:

$$(\tilde{x}-x+1/2)^2\ [1/\tilde{x} + 1/(m-\tilde{x}) + 1/(t-\tilde{x}) + 1/(n-t+\tilde{x})] = X^2_\alpha$$

onde X^2_α é o ponto α-percentual superior da distribuição Qui-Quadrado com 1 grau de liberdade;

x é a observação amostral da casela convencionalmente adotada.

Temos o seguinte limite assintótico:

$$Pr(\tilde{x}_1 <= \tilde{x} <= \tilde{x}_2) = 1 - \alpha$$

Desde que Θ é uma função monotônica de $\tilde{x}$, podemos derivar do resultado acima, o correspondente limite assintóti

co para o parâmetro razão de produtos cruzados:

$$Pr(\theta_1 < \theta < \theta_2) = 1 - \alpha$$

onde θ_1 e θ_2 são calculados através da substituição de $\tilde{x}_1$ e $\tilde{x}_2$ em (27), respectivamente. No Apêndice 4, chamamos a atenção para algumas restrições que devem ser estabelecidas para que a expressão (27) defina uma relação monotônica crescente entre θ e $\tilde{x}$.

VI.2.2.2. Método proposto por Cox (1958)

Neste caso, o autor aborda o problema de estimação por intervalo do parâmetro razão de produtos cruzados adotando um **modelo logístico** para a probabilidade de sucesso em uma seqüência mutuamente independente de variáveis aleatórias binárias.

Para nossa finalidade, vamos compor simplesmente os resultados que conduzem diretamente à obtenção das estimativas por intervalo assintóticas do parâmetro de interesse.

O modelo condicional $P(x/N,m,t,\theta)$ pode ser escrito em termos de $\lambda = \ln \theta$, o logarítmo neperiano do parâmetro razão de produtos cruzados. Assim:

$$P(x/N,m,t,\lambda) = \frac{\binom{m}{x}\binom{n}{y}e^{x\lambda}}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}e^{u\lambda}}$$

onde $n = N-m$ $\qquad$ $k = \max(0, t-n)$

$y = t-x$ $\qquad$ $l = \min(m, t)$

Deste modo, $P(x/N,m,t,\lambda=0)$ é da forma Hipergeométri

ca.

Vamos denotar o j-ésimo cumulante da distribuição condicional de X por $K^{(j)}(\lambda)$. A função geratriz dos cumulantes é dada por:

$$K(w;\lambda) = \log E\ (e^{wx})$$

No Apêndice 2, estabelecemos o seguinte resultado:

$$K(w;\lambda) = K(w+\lambda\ ;\ 0) - K(\lambda\ ;\ 0)$$

Isto significa, que os cumulantes da distribuição Hipergeométrica Generalizada, ou seja, do modelo $P(x/N,m,t,\lambda)$ para valores gerais de λ, são obtidos a partir da função geratriz dos cumulantes da distribuição Hipergeométrica, $P(x/N,m,t,\lambda=0)$, avaliada nos pontos $(w+\lambda)$ e λ.

Desta forma, uma expansão em série para os cumulantes da distribuição condicional de X é dada por:

$$K^{(1)}(w\ ;\ \lambda) = E\ (x/N,m,t,\lambda) = K_1 + \lambda\ K_2 + \ldots \qquad (28)$$

$$K^{(2)}(w\ ;\ \lambda) = V\ (x/N,m,t,\lambda) = K_2 + \lambda\ K_3 + \ldots \qquad (29)$$

onde K_j = é o j-ésimo cumulante da distribuição Hipergeométrica.

Considere a notação da Tabela 7 e suponha o seguinte experimento: de uma população de tamanho N, constituída de m indivíduos com a característica A e n indivíduos com a característica A', foi extraída aleatoriamente e sem reposição uma amostra de tamanho t. Seja x o número de indivíduos na amostra com a característica A. Esta situação é adequada para descrever o procedimento de fixar marginais aleatórias, t no

modelo Binomial ou (m, t) no modelo Multinomial. Ainda mais, caracteriza claramente o modelo Hipergeométrico para a distribuição de probabilidades de X. Assim:

$$K_1 = E\ (x/N,m,t,\lambda=0) = t\ (m/N)$$

$$K_2 = V\ (x/N,m,t,\lambda=0) = t\ (m/N)\ (n/N)\ [(N-t)/(N-1)]$$

Segue da normalidade assintótica da distribuição Hipergeométrica e dos resultados indicados em (28) e (29) que a distribuição da variável $(X/N,m,t,\lambda)$ converge para a Normal, respectivamente, com média e variância dadas por:

$$(K_1 + \lambda K_2) \quad e \quad K_2$$

Portanto, com um valor observado x, um limite de confiança a $(100 - \alpha)\%$ para a verdadeira média, com base na aproximação Normal e usando correção de continuidade, é obtido por:

$$(x \pm z_{\alpha/2}\ (K_2)^{1/2} \pm 1/2)$$

onde, o sinal positivo (+) está associado com o limite superior e o sinal negativo (-) com o limite inferior do intervalo. Desde que, a verdadeira média da distribuição de $(X/N,m,t,\lambda)$ é aproximadamente $(K_1 + \lambda K_2)$ segue o limite correspondente para λ e, consequentemente, para θ, a razão de produtos cruzados ($\theta = \exp \lambda$).

Enfatizamos, que Birch (1964) com base nestes resultados propõe a seguinte aproximação para o estimador do parâmetro λ, quando λ é pequeno (isto é, sob a hipótese H_0: $\lambda=0$):

$$\tilde{\lambda} = \frac{x - K_1}{K_2}$$

Desta forma, sob H_0,

$$\frac{x - K_1}{\sqrt{K_2}} \sim N(0;1)$$

ou, equivalentemente, sob H_0,

$$\frac{(x - K_1)^2}{K_2} \sim \chi^2_1$$

Assim, Birch (1964) estabelece o seguinte teste da hipótese $H_0: \lambda = 0$ x $H_1: \lambda > 0$. Usando correção de continuidade a hipótese H_0 é rejeitada em favor da hipótese H_1 se, e somente se:

$$(|x-K_1| - 1/2)^2 / K_2 >= \chi^2_\alpha$$

onde χ^2_α é o ponto α percentual superior da distribuição Qui-Quadrado com 1 grau de liberdade.

VI.2.3. Intervalos de Confiança Assintóticos Baseados em Modelos não Condicionais

Esta seção tem por finalidade apresentar duas outras propostas de obtenção de intervalos de confiança para o parâmetro θ sem, contudo, empregar o modelo reduzido $P(x/N,m,t,\theta)$. Em síntese, tais métodos se utilizam das propriedades assintóticas dos estimadores de máxima verossimilhança **não condicionais.**

Do nosso ponto de vista e, levando em consideração toda a argumentação desenvolvida nos capítulos anteriores, questionamos, não somente a vantagem real de aplicação destes procedimentos na análise de tabelas de contingência 2x2, mas,

principalmente, se é possível estabelecer pontos de comparação entre os limites obtidos a partir de métodos condicionais e não condicionais.

Realmente, é bastante frequente encontrarmos na literatura Estatística alternativas deste tipo, principalmente, quando o espaço paramétrico envolve muitos parâmetros e estamos interessados apenas em uma única função deles. Este é o caso:

- nos estudos prospectivos ou retrospectivos, (p_1, p_2) parâmetros do modelo Produto de Binomiais, definem o espaço paramétrico e desejamos realizar inferências apenas sobre $\theta = p_1(1-p_2)/p_2(1-p_1)$. Esta não é uma transformação biunívoca. Como está definido o parâmetro "nuisance"?

- ou ainda, nos estudos transversais, (p,q,r) parâmetros do modelo Multinomial, definem o espaço paramétrico e desejamos realizar inferências apenas sobre $\theta = ps/qr$. Esta também não é uma transformação biunívoca. Como estão definidos os parâmetros "nuisance"?

Neste sentido, ao se abordar o problema fazendo uso das propriedades assintóticas dos estimadores de máxima verossimilhança **não condicionais:**

$\hat{\theta} = \hat{p}_1(1-\hat{p}_2)/\hat{p}_2(1-\hat{p}_1)$ no caso do modelo Produto de Binomiais

ou $\hat{\theta} = \hat{p}\hat{s}/\hat{q}\hat{r}$ no caso Multinomial,

não se está controlando qualquer tipo de "relação" que possa estar ligando o parâmetro θ ao(s) parâmetro(s) "nuisance". Toda transformação introduzida no espaço paramétrico original deve ser analisada em detalhe, identificando quais são e como estão definidos os novos parâmetros.

VI.2.3.1. Método proposto por Gart (1962)

O autor deriva limites de confiança para o parâmetro θ considerando tamanhos amostrais grandes e pequenos.

Para **grandes amostras** o método se apresenta como uma modificação da estatística Qui-Quadrado comumente utilizada em tabelas 2x2:

$$\frac{[\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)]^2}{(\hat{p}_1\,\hat{q}_1)/m + (\hat{p}_2\,\hat{q}_2)/n} \sim \chi^2_1$$

onde χ^2_1 é a distribuição Qui-Quadrado com 1 grau de liberdade;

p_1 e p_2 são os parâmetros **Binomiais** correspondentes às duas populações em estudo;

$\hat{p}_1$ e $\hat{p}_2$ são as estimativas de máxima verossimilhança não condicionais dos parâmetros p_1 e p_2;

$q_1 = (1 - p_1)$ e $q_2 = (1 - p_2)$.

O autor propõe uma modificação desta expressão definindo a seguinte variável:

$$z^2 = (\hat{p}_1\,\hat{q}_2 - \hat{p}_2\,\hat{q}_1\,\theta)^2$$

Assim, não é difícil verificar que o valor esperado da variável z é nulo, isto é:

$$E(z) = 0 \quad \text{para todo } \theta$$

Ainda, $z / (\hat{V}(z))^{1/2}$ tem uma distribuição assintótica Normal Padrão, isto é:

$$\frac{z}{\hat{V}(z)^{1/2}} \sim N(0,1)$$

e portanto:

$$\frac{z^2}{\hat{V}(z)} \sim \chi^2_1$$

onde a variância assintótica de z é:

$$V(z) = \frac{p_2 q_2}{n}(p_1+\theta q_1)^2 + \frac{p_1 q_1}{m}(\theta p_2+q_2)^2 + \frac{p_1 q_1 p_2 q_2}{m n}(1-\theta)^2$$

Pode ser estabelecido o seguinte limite de confiança com correção de continuidade:

$$[z \pm 1/2\ (1/m + 1/n)]^2 = \chi^2_\alpha\ \hat{V}(z) \qquad (30)$$

onde, χ^2_α é o ponto α-percentual superior da distribuição Qui Quadrado com 1 grau de liberdade;

$\hat{V}(z)$ é a estimativa de $V(z)$, obtida através da substituição dos parâmetros p_1 e p_2 por suas estimativas.

Os sinais positivos e negativos estão associados com os limites superior e inferior do intervalo, respectivamente.

A equação (30) é quadrática em θ e suas duas raízes são os limites de um intervalo aproximado para o parâmetro razão de produtos cruzados com coeficiente de confiança igual a $(100 - \alpha)\%$.

Para tabelas em que:

$$\frac{x(t-x)}{m+n} > 1$$

as seguintes expressões aproximadas são derivadas para o limite inferior (θ_1) e superior (θ_2), respectivamente:

$$\theta_1 = \frac{\hat{\theta}\{1+(\frac{1}{m}+\frac{1}{n})(\chi_\alpha^2 - \frac{1}{2\hat{p}_1\hat{q}_2}) - \chi_\alpha[\frac{1}{(m-1)\hat{p}_1\hat{q}_1} + \frac{1}{(n-1)\hat{p}_2\hat{q}_2} - \frac{A}{mn} - B(\frac{1}{m}+\frac{1}{n})]^{1/2}\}}{1 - \chi_\alpha^2[\hat{p}_1/\hat{q}_1 m + \hat{q}_2/\hat{p}_2 n]}$$

$$\theta_2 = \frac{\hat{\theta}\{1+(\frac{1}{m}+\frac{1}{n})(\chi_\alpha^2 + \frac{1}{2\hat{p}_1\hat{q}_2}) + \chi_\alpha[\frac{1}{(m-1)\hat{p}_1\hat{q}_1} + \frac{1}{(n-1)\hat{p}_2\hat{q}_2} - \frac{A}{mn} + B(\frac{1}{m}+\frac{1}{n})]^{1/2}\}}{1 - \chi_\alpha^2(\hat{p}_1/\hat{q}_1 m + \hat{q}_2/\hat{p}_2 n)}$$

onde:

$$A = 1/\hat{\theta}[(1-\hat{\theta})^2 + \chi_\alpha^2(\hat{p}_1/\hat{q}_1 - \hat{q}_2/\hat{p}_2)^2]$$

$$B = (\hat{q}_1/n + \hat{p}_2/m) / (\hat{p}_1\,\hat{q}_1\,\hat{p}_2\,\hat{q}_2)$$

Quando:

$$\frac{x(t-x)}{m+n} >= 5$$

as seguintes fórmulas mais simples podem ser usadas:

$$\theta_1 = \hat{\theta}[1 \pm \chi_\alpha(\frac{1}{m\,\hat{p}_1\,\hat{q}_1} + \frac{1}{n\,\hat{p}_2\,\hat{q}_2})^{1/2}]$$

$$\theta_1 = \hat{\theta}[1 \pm \chi_\alpha(1/x + 1/(m-x) + 1/(t-x) + 1/(n-t+x))^{1/2}]$$

onde:

Θ_1 é o limite inferior do intervalo e corresponde ao sinal (-) da equação;

Θ_2 é o limite superior do intervalo e corresponde ao sinal (+) da equação.

No caso de **pequenas amostras** o autor modifica os limites de confiança exatos para o parâmetro Θ, expressões (25) e (26), propondo uma distribuição Hipergeométrica Generalizada com coeficientes Θ, isto é, dividindo o numerador e denominador de (25) por exemplo, por:

$$\binom{N}{t}$$

A seguir, o método está baseado em uma aproximação Binomial para a distribuição Hipergeométrica e, correspondentemente, a soma de Binomiais é expressa em termos da função Beta incompleta. Finalmente, é feito uso do relacionamento entre esta e a distribuição F.

É possível estabelecer o seguinte limite de confiança para Θ a $(100 - \alpha)\%$, que tem se apresentado como uma boa aproximação (Gart, 1962), para tabelas em que:

$$\frac{x\,(t-x)}{m+n} <= 1$$

$$\Theta_1 = \frac{(2n - H_1 + 1)\,x}{(2m + H_1 - t)(t - x + 1)} \quad \frac{1}{F_{1-\alpha/2}(2t - 2x + 2\ ;\ 2x)}$$

$$\Theta_2 = \frac{(2n - t + H_2)(x + 1)}{(2m - H_2 + 1)(t - x)}\, F_{1-\alpha/2}(2x + 2\ ;\ 2t - 2x)$$

onde:

$$H_1 = \frac{(t - x + 1)^2 + (t + 1)^2}{(2t - x + 2)}$$

$$H_2 = \frac{(x + 1)^2 + (t + 1)^2}{(x + t + 2)}$$

Observe, que o método de Gart para **grandes amostras** considera procedimentos **não condicionais**; é derivado de uma modificação na estatística Qui-Quadrado definida como função dos parâmetros Binomiais p_1 e p_2. Assim, tais resultados **somente** podem ser aplicados na análise de dados gerados dos estudos epidemiológicos **prospectivo e retrospectivo** que correspondem ao esquema amostral de duas populações Binomiais independentes.

Por outro lado, para o caso de **pequenas amostras** o autor deriva seus limites com base no modelo de distribuição **condicional** para X. Logo, os resultados obtidos **independem** do tipo de estudo epidemiológico.

VI.2.3.2. Método da Transformação Logito

Este método utiliza as propriedades assintóticas do estimador de máxima verossimilhança **não condicional** para o parâmetro razão de produtos cruzados.

Considere em um estudo prospectivo ou retrospectivo (modelo Produto de Binomiais) a seguinte variável indicada como a diferença entre logitos:

$$\hat{\beta} \quad \ln \hat{\Theta} \quad \ln \frac{\hat{p}_1}{(1 - \hat{p}_1)} - \ln \frac{\hat{p}_2}{(1 - \hat{p}_2)}$$

$$\ln \frac{x}{(m - x)} - \ln \frac{(t - x)}{(n - t + x)}$$

onde $\hat{p}_1$ e $\hat{p}_2$ são as estimativas de máxima verossimilhança **não condicionais** dos parâmetros p_1 e p_2.

Observe, que considerando um estudo transversal (modelo Multinomial) a mesma estimativa **não condicional** para o parâmetro β é obtida:

$$\hat{\beta} = \ln \hat{\Theta} = \ln \frac{\hat{p}}{\hat{r}} - \ln \frac{\hat{q}}{\hat{s}}$$

$$= \ln \frac{x/N}{(m-x)/N} - \ln \frac{(t-x)/N}{(n-t+x)/N}$$

A variância assintótica estimada de $\hat{\beta}$ é:

$$\hat{V}(\hat{\beta}) = \frac{1}{m \hat{p}_1 (1 - \hat{p}_1)} + \frac{1}{n \hat{p}_2 (1 - \hat{p}_2)}$$

$$= 1/x + 1/(m-x) + 1/(t-x) + 1/(n-t+x)$$

Novamente, observe que o mesmo resultado pode ser estabelecido considerando um estudo transversal; assintoticamente pode ser mostrado que a matriz de variância e covariância dos estimadores do logito dos parâmetros Multinomiais tende para uma matriz diagonal constituída pelas variâncias dos estimadores.

Assim, independente do modelo epidemiológico que gerou os dados; um limite de confiança aproximado para $\beta = \ln \Theta$

com coeficiente de confiança (100 - α)%, é dado por:

$$\ln \left[\frac{(x + 1/2)(n - t + x + 1/2)}{(t - x + 1/2)(m - x + 1/2)} \right] \pm z_{\alpha/2} (\hat{V}(\hat{\beta}))^{1/2}$$

onde $z_{\alpha/2}$ é o ponto α-percentual da distribuição N(0,1).

Tomando antilogaritmos, obtêm-se o limite correspondente para o parâmetro θ:

$$\hat{\theta} \exp \left[\pm z_{\alpha/2} (\hat{V}(\hat{\beta}))^{1/2} \right]$$

Estes limites sem correção de continuidade foram usados por Woolf (1954), citado por Gart (1962).

VI.3. Aplicações

VI.3.1. Exemplo 1: Um Estudo Retrospectivo

Os dados da Tabela 8 são citados e analisados por Cornfield (1956) è reanalisados por Cox (1958) e Gart (1962).

TABELA VI.1. Distribuição de pacientes com e sem câncer pulmonar de acordo com o hábito de fumar

RESPOSTA / GRUPO	F	F'	TOTAL
DOENTE	60	3	63
CONTROLE	32	11	43
TOTAL	92	14	106

O estudo considerou uma amostra aleatória de 63 indivíduos doentes, com câncer pulmonar, e uma amostra aleatória de 43 indivíduos controle, sem câncer pulmonar. Em cada grupo observou-se o número de indivíduos fumantes (F). Trata-se de um **estudo retrospectivo** em que os grupos de casos e controles são fixados e a resposta anotada é a presença ou ausência do possível fator de risco, o hábito de fumar.

A pesquisa tem por finalidade investigar se o hábito de fumar está associado com a ocorrência de câncer pulmonar.

Em estudos deste tipo, em que a prevalência da doença na população é baixa, a medida razão de produtos cruzados (θ) estima o risco relativo (RR):

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \cong RR = \frac{\text{risco da doença na população exposta}}{\text{risco da doença na população não exposta}}$$

onde

p_1 = probabilidade de um indivíduo do grupo com câncer pulmonar ser fumante;

p_2 = probabilidade de um indivíduo do grupo sem câncer pulmonar ser fumante.

Cornfield (1956), com o objetivo de mostrar a adequação numérica de seus limites assintóticos para o parâmetro razão de produtos cruzados quando aplicados a amostras pequenas, por conveniência, adota o **valor observado da variável X como 3**, portanto, condicionando a distribuição no **menor total marginal t=4**. Esta abordagem do problema apesar de introduzir uma modificação na definição do parâmetro θ, de acordo com nossa notação nos capítulos anteriores, conduz ao mesmo tipo

de análise dos dados.

Considerando a variável X como o número observado de indivíduos não fumantes (possível fator de risco **ausente**) no grupo com câncer pulmonar (grupo doente), a **distribuição condicional** de X dado que uma amostra de t indivíduos não fumantes é extraída da população geral (grupos doente e controle) é o modelo Hipergeométrico Generalizado parametrizado por:

$$\theta' = \frac{p_2/(1-p_2)}{p_1/(1-p_1)} = \frac{1}{\theta}$$

$$\cong \frac{\text{risco da doença na população não exposta}}{\text{risco da doença na população exposta}}$$

De acordo com este critério de definição do parâmetro θ', limites de confiança assintóticos a 95 por cento foram obtidos por Cornfield (1956), Cox (1958) e Gart (1962). Os resultados estão apresentados na Tabela VI.2.

TABELA VI.2. Intervalos de confiança assintóticos (LI e LS) para o parâmetro θ' considerando os dados da Tabela VI.1. Indicação dos níveis de confiança unicaudais exatos (α_1 e α_2), do comprimento dos intervalos (C) e dos coeficientes de confiança exatos.

MÉTODO	COEF. CONF. NOMINAL (%)	LI	α_1	LS	α_2	C	COEF. CONF. EXATO (%)
CORNFIELD	95	0.0296	0.0383	0.6229	0.0237	0.5933	93.80
COX	95	0.0450	0.0978	0.6130	0.0253	0.5680	87.69
GART	95	0.0285	0.0350	0.7651	0.0095	0.7360	95.54
LOGITO	95	0.0440	0.0933	0.6070	0.0263	0.5630	88.04

No Apêndice 4 estão indicados os cálculos para obtenção destes limites. O comprimento dos intervalos (C) é obtido simplesmente pela diferença entre os limites superior (LS) e inferior (LI). Os níveis de confiança unicaudais exatos (α_1 e α_2) são obtidos pelo cálculo do lado esquerdo das expressões (25) e (26) para as estimativas dos limites dos intervalos assintóticos em questão. Assim, se $\hat{\theta}_1$ e $\hat{\theta}_2$ representam os limites inferior e superior estimados para um certo intervalo, α_1 e α_2 são soluções das equações:

$$\sum_{v=k}^{x} \frac{\binom{m}{v}\binom{n}{t-v}\hat{\theta}_2^v}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\hat{\theta}_2^u} = \alpha_2$$

$$\sum_{v=x}^{l} \frac{\binom{m}{v}\binom{n}{t-v}\hat{\theta}_1^v}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\hat{\theta}_1^u} = \alpha_1$$

k e l definidos como anteriormente

onde $\alpha_1 + \alpha_2 = \alpha$

Logo, os coeficientes de confiança exatos são obtidos pelo cálculo $(1-\alpha)$.

Para estes dados e segundo a notação adotada:

$x = 3$ $\qquad$ $m = 63$

$t = 14$ $\qquad$ $n = 43$

onde,

$$\hat{\theta}' = \frac{x\,(n-t+x)}{(t-x)(m-x)} = 0.145$$

é a estimativa de máxima verossimilhança **não condicional** do parâmetro θ' e,

$$\tilde{\theta}' = \exp(\tilde{\lambda}') = \exp\left(\frac{x - K_1}{K_2}\right) = 0.165$$

é a estimativa **condicional** do parâmetro θ', sob a hipótese de θ' estar próximo da unidade.

Considerando os resultados apresentados na Tabela VI.2, o intervalo obtido por Cornfield, por exemplo, indica que com uma confiança de 95 por cento, o risco de câncer pulmonar para não fumantes é no máximo 62 por cento do risco para fumantes e, pode ser tão pequeno quanto 3 por cento. Esta é uma evidência de que o hábito de fumar é um fator que altera a ocorrência da doença na população, isto é, o risco de câncer pulmonar para fumantes é maior do que o risco para não fumantes. Ainda, o intervalo não inclue a unidade e, portanto, os grupos doente e controle não são homogêneos quanto ao hábito de fumar. O mesmo tipo de conclusão pode ser extraída a partir dos limites obtidos através dos outros métodos.

No artigo de Gart (1962) estes métodos são comparados de acordo com o comprimento dos intervalos, os coeficientes de confiança baseados nos limites exatos, bem como, na complexidade dos cálculos envolvidos. O autor discute, mediante a análise de várias tabelas, que o método de Gart é o mais **preciso**, isto é, apresentando coeficientes de confiança nominais bastante próximos dos valores calculados a partir dos limites exatos. Por outro lado, os métodos de Cox e Logito oferecem limites mais estreitos. O método de Cornfield apresentando limites também estreitos, é certamente adequado, porém, não atinge a precisão do método de Gart, além de envolver cálculos iterativos.

Apesar da colocação destes resultados, novamente chamamos a atenção para o fato de que os métodos de Cornfield e Cox são baseados no modelo condicional $P(x/N,m,t,\theta)$ ao passo que os métodos de Gart, para grandes amostras, e Logito consideram técnicas não condicionais. Fundamentalmente, do ponto de vista teórico, qualquer critério de comparação que seja definido não pode ser admitido sem levar em conta tais características metodológicas. Contudo, do ponto de vista prático, é preciso comparar e analisar criticamente os resultados obtidos através de todos os métodos propostos para, finalmente, escolher qual deve ser utilizado.

VI.3.2. Exemplo 2: Um Estudo Prospectivo

Os dados da Tabela VI.3 se referem a um estudo longitudinal realizado em Framingham para investigar o efeito de um grande número de variáveis como fatores de risco no desenvolvimento de doenças coronarianas (Truett, Cornfield e Kannel, 1967).

A pesquisa considerou o seguimento por doze anos de 2187 homens e 2669 mulheres com idades entre 30 e 62 anos, no início, livres de doenças coronarianas, observando-se assim, o desenvolvimento da doença. Passaremos a analisar parte destes dados.

TABELA VI.3. Classificação de 2187 homens e 2669 mulheres de acordo com a taxa de incidência de doenças coronarianas durante o período de 12 anos

GRUPO \ RESPOSTA	D	D'	TOTAL
F	129	2540	2669
M	258	1929	2187
TOTAL	387	4469	4856

Trata-se de um **estudo prospectivo** que tem por finalidade investigar se o desenvolvimento de doenças coronarianas é diferente no grupo feminino em relação ao grupo masculino. Ressaltamos, que devido à metodologia em consideração, não levamos em conta a estratificação dos dados segundo classes de idade.

Com os dados da Tabela VI.3, por conveniência, considere a seguinte notação:

$x = 129$ $\qquad$ $m = 2669$

$t = 387$ $\qquad$ $n = 2187$

onde,

$$\Theta = \frac{\text{risco de doenças coronarianas em mulheres}}{\text{risco de doenças coronarianas em homens}}$$

e, assim,

$$\hat{\Theta} = \frac{129 * 2540}{258 * 1929} = 0.3727$$

é a estimativa de máxima verossimilhança **não condicional** do parâmetro θ e,

$$\hat{\theta} = \exp\left(\frac{129 - 212.7065}{88.18}\right) = 0.3870$$

é a estimativa **condicional** do parâmetro θ, sob a hipótese de θ estar próximo da unidade.

Na tabela a seguir estão indicados os limites de confiança assintóticos a 95 por cento de acordo com os métodos de Cornfield (1956), Cox (1958), Gart (1962) e Logito. Não é apresentado o mesmo formato da Tabela VI.2 devido aos totais marginais para estes dados serem grandes e levarem a problemas de "over-flow" na execução do programa para o cálculo dos coeficientes de confiança exatos.

TABELA VI.4. Intervalos de confiança assintóticos (LI e LS) a 95 por cento para o parâmetro θ considerando os dados da Tabela VI.3.

MÉTODO	LI	LS
CORNFIELD	0.3030	0.4756
COX	0.3123	0.4796
GART	0.2963	0.4630
LOGITO	0.3055	0.4738

Com estes resultados, observamos que os métodos condicionais de Cornfield e Cox fornecem limites bastante próximos. O mesmo ocorrendo com os métodos não condicionais de Gart e Logito.

Analisando, por exemplo, os limites segundo Cornfield para estes dados, verificamos que o risco de doenças coronarianas para mulheres é no máximo 47 por cento do risco de doenças coronarianas para homens e, pode ser tão pequeno quanto 30 por cento. Isto indica que a incidência de doenças coronarianas é maior nos homens. Realmente, para os dados deste estudo prospectivo, não caracterizando a estratificação pela idade, a estimativa da taxa de incidência de doenças coronarianas nas mulheres é menor do que nos homens:

$\hat{p}_1 = x / m = 129 / 2669 = 0.0483$

$\hat{p}_2 = (t-x) / n = 258 / 2187 = 0.1179$

onde

$\hat{p}_1$ é a estimativa de máxima verossimilhança não condicional da proporção de doenças coronarianas nas mulheres;

$\hat{p}_2$ é a estimativa não condicional da proporção de doenças coronarianas nos homens.

VI.3.3. Exemplo 3: Um Estudo Transversal

Os dados da Tabela 12 correspondem à classificação de 20878 nascimentos ocorridos em 31 maternidades do município de São Paulo em 1978, de acordo com o hábito de fumar da mãe (presença ou ausência) e peso do recém nascido (baixo peso ou não).

TABELA VI.5. Classificação de 20878 recém nascidos de acordo com o hábito de fumar da mãe e o peso do recém nascido

PESO(gr) / HÁBITO FUMAR	<= 2500	> 2500	TOTAL
F	752	6756	7508
F'	679	12691	13370
TOTAL	1431	19447	20878

Considerando o tipo de delineamento utilizado na coleta dos dados, observamos que trata-se de um **estudo transversal** em que o único total marginal fixo é o número 20878 nascimentos observados. Desta forma, cada unidade experimental foi classificada segundo o hábito de fumar da mãe e o peso do recém nascido.

A pesquisa tem por finalidade investigar se as variáveis materna e do recém nascido em estudo estão associadas ou ocorrem independentemente.

Para estes dados considere a seguinte notação:

$x = 752$ $\qquad$ $m = 7508$

$t = 1431$ $\qquad$ $n = 13370$

Adotando como medida de associação o parâmetro razão de produtos cruzados (θ), onde:

$$\theta = \frac{\text{risco de RN de baixo peso para mães fumantes}}{\text{risco de RN de baixo peso para mães não fumantes}}$$

Temos:

$$\hat{\theta} \quad \frac{752 * 12691}{679 * 6756} = 2.08$$

é a estimativa de máxima verossimilhança **não condicional** de θ

$$\tilde{\theta} = \exp\left(\frac{752 - 514.606}{306.9743}\right) = 2.16$$

é a estimativa **condicional** de θ, sob a hipótese de θ estar próximo da unidade.

Na tabela a seguir estão indicados os limites de confiança assintóticos a 95 por cento, de acordo com os métodos de Cornfield (1956), Cox (1958) e Logito. Neste caso, o método de Gart (1962) para grandes amostras não é apropriado. Também, devido aos totais marginais dos dados serem grandes não foram possíveis os cálculos dos coeficientes de confiança exatos.

TABELA VI.6. Intervalo de confiança assintóticos (LI e LS) a 95 por cento para o parâmetro θ, considerando os dados da Tabela VI.3.

MÉTODO	LI	LS
CORNFIELD	1.8649	2.3208
COX	1.9344	2.4274
LOGITO	1.8674	2.3171

Observamos que os métodos condicionais de Cornfield e Cox, bem como o método não condicional Logito, fornecem limites de confiança bastante próximos. Na prática, este resul

tado é uma indicação de que todos os métodos em questão fornecem o mesmo intervalo para o parâmetro θ.

Os intervalos não incluem a unidade e, portanto, concluímos que existe associação entre o hábito de fumar da mãe e o peso do recém nascido. Analisando, por exemplo, os limites segundo Cornfield, verificamos que o risco de recém nascidos de baixo peso para mães fumantes é no mínimo 1.86 e no máximo 2.32 vezes o risco de recém nascidos de baixo peso para mães não fumantes. Assim, os dados deste estudo transversal indicam que a presença de hábito de fumar em mães é um fator que concorre efetivamente para a ocorrência de recém nascidos de baixo peso.

APÊNDICE 1

Temos que mostrar que as condições da definição de G-ancilaridade segundo Godambe (1980), são satisfeitas pelas estatísticas marginais (m,t) derivadas do modelo Multinomial com respeito ao parâmetro Θ, razão de produtos cruzados.

Verificamos a seguinte fatoração (expressão 23):

$$P(x,m,t/N,\Theta,\psi,\Phi) = P(m,t/N,\Theta,\psi,\Phi)\ P(x/N,m,t,\Theta)$$

de onde decorre que a estatística x é **especificamente suficiente** para os parâmetros (ψ,Φ). Deste modo, para que o conceito de G-ancilaridade seja válido para as estatísticas (m,t) em relação a Θ, resta saber se para Θ pertencente a $\oplus$ fixado, a família de distribuições marginais $\{P(m,t/N,\Theta,\psi,\Phi)\ ;\ 0 < \Phi,\psi < 1\}$ é **completa**, onde:

$$P(m,t/N,\Theta,\psi,\Phi) = \frac{\psi^t\ (1-\psi)^{(N-t)}}{(1\ -\ \psi\ +\ \Theta\psi)^m}\ \ \Phi^m\ (1-\Phi)^{(N-m)}\ \binom{N}{m}\ \sum_{u=k}^{l}\ \binom{m}{u}\ \binom{n}{t-u}\ \Theta^u$$

$$k = \max\ (0, t-n)$$
$$l = \min\ (m, t)$$

É necessário mostrar que:

$$E\ [f(m,t)] = 0 \quad \Longrightarrow \quad f(m,t) = 0 \quad \text{para todo } (m,t)$$

Assim:

$$E[f(m,t)] = \sum_{m=0}^{N}\ \sum_{t=0}^{N}\ f(m,t)\ \frac{\psi^t(1-\psi)^{(N-t)}}{(1-\psi+\Theta\psi)^m}\ \Phi^m\ (1-\Phi)^{(N-m)}\ \binom{N}{m}\ \sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\ \Theta^u =$$

$$\sum_m \left\{ \frac{\phi^m (1-\phi)^{(N-m)} \binom{N}{m}}{(1-\psi+\theta\psi)^m} \left[\sum_t f(m,t)\, \psi^t (1-\psi)^{(N-t)} \sum_u \binom{m}{u} \binom{n}{t-u} \theta^u \right] \right\} = 0$$

Para θ e ψ fixados, a equação anterior pode ser vista como a esperança da função:

$$h(m) = \frac{1}{(1-\psi+\theta\psi)^m} \left[\sum_t f(m,t)\, \psi^t (1-\psi)^{(N-t)} \sum_u \binom{m}{u} \binom{n}{t-u} \theta^u \right]$$

isto é:

$$E[f(m,t)] = \sum_m h(m) \underbrace{\binom{N}{m} \phi^m (1-\phi)^{(N-m)}}_{\text{Bin}(N;\phi)} = E[h(m)] = 0$$

Pela propriedade de **completividade** da distribuição Binomial temos que:

$$E[h(m)] = 0 \Longrightarrow h(m) = 0 \quad , \quad m = 0,\ldots,N$$

Seja então, m fixado e arbitrário. Temos que:

$$\frac{1}{(1-\psi+\theta\psi)^m} \left[\sum_{t=0}^{N} f(m,t)\, \psi^t (1-\psi)^{(N-t)} \sum_{u=k}^{l} \binom{m}{u} \binom{n}{t-u} \theta^u \right] = 0$$

Como $[1 + \psi(\theta-1)] > 0$, podemos analisar apenas o somatório. Portanto:

$$\sum_{t=0}^{N} f(m,t)\, \psi^t (1-\psi)^{(N-t)} \sum_{u=k}^{l} \binom{m}{u} \binom{n}{t-u} \theta^u = 0$$

Novamente pela completividade da distribuição Binomial:

$$f(m,t) \quad \sum_{u=k}^{l} \binom{m}{u} \binom{n}{t-u} \theta^u = 0$$

Como:

$$\sum_{u=k}^{l} \binom{m}{u} \binom{n}{t-u} \cdot \theta^u > 0$$

temos que:

$f(m,t) = 0$ para $m = 0,\ldots,N \quad 0 < t < m$

Logo, a família $\{P(m,t/N,\theta,\psi,\Phi) ; 0 < \psi,\Phi < 1\}$ é completa e, portanto, a estatística (m,t) é G-ancilar com respeito a θ.

APÊNDICE 2

Podemos reescrever a distribuição condicional de X da seguinte forma:

$$P(x/N,m,t,\theta) = \frac{\binom{m}{x}\binom{n}{t-x}\theta^x}{\sum_{u=k}^{l}\binom{m}{u}\binom{n}{t-u}\theta^u}$$

$$= \frac{e^{x\lambda} / x!\,(m-x)!\,(t-x)!\,(n-t+x)!}{\sum_{u=k}^{l} e^{u\lambda} / u!\,(m-u)!\,(t-u)!\,(n-t+u)!}$$

$$= \frac{e^{x\lambda} / x!\,(m-x)!\,(t-x)!\,(n-t+x)!}{f(\lambda)}$$

onde:

$k = \max(0, t-n)$ $\qquad$ $\lambda = \ln \theta$

$l = \min(m, t)$ $\qquad$ $n = N-m$

$$f(\lambda) = \sum_{u=k}^{l} e^{u\lambda} / u!\,(m-u)!\,(t-u)!\,(n-t+u)!$$

Vamos denotar o j-ésimo **cumulante** de distribuição condicional de X por $K^{(j)}(\lambda)$. A **função geratriz dos cumulantes** é dada por:

$$K(w\ ;\ \lambda) = \log E\,(e^{wx})$$

Podemos expressar a função $K(w\ ;\ \lambda)$ em termos de $f(.)$:

$$K(w\ ;\ \lambda) = \log E(e^{wx}) = \log \sum_x e^{wx}\ P(x/N,m,t,\lambda)$$

$$\log 1/f(\lambda) \sum_x e^{x(w+\lambda)} / x!\ (m-x)!\ (t-x)!\ (n-t+x)!$$

$$= \log [f(w+\lambda) / f(\lambda)] = \log f(w+\lambda) - \log f(\lambda)$$

Por outro lado, a função geratriz de momentos da variável $(X/N,m,t,\lambda=0)$ que tem distribuição Hipergeométrica, é obtida por:

$$M(w;0) = E\ (e^{wx}) = \sum_x e^{wx}\ P(x/N,m,T,\lambda=0)$$

$$= \sum_x e^{wx}\ \frac{1\ /\ x!\ (m-x)!\ (t-x)!\ (n-t+x)!}{f(0)}$$

$$= 1/f(0) \sum_x e^{wx} / x!\ (m-x)!\ (t-x)!\ (n-t+x)!$$

$$= f(w) / f(0)$$

Portanto,

$$K(w;0) = \log M(w;0) = \log [\ f(w) / f(0)\ ]$$

Do mesmo modo, pode ser encontrado que a função geratriz de momentos da variável $(X/N,m,t,\lambda)$ para valores gerais de λ, é da forma:

$$M(w;\lambda) = M(\lambda+w;0) / M(\lambda;0)$$

$$\frac{f(\lambda+w) \;/\; f(0)}{f(\lambda) \;/\; f(0)} \quad f(\lambda+w) \;/\; f(\lambda)$$

Assim, podemos estabelecer o seguinte resultado:

$$K(w;\lambda) = \log M(w;\lambda) = \log [f(\lambda+w) \;/\; f(\lambda)]$$

$$= \log [f(\lambda+w) \;/\; f(0)] \;-\; \log [f(\lambda) \;/\; f(0)]$$

$$= \log M(w+\lambda;0) \;-\; \log M(\lambda;0)$$

Portanto, a função geratriz dos cumulantes da distribuição Hipergeométrica Generalizada é dada por:

$$K(w;\lambda) = K(w+\lambda;0) \;-\; K(\lambda;0)$$

APÊNDICE 3

Seja $X = (X_1, X_2, \ldots, X_k)$ uma variável aleatória Multinomial com vetor de probabilidades $p = (p_1, p_2, \ldots, p_k)$. A função de probabilidades de X é dada por:

$$P(x/N,p) = N! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}$$

onde, os x_i's de $x = (x_1, x_2, \ldots, x_k)$ tomam os valores 0,1, ...,N, sujeitos à restrição

$$\sum_{i=1}^{k} x_i = N$$

supondo o valor de N conhecido. Os valores dos p_i's pertencem ao intervalo [0,1] sujeitos à restrição

$$\sum_{i=1}^{k} p_i = 1.$$

Temos o seguinte resultado:

$$E(X_i) = Np_i \qquad i = 1,2,\ldots,k$$

$$V(X_i) = Np_i(1-p_i) \qquad i = 1,2,\ldots,k$$

$$Cov\,(X_i,X_j) = -Np_ip_j \qquad i \neq j \;;\; i,j = 1,2,\ldots,k$$

Vamos aplicar a seguinte transformação na variável aleatória X:

$$X \xrightarrow{T} \ln X$$

A expansão em série de Taylor ao redor do ponto $E(X_i)$ com aproximação até primeira ordem é dada por:

$$\ln X_i \cdot \ln Np_i + [(X_i - Np_i) / Np_i]$$

Logo,

$$E (\ln X_i) \cong \ln Np_i$$

$$V (\ln X_i) \cong E [\ln X_i - \ln Np_i]^2$$

$$\cong 1 / (Np_i)^2 \; V(X_i) = 1 / Np_i - 1 / N$$

$$Cov (\ln X_i ; \ln X_j) \cong E [(\ln X_i - \ln Np_i) (\ln X_j - \ln Np_j)]$$

$$\cong E \left[\frac{X_i - Np_i}{Np_i} \quad \frac{X_j - Np_j}{Np_j} \right] = -1 / N$$

Este resultado estabelece que para N suficientemente grande ($N \longrightarrow \infty$), a matriz de variância e covariância dos estimadores multinomiais tende para uma matriz diagonal constituída pelas variâncias dos estimadores.

APÊNDICE 4

Para os dados da Tabela 8 e utilizando a notação adotada temos:

$x = 3$ $\qquad$ $m = 63$

$t = 14$ $\qquad$ $n = 43$

Podemos derivar os seguintes resultados:

- MÉTODO DE CORNFIELD (1956)

De acordo com o método descrito na seção VI.2.2.1, precisamos obter as soluções das seguintes equações quárticas em $\tilde{x}_i$:

$$(\tilde{x}_i - 3 \pm 1/2)^2 [1/\tilde{x}_i + 1/(63-\tilde{x}_i) + 1/(14-\tilde{x}_i) + 1/(29+\tilde{x}_i)] = 3.841$$

onde,

$\tilde{x}_1$ corresponde à **menor raiz real positiva** para o sinal (+) da equação;

$\tilde{x}_2$ corresponde à **maior raiz real positiva** para o sinal (-) da equação.

$X^2_{1;0.05} = 3.841$

Na Tabela 14 estão apresentadas as quatro raízes de cada equação, obtidas pelo método iterativo de Newton Raphson. Para tanto, utilizamos o "software" SOC (EMBRAPA).

TABELA 14. Indicação das soluções aproximadas das equações quárticas em $\tilde{x}_1$ e $\tilde{x}_2$.

EQUAÇÃO EM $\tilde{x}_1$		EQUAÇÃO EM $\tilde{x}_2$	
RAIZ	VALOR DA FUNÇÃO	RAIZ	VALOR DA FUNÇÃO
-11.2333	$-4.17\ 10^{-5}$	-11.009	$-4.1\ 10^{-6}$
0.8148	$2.1\ 10^{-6}$	1.3738	$2.99\ 10^{-2}$
5.8628	$-2.4\ 10^{-6}$	6.9042	$-8.2\ 10^{-6}$
27.1188	$0.3\ 10^{-6}$	27.2193	$7.2\ 10^{-6}$

A solução de $\tilde{x}_1$ que procuramos corresponde ao valor 0.815, isto é, a menor raiz real positiva. No entanto, para obtermos a solução de $\tilde{x}_2$ não basta encontrarmos a maior raiz real positiva, esta deve estar **restrita** a valores de $\tilde{x}$ que sejam **menores que o mínimo entre (m,t)**. Esta condição deve estar satisfeita para que, de acordo com a expressão (27), o parâmetro $\theta(\theta')$ seja uma **função monotônica crescente** de $\tilde{x}$.

Assim,

$$\tilde{x}_2 = 6.904 \quad < \quad \min\ (63;14)$$

é a solução que procuramos.

Temos o seguinte resultado:

IC ($\tilde{x}$) a 95% = (0.815 ; 6.9904)

e, portanto, substituindo estes limites em (27):

IC (θ') a 95% = (0.0296 ; 0.6229)

- **MÉTODO DE COX (1958)**

Considerando os resultados apresentados na seção VI.2.2.2, para um valor observado X=3, um limite de confiança para a verdadeira média da distribuição condicional de X baseada na aproximação Normal, com correção de continuidade é dada por:

$$(x \pm z \sqrt{K_2} \pm 1/2) = (3 \pm (1.96*1.72) \pm 1/2)$$

$$= (-0.871 ; 6.871)$$

onde:

$$z_{0.025} = 1.96$$

$$K_2 = t\ m/N\ n/N\ [(N-t) / (N-1)]$$

$$= 14\ 63/106\ 43/106\ [\ (106-14) / (106-1)\ ]$$

$$= 2.9574$$

e, desde que a verdadeira média da distribuição condicional de X é aproximadamente $(K_1 + \lambda K_2)$, segue o correspondente limite para o parâmetro λ:

$$K_1 + \lambda_1 K_2 = -0.871 \longrightarrow \lambda_1 = -3.109$$

$$K_1 + \lambda_2 K_2 = 6.871 \longrightarrow \lambda_2 = -0.490$$

onde:

$$K_1 = t\ m/N = 14\ 63/106 = 8.321$$

e, como $\Theta' = \exp(\lambda)$:

IC (θ') a 95% = (0.045 ; 0.613)

- MÉTODO DE GART (1962)

Considerando os resultados descritos na seção VI.2.3.1 temos o seguinte critério:

$$\frac{x\ (t - x)}{m + n} = \frac{3\ (14-3)}{63 + 43} = 0.3113 < 1$$

e, portanto, temos que calcular:

$$H_1 = \frac{(t-x+1)^2 + (t+1)^2}{2t - x + 2} = \frac{369}{27} = 13.6667$$

$$H_2 = \frac{(x+1)^2 + (t+1)^2}{x + t + 2} = \frac{241}{19} = 12.6842$$

$$\theta'_1 = \frac{(2n - H_1 + 1)\ x}{(2m+H_1-t)\ (t-x+1)} \quad \frac{1}{F_{0.975}(2t-2x+2;2x)}$$

$$= \frac{219.9999}{1508.0004} \quad \frac{1}{5.1172} = 0.0285$$

$$\theta'_2 = \frac{(2n-t+H_2)\ (x+1)}{(2m-H_2+1)\ (t-x)} \quad F_{0.975}(2x+2;2t-2x)$$

$$= \frac{338.7368}{1257.4738} \quad 2.8402 = 0.7651$$

Logo,

IC (θ') a 95% = (0.0285 ; 0.7651)

- MÉTODO DA TRANSFORMAÇÃO LOGITO

Considerando os resultados descritos na seção VI.2.3.2 um limite de confiança a 95 por cento para o parâmetro β é dado por:

$$\ln \left[\frac{(x+1/2)\ (n-t+x+1/2)}{(t-x+1/2)\ (m-x+1/2)} \right] \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{B})}$$

onde:

$$z_{\alpha/2} = z_{0.025} = 1.96$$

$$\hat{V}(\hat{B}) = 1/x + 1/(m-x) + 1/(t-x) + 1/(n-t+x)$$

$$= 0.4721$$

Logo,

IC (β) a 95% = In 0.1634 $\pm$ 1.3467

= (-3.1576 ; -0.4642)

e, portanto, como β = In θ', o correspondente intervalo para o parâmetro θ' é obtido:

IC (θ') a 95% = (0.0425 ; 0.6285)

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSEN, E.B. (1967). On Parcial Sufficiency and Partial Ancillarity. Aktuartidskr: 137-152.

ANDERSEN, E.B. (1970). Asymptotic Properties of Conditional Maximum-likelihood Estimators. J.R.Statist.Soc., Series B, **32**: 283-301.

BARNDORFF-NIELSEN, O. (1978). **Information and Exponential Families in Statistical Theory.** John Wiley, New York.

BASU, D. (1975). Statistical Information and Likelihood (with Discussions). Sankyã, Series A, **37**: 1-71.

BASU, D. (1977). On the Elimination of Nuisance Paramaters. J.Am.Statist.Assoc., **72**: 355-366.

BASU, D. (1979). Discussion of Berkson's Paper "In Dispraise of the exact test". J.Statist.Plan.Inf., **3** : 189-197.

BHAPKAR, V.P.; KOCH, G.G. (1968). Hypotheses of "No Interaction" in Multidimensional Contingency Tables. Technometrics, **10**(1): 107-123.

BIRCH, M.W. (1963). Maximum Likelihood inThree – Way Contingency Tables. J.R.Statist.Soc., Series B, **25**: 220-233.

BIRCH, M.W. (1964). The Detection of Partial Association, I: The 2x2 Case. J.R.Statist.Soc., Series B, **26**: 313-324.

BIRCH, M.W. (1965). The Detection of Partial Association, II: The General Case. J.R.Statist.Soc., Series B, **27**: 111-124.

CORNFIELD, J. (1956). A Statistical Problem Arising from Retrospective Studies. Proc.Third Berkeley Symp., **4**: 135-148.

CORNFIELD, J.; HAENSZEL, W. (1960). Some Aspects of Retrospective Studies. J.Chron.Dis., **11**: 523.

COX, D.R. (1958). The Regression Analysis of Binary Sequences. J.R.Statist.Soc., Series B, **20**: 215-242.

COX, D.R. (1970). **The Analysis of Binary Data.** London, Methuen.

COX, D.R. (1975). Partial Likelihood. Biometrika, **62** (2): 269-276.

DARROCH, J.N. (1962). Interactions in Multi-Factor Contingency Tables. J.R.Statist.Soc., Series B, **24**: 251-263.

EDWARDS, A.W.F. (1963). The Measure of Association in a 2x2 Table. J.R.Statist.Soc., Series A, **126**: 109-114.

FISHER, R.A. (1935). The Logic of Inductive Inference. J.R. Statist.Soc., Series A, **98** (1): 39-54.

FISHER, R.A. (1962). Confidence Limits for a Cross-Product Ratio. Austr.J.Statist., **4**: 41.

FRASER, D.A.S. (1956). Sufficient Statistics with Nuisance Parameters. Ann.Math.Statist., **27**: 838-842.

GART, J.J. (1962). On the Combination of Relative Risks. Biometrics, **18**: 601-610.

GART, J.J. (1971). The Comparison of Proportions: A Review of Significance Tests, Confidence Intervals and Adjustments for Stratification. Nat.Cancer Inst., **39**(2): 148-169.

GODAMBE, V.R. (1980). On Sufficient and Ancilearity in Presence of a Nuisance Parameters. Biometrika, **67**: 155-162.

GOODMAN, L.A. (1964). Simple Methods for Analysing Three-Factor Interactions in Contingency Tables. J.Am.Statist.Assoc., **59**(306): 210-353.

GOODMAN, L.A. (1969). On Partitioning X^2 and Detecting Partial Association in Three-Way Contingency Tables. J.R. Statist. Soc., Series B, **31**: 485-498.

GOODMAN, L.A.; KRUSKAL, W.H. (1954). Measures of Association for Cross Classifications. J.Am.Statist.Assoc., **49**: 734-764.

GOODMAN, L.A.; KRUSKAL, W.H. (1959). Measures of Association for Cross Classifications. II: Further Discussion and References. J.Am.Statist.Assoc., **54**: 123-163.

GUENTHER, W.G. (1968). **Concepts of Probability.** McGraw Hill Book Company.

HANNAN, J.; HARKNESS, W.L. (1963). Normal Aproximation to the Distribution of two Independent Binomials, Conditional on Fixed Sum.Ann.Math.Statist., **34**: 1593-1595.

HARKNESS, W.L. (1965). Properties of the Extended Hipergeometric Distribution. Ann.Math.Statist., **36**: 938-945.

IRONY, T.Z. (1984). **Testes Exatos para Tabelas 2x2: Bayes x Fisher.** São Paulo, 133p. Dissertação (Mestrado) - IME/USP.

KALBFLEISCH, J.D.; SPROTT, D.A. (1973). Marginal and Conditional Likelihoods. Sankhyã, Series A, **35**: 311-328.

KLEINBAUM, D.G.; KUPPER, L.L.; MORGENSTERN, H. (1982). **Epidemiologic Research. Principlels and Quantitative Methods.** Lifetime Learning Publications. Belmont, California.

KSHIRSAGAR, A.M. (1972). **Multivariate Analysis.** New York, Marcel Dekker, INC., vol. 2.

LEHMANN, E.L. (1959). **Testing Statistical Hypotheses.** New York, John Wiley & Sons.

MACMAHON, B.; PUGH, R.F. (1970). **Epidemiology. Principles and Methods.** Boston, Little, Brown and Company.

MANTEL, N.; HANKEY, B.F. (1971). Programed Analysis of a 2x2 Contingency Table. Am. Statist., 25(5): 40-44.

MANTEL, N.; HAENSZEL, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. J. Nat. Cancer Inst., 22: 719-748.

MAUSNER, J.S.; KRAMER, S. (1985). **Epidemiology - An Introductory Text.** W.B.Saundess Company.

PLACKETT, R.L. (1962). A Note on Interactions in Contingency Tables. J.R.Statist.Soc., Series B, **24**: 162-166.

SAS Institute Inc. (1985). Reference Manual.

SOC (1987). Manual do Usuário, módulo CM. NTIA, EMBRAPA.

SPROTT, D.A. (1975). Marginal and Conditional Sufficiency. Biometrika, **62**(3): 599-605.

SVERDRUP, E. (1966). The Present State of the Decision Theory and the Neyman - Pearson Theory. Rev.Int.Statist.Inst., **34**: 309-333.

THOMAS, D.G. (1971). Exact Confidence Limits for the Odds Ratio in a 2x2 Table. Appl.Statist., **20**: 105-110.

TRUETT, J.; CORNFIELD, J.; KANNEL, W. (1967). A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. J.Chron.Dis., **20**: 511-524.

WOLF, B. (1954). On Estimating the Relation Betwen Blood Group and Disease. Ann.Hum.Genetics, **19**: 251-253.