APLICAÇÃO DO MODELO LOG-LINEAR NA ESTIMAÇÃO DO RISCO RELATIVO NOS ESTUDOS CASO-CONTROLE

TAKUMI IGUCHI

DISSERTAÇÃO APRESENTADA

AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR:

PROF. DR. DAVID MARTINS DORIGO

- SÃO PAULO, ABRIL DE 1979 -

Quero dedicar este trabalho a meus pais,
Yoshimi e Namie,
pela formação que me propiciaram, a
minha esposa Adélia Kazuko
e ao meu filho Fábio,
por sua dedicação e compreensão pelas
horas roubadas de seu convívio durante
as nossas atividades ...

AGRADECIMENTOS

Gostaríamos de agradecer a todos que, direta ou indiretamente, contribuiram decisivamente, quer incentivando, quer orientando, para a realização deste trabalho. Em especial, desejamos destacar:

– O Professor Doutor David Martins Dorigo – IM-UFRJ – que sugeriu o tema desta dissertação e foi, além de orientador efetivo, um amigo incetivador e paciente em todas as situações, transmitindo com dedicação seus conhecimentos e experiência;

– O Professor Doutor Euclides Ayres de Castilho – FM-USP – por seus valiosos comentários e sugestões, orientando-nos, especialmente, nos enfoques epidemiológicos;

– O Professor Doutor Adolpho Walter Pimazoni Canton – IME-USP – pela leitura paciente e pelas sugestões;

– Os membros do Departamento de Estatística – IME-USP – particularmente, o Professor Doutor Carlos Alberto Barbosa Dantas que possibilitou nossos primeiros contactos com esse Departamento e orientou nossos primeiros passos na área e o Professor Doutor Flávio Wagner Rodrigues que nos orientou na fase de obtenção de créditos, quando fomos bolsistas da CAPES;

– O Professor Doutor Djalma Galvão Carneiro Pessoa – IMPA – que nos orientou no primeiro ano como bolsita da CAPES;

– O Professor Doutor Clóvis de Araújo Peres e a Professora Maria Takishita – IME-USP – pela orientação e incentivo;

– Os colegas do Departamento de Epidemiologia e Métodos Quantitativos em Saúde – ENSP-FIOCRUZ, nas pessoas dos Professores Doutores Joir

Gonçalves da Fonte e Eduardo de Azeredo Costa;

— Os Professores Marina Wagner Osanai e Carlos Hiroyuki Osanái — ENSP - FIOCRUZ por suas pacientes leituras e críticas dos textos;

— A Senhora Lisabel Espellet Klein — PEPPE - FIOCRUZ — pela incansável dedicação na busca e obtenção de material bibliográfico utilizado;

— A Senhora Maria Eugenia Rocha Valente por seu paciente e árduo trabalho inicial de datilografia dos manuscritos;

— A Senhorita Angela Aparecida Macedo e Senhora Ione de Macedo Sinieghi pela dedicação e eficiência no trabalho final de datilografia;

— O Senhor Jesus de Souza, pelo eficiente trabalho gráfico que realizou.

Finalmente, não poderíamos deixar de agradecer o apoio financeiro recebido da Coordenação do Aperfeiçoamento de Pessoal de Nível Superior (CAPES) durante o período de 1972 a 1974.

ÍNDICE

INTRODUÇÃO

A estatística, como instrumental auxiliar na análise de dados em quase todas as áreas de conhecimento humano, não poderia deixar de contribuir com a área de *Saúde Pública*, particularmente com a *Epidemiologia* que se preocupa com as distribuições e os determinantes da frequência da doença no homem. Neste trabalho serão abordados alguns aspectos dos estudos epidemiológicos destinados a medir os *graus de associação* entre fatores associados com as doenças em estudo.

As estratégias epidemiológicas mais conhecidas para essa finalidade são: *Estudos Naturalístico - Descritivo (Seccional)*, *Quase-Experimental Prospectivo (Coorte)* e *Quase-Experimental Retrospectivo (Caso-Controle)*. Mas, como o desenvolvimento da doença depende de vários fatores, não somente daquele considerado como *fator "causal" principal*, necessita-se utilizar métodos de controle dessas *variáveis interferentes ou efeito-modificantes*, visando evitar *associações espúrias*, principalmente nos estudos

retrospectivos (caso-controle).

Para minimizar essas associações espúrias, frequentemente utilizam-se a *subclassificação ou o pareamento*.

Com isso, é possível obter estimativas individuais de graus de associação em cada subclasse, medidas essas que tem como objetivo avaliar uma possível influência de um fator "causal" no surgimento de alguma alteração no estado de saúde de um indivíduo Encontradas estas *medidas de riscos* em cada subclassificação, testa-se a igualdade delas e caso isso se verifique, estima-se o *grau de associação ou o risco global* ("*over-all risk*").

As análises mais tradicionais acerca destes riscos, não levam em consideração a influência das variáveis que serviram de suporte na formação destes estratos, porém neste trabalho, será apresentada uma metodologia de análise que engloba todas as informações disponíveis para descrever estes riscos por meio de *modelo log-linear*.

No Capítulo zero serão apresentados alguns aspectos dos estudos epidemiológicos, definições de *risco relativo* e de "*odds ratio*" , com respectivos estimadores.

Os métodos de detecção de associação não causal em estudos caso-controle serão abordados no Capítulo 1.

tratada no Capítulo 2, que terá ilustrações a respeito. Ainda nesse capítulo será apresentado um *modelo condicionado de risco*, com enfoque diferente dos anteriores.

No último capítulo será feita a análise da *"razão dos produtos cruzados"* (*"cross-product ratio"*) nos estudos epidemiológicos retrospectivos através do modelo log-linear que permite descrever a influência das *variáveis intervenientes* nos riscos a uma doença sob estudo.

CAPÍTULO 0

ALGUNS ASPECTOS DOS ESTUDOS EPIDEMIOLÓGICOS

Este trabalho será iniciado com alguns conceitos básicos em epidemiologia que serão úteis para compreensão do tema proposto. Serão descritos três modelos de estudos epidemiológicos mais usuais e, serão definidos o *risco relativo* ("*relative risk*") e o *risco adicional* ("*odds ratio*"), com alguns comentários complementares.

0.a - CONCEITOS E ENFOQUES GERAIS

A *epidemiologia*, segundo MacMahon e Pugh (1970) é definida como "o estudo da distribuição e dos determinantes da frequência da doença no homem".

Entende-se por estudar a distribuição da doença, descrever a distribuição dos estados de doença segundo idade, sexo, raça, local, época e outras variáveis de interesse. Este tipo de estudo pode ser considerado como uma extensão da *demografia para a saúde e doença*. Para formular hipóteses e mesmo para planejar investigações epidemiológicas, as in-

formações disponíveis sobre as alterações na saúde da população ou variações na incidência de qualquer doença são apresentadas, mais comumente, em termos de variáveis descritivas de pessoa, local e época.

"Uma associação entre categorias de eventos ou características, na qual uma alteração na frequência ou qualidade de uma categoria é seguida por uma mudança na outra", MacMahon e Pugh (1970) define uma associação *causal*.

Este é outro interesse da epidemiologia: o conhecimento da etiologia da doença, ou seja, procurar explicar as distribuições em termos de fatores causais, o que na prática é muito difícil e mesmo, na maioria das vezes, impossível. Isto porque, o caminho mais lógico de verificar a associação causal seria o experimento direto, porém, no caso da epidemiologia, cujo objeto de estudo são as populações humanas, se torna inviável, por motivos óbvios.

Assim, a estratégia utilizada pelo epidemiologista é planejar estudos que sejam bem próximos daqueles ditos experimentais, razão pela qual muitas vezes são chamados *quase-experimentais*.

Realiza-se este tipo de estudo, selecionando as variáveis de interesse e procurando verificar se a associação é *casual* ou não. Em caso de não ser casual, mede-se o grau de associação, medida esta que em epidemiologia representa o risco de adquirir uma doença em função da exposição ou não

a algum fator suspeito.

Para medir o grau de associação entre o fator suspeito e a doença em estudo, os estudos são planejados conforme as possibilidades metodológicas oferecidas em cada situação.

A seguir, serão apresentados os três enfoques mais usuais em epidemiologia:

(i) modelo *naturalístico-descritivo ou seccional*

(ii) modelo *quase-experimental prospectivo ou coorte*

(iii) modelo *quase-experimental retrospectivo ou caso-controle.*

(i) Modelo Naturalístico Descritivo (Seccional):
É o tipo de estudo que consiste em selecionar uma amostra de T indivíduos da população a ser estudada e pesquisar a presença ou ausência dos fatores e acontecimentos epidemiológicos de interesse para a pesquisa. Não se faz qualquer estratificação antes da obtenção dos dados, nem a categorização dos fatores antecedentes. Muito frequentemente, nem se chega mesmo a explicitar qualquer hipótese causal, a priori.

EXEMPLO 0.a.1.

Selecionam-se por exemplo, T = 200 recém- nascidos para estudar a insuficiência ponderal do recém-nascido (IPRN). Para tal, coletam-se as informações sobre o peso do recém-

nascido, bem como o peso, idade, paridade, hábito de fumar e outras características maternas relevantes. Cada uma das características acima poderia ser considerada como um acontecimento antecedente e o peso do recém-nascido como o consequente. Para fins de exemplificação consideremos o tabagismo materno como fator antecedente de interesse, estratificando-o em dois sub-grupos - fumantes e não fumantes - e peso do recém-nascido em duas categorias - 2500 gramas ou menos (IPRN) e mais de 2500 gramas (normal) -, como se ilustra a seguir:

TABELA 0.a.A - Estudo naturalístico-descritivo, Tabagismo Materno e IPRN

Tabagismo Materno	Peso ao nascer		Total
	2500g ou menos (E)	mais de 2500g ($\bar{E}$)	
Fumantes (F)	10	40	50
Não fumantes ($\bar{F}$)	15	135	150
Total	25	175	200

Fonte: Adaptado de Fleiss (1973).

Neste tipo de estudo é possível estimar:

(a) A proporção de IPRN entre mães fumantes (*taxa de incidência* de IPRN para fumantes):

$$I_F = P(E/F) = \frac{P(EF)}{P(F)} \text{ ou } \widehat{P(E/F)} = \frac{0,05}{0,25} = 0,20$$

(b) A proporção de IPRN entre mães não fumantes (*taxa de incidência* de IPRN para não fumantes):

$$I_{\bar{F}} = P(E/\bar{F}) = \frac{P(E\bar{F})}{P(\bar{F})} \quad \text{ou} \quad \widehat{P(E/\bar{F})} = \frac{0,075}{0,75} = 0,10$$

(c) A proporção de mães fumantes no sub-grupo de crianças com IPRN:

$$P(F/E) = \frac{P(FE)}{P(E)} \quad \text{ou} \quad \widehat{P(F/E)} = \frac{0,05}{0,125} = 0,4$$

(d) A proporção de mães fumantes no sub-grupo de crianças com peso normal:

$$P(F/\bar{E}) = \frac{P(F\bar{E})}{P(\bar{E})} \quad \text{ou} \quad \widehat{P(F/\bar{E})} = \frac{0,20}{0,875} = 0,23$$

(ii) Modelo quase-experimental prospectivo (coorte).

É o modelo que se fundamenta, essencialmente, na comparação entre dois sub-grupos (de tamanhos M_1 e M_2) caracterizados antes da coleta de dados pela presença ou ausência do fator antecedente. Especificamente, a comparação é entre as frequências do acontecimento consequente, que no caso poderia ser a doença ou condição de saúde em estudo nos sub-grupos, frequências estas obtidas prospectivamente. A designação quase-experimental é no sentido de que se procura maior ou me-

nor aproximação do modelo experimental, sem contudo ser possível um controle completo sobre os fatores interferentes. Este tipo de estudo é também denominado de coorte.

EXEMPLO 0.a.2

Para verificar a relação entre o tabagismo materno e IPRN, o grupo de gestantes será subdividido segundo ser fumante ou não fumante, antes da coleta de informações acerca do peso do recém-nascido de mulheres de cada uma destas amostras. É, portanto, previamente estabelecida a caracterização dos sub-grupos quanto ao fator de risco considerado. Partindo de $M_1 = M_2 = 100$ obtém-se as informações como mostra a TABELA 0.a.B.

TABELA 0.a.B Estudo Prospectivo - Tabagismo Materno e IPRN

Tabagismo Materno	Peso ao nascer		Total
	2500 g ou menos (E)	mais de 2500 g ($\bar{E}$)	
Fumantes (F)	20	80	100
Não fumantes ($\bar{F}$)	10	90	100
Total	30	170	200

Fonte: Adaptado de Fleiss (1973)

Aqui é possível estimar a taxa de incidência de IPRN para fumantes.

$$I_F = P(E/F) = \frac{P(EF)}{P(E)} \quad \text{ou} \quad \widehat{P(E/F)} = \frac{0,1}{0,5} = 0,20$$

e a taxa de incidência de IPRN para não fumantes

$$I_{\bar{F}} = P(E/\bar{F}) = \frac{P(E\bar{F})}{P(\bar{F})} \quad \text{ou} \quad \widehat{P(E/\bar{F})} = \frac{0,05}{0,5} = 0,10$$

Mas não é possível estimar:

(a) A proporção de mães fumantes no sub-grupo de crianças com IPRN, e

(b) A proporção de mães fumantes no sub-grupo de crianças com peso normal.

Pois, a população base, onde as amostras são colhidas, é definida a partir de variáveis antecedentes, que no caso são mães fumantes e não fumantes e não a partir da população de recém-nascidos com IPRN e com peso normal.

(iii) Modelo quase-experimental retrospectivo (caso-controle) - É um modelo calcado na comparação entre dois ou mais grupos previamente definidos. É também denominado de estudo caso-controle e difere do modelo prospectivo no critério de caracterizar os sub-grupos. Seleciona-se um grupo de elementos com fator consequente (efeito) e um outro que não tenha o fator consequente, de tal forma que estes grupos sejam bem similares, tanto quanto possível, com exceção da presença ou

ausência do fator consequente, para que possa ser detectado o fator antecedente, eliminando ao máximo, a presença de associações espúrias.

EXEMPLO 0.a.3

No estudo da IPRN, selecionam-se adequadamente N_1= 100 recém-nascidos com peso igual ou abaixo de 2500 gramas e N_2 = 100 recém-nascidos com peso acima de 2500 gramas, e será estudada a história de tabagismo materno, resultando a TABELA 0.a.C:

TABELA 0.a.C Estudo Retrospectivo - Tabagismo Materno e IPRN

Tabagismo Materno	Peso ao nascer		Total
	2500 g ou menos (E)	mais de 2500g ($\bar{E}$)	
Fumantes (F)	40	23	63
Não fumantes ($\bar{F}$)	60	77	137
Total	100	100	200

Fonte: Adaptado de Fleiss (1973)

Este modelo não nos proporciona a estimação direta das taxas de incidência de IPRN como nos modelos anteriores, pois não se dispõe da população de mães com história de tabagismo.

Pode-se no entanto estimar a proporção de mães com história positiva de tabagismo nos sub-grupos de crianças com

IPRN e com peso normal, respectivamente:

$$P\ (F/E) = \frac{P(EF)}{P(E)} \quad \text{ou} \quad \widehat{P(F/E)} = \frac{0,2}{0,5} = 0,40$$

e

$$P(F/\bar{E}) = \frac{P(\bar{E}F)}{P(\bar{E})} \quad \text{ou} \quad \widehat{P(F/\bar{E})} = \frac{0,115}{0,5} = 0,23$$

De um modo geral, poder-se-ia apresentar o resumo das informações obtidas em qualquer das três abordagens citadas, por meio da uma tabela como:

TABELA 0.a.D Estudos Epidemiológicos - Fator e Efeito

Fator	Efeito		Total
	Presente (E)	Ausente ($\bar{E}$)	
Presente (F)	A (p_{11})	C (p_{12})	A+C= M_1 ($p_{1\cdot}$)
Ausente ($\bar{F}$)	B (p_{21})	D (p_{22})	B+D= M_2 ($p_{2\cdot}$)
Total	A+B=N_1 ($p_{\cdot 1}$)	C+D=N_2 ($p_{\cdot 2}$)	T (1)

A seguir, serão descritas as medidas de graus de associação que são de interesse neste trabalho.

0.b MEDIDAS DE GRAUS DE ASSOCIAÇÃO

É de conhecimento amplo que há várias medidas de graus de associação para dados representados por tabelas de

contingência. Neste momento, a TABELA 0.a.D será retomada, para dar base ao desenvolvimento de alguns conceitos acerca destas medidas.

Seja A o número de elementos com fator consequente dentre aqueles com o fator antecedente, C o número daqueles sem o fator consequente dentre aqueles com o fator antecendente, enquanto que B representa o número de elementos com o fator consequente dentre aqueles sem o fator antecedente e D indica o número de elementos sem o fator consequente dentre aqueles sem o fator antecedente.

0.b.1 Risco Relativo ("Relative Risk")

Num estudo prospectivo (coorte), após um certo período de tempo, pode-se obter a taxa de incidência da doença entre os indivíduos expostos a um fator em estudo:

$$I_F = P_{\mathcal{A}}\{\text{risco da doença na presença do fator }\} \quad (0.b.1)$$

enquanto que a taxa de incidência da doença entre os não expostos ao fator é:

$$I_{\bar{F}} = P_{\mathcal{A}}\{\text{risco da doença na ausência do fator}\} \quad (0.b.2)$$

e podem ser estimadas, respectivamente, por

$$\hat{I}_F = A/M_1 \quad \text{e} \quad \hat{I}_{\bar{F}} = B/M_2 \quad (0.b.3)$$

segundo as considerações da TABELA 0.a.D.

Comparando estas taxas citadas acima, Cornfield (1951) definiu um tipo de risco bastante usual na epidemiologia, o *Risco Relativo* (R) de desenvolver a doença como a razão de taxas de incidência entre aqueles com e sem fator:

$$R = \frac{P_{\mathcal{R}} \{\text{risco da doença na presença do fator}\}}{P_{\mathcal{R}} \{\text{risco da doença na ausência do fator}\}} = \frac{I_F}{I_{\bar{F}}} \qquad (0.b.4)$$

que pode ser estimado pelo quociente das expressões(0.b.3):

$$\hat{R} = \frac{A/M_1}{B/M_2} = \frac{A\ M_2}{B\ M_1} \qquad (0.b.5)$$

Como já foi visto anteriormente, no enfoque do estudo caso-controle é possível estimar somente as proporções das pessoas com e sem doença e que sofreram a ação do fator e não estimar o risco relativo. Porém, tal estimativa poderá ser derivada segundo Cornfield (1951 e 1956), em situações particulares.

Denotando por P, a proporção da população que apresenta a doença durante um período de interesse, também chamada de taxa de prevalência e, por p_1 e p_2, as proporções daquelas que possuiam a característica dentre as que desenvolveram e não desenvolveram a doença, respectivamente:

$p_1 = Pr$ {sofrer a ação do fator, dentre as que desenvolvem a doença} =

$= P\ (F/E)$

$p_2 = P_{\bar{h}}$ {sofrer a ação do fator, dentre as que não desenvolvem a doença} = $P\,(F/\bar{E})$

É fácil ver pelo enfoque de um estudo retrospectivo, que p_1 e p_2 são possíveis de serem estimados, mas a prevalência P não é. Porém, pelo tratamento dado por Cornfield (1951 e 1956) isso será contornado. Os estimadores são:

$$\hat{P} = N_1/T \qquad (0.b.6)$$

$$\hat{p}_1 = \hat{P}(F/E) = A/N_1 \qquad (0.b.7)$$

$$\hat{p}_2 = \hat{P}(F/\bar{E}) = C/N_2 \qquad (0.b.8)$$

A incidência da doença para aquelas que sofreram a ação do fator é:

$$I_F = \frac{p_1 P}{p_1 P + p_2(1-P)} \qquad (0.b.9)$$

e para aquelas sem a influência do fator é:

$$I_{\bar{F}} = \frac{(1-p_1)\,P}{(1-p_1)P + (1-p_2)(1-P)} \qquad (0.b.10)$$

O risco relativo, definido por (0.b.4), passará a ter agora uma outra formulação, devido ao novo enfoque, com o uso das equações (0.b.9) e (0.b.10):

$$R = \frac{I_F}{I_{\bar{F}}} = \frac{p_1}{1-p_1} \cdot \frac{(1-p_1)P + (1-p_2)(1-P)}{p_1\,P + p_2(1-P)} \qquad (0.b.11)$$

É fácil observar que, se P é pequena, a expressão (0.b.11) pode ser reduzida a

$$R_a = \frac{p_1}{1-p_1} \cdot \frac{1-p_2}{p_2} = \frac{P(F/E)\ P(\bar{F}/\bar{E})}{P(\bar{F}/E)\ P(F/\bar{E})} \qquad (0.b.12)$$

A expressão (0.b.12), obtida por Cornfield (1951) como uma aproximação do risco relativo, pode ser estimada por

$$\hat{R}_a = \frac{A/N_1}{B/N_1} \Big/ \frac{C/N_2}{D/N_2} = \frac{AD}{BC} \qquad (0.b.13)$$

Como a expressão (0.b.12) depende somente de p_1 e p_2, o estudo retrospectivo permite estimá-lo através de (0.b.13) ou do uso adequado de (0.b.7) e (0.b.8).

EXEMPLO 0.b.1

Considerem-se os dados de Breslow et al.(1954) onde de 518 pacientes com carcinoma pulmonar e de 518 controles, 499 e 462 respectivamente, eram fumantes.

Estes dados são apresentados na TABELA 0.b.A:

TABELA 0.b.A Estudo da influência do Hábito de Fumar e Câncer de Pulmão.

Hábito de fumar	Carcinoma Pulmonar	Controle	Total
Fumantes	499	462	961
Não Fumantes	19	56	75
Total	518	518	1036

Fonte: Breslow et al. (1954).

Utilizando (0.b.7) e (0.b.8) estimam-se:

$$\hat{p}_1 = 499/518 = 0,9633$$

$$\hat{p}_2 = 462/518 = 0,8919$$

Daí obteremos:

$$\hat{Ra} = \frac{0,9633}{0,0367} \quad \frac{0,1081}{0,8919} = 3,2$$

Assim, o risco relativo aproximado estimado de câncer de pulmão entre fumantes é de 3,2, num estudo tipo caso-controle.

0.b.11 Risco Adicional ("Odds Ratio")

Uma outra medida de grau de associação, dentre as várias existentes, é uma apresentada por Goodman e Kruskal (1954,1959), citada por Fleiss (1973), chamada de *"odds ratio"*, que Guedes (1976) chama de *risco adicional*, denominação esta que será mantida neste trabalho.

Trata-se de uma medida que depende de duas outras, definidas por:

(i) $\Omega_F = \dfrac{P_h\{\text{desenvolver a doença nos expostos ao fator}\}}{P_h\{\text{não desenvolver a doença nos expostos ao fator}\}} = \dfrac{P(E/F)}{P(\bar{E}/F)}$

(ii) $\Omega_{\bar{F}} = \dfrac{P_h\{\text{desenvolver a doença nos não expostos ao fator}\}}{P_h\{\text{não desenvolver a doença nos não expostos ao fator}\}} = \dfrac{P(E/\bar{F})}{P(\bar{E}/\bar{F})}$

que sendo reescritas, utilizando a notação da TABELA 0.a.D. se tornaram:

$$\text{(i)}\ \Omega_F = \frac{p_{11}/p_{1\cdot}}{p_{12}/p_{1\cdot}} = \frac{p_{11}}{p_{12}} \qquad (0.b.14)$$

$$\text{(ii)}\ \Omega_{\bar{F}} = \frac{p_{21}/p_{2\cdot}}{p_{22}/p_{2\cdot}} = \frac{p_{21}}{p_{22}} \qquad (0.b.15)$$

Assim, o risco adicional foi definido como o quociente de (0.b.14) por (0.b.15)

$$\Omega = \frac{\Omega_F}{\Omega_{\bar{F}}} = \frac{p_{11}\ \ p_{22}}{p_{12}\ \ p_{21}} \qquad (0.b.16)$$

O risco adicional é também chamado por alguns autores de "*razão dos produtos cruzados* ("*cross-product ratio*")-"RPC"- como é fácil deduzir pela expressão (0.b.16), observando a TABELA 0.a.D.

Conforme o enfoque dado na definição Ω_F e $\Omega_{\bar{F}}$, só os estudos naturalístico-descritivo e quase experimental prospectivo permitirão estimar o risco adicional. Porém, se for reescrito de outra forma, por um outro ângulo de abordagem, será possível estimá-lo num estudo quase-experimental retrospectivo:

$$\Omega = \frac{\Omega_F}{\Omega_{\bar{F}}} = \frac{P(E/F)\ .\ P(\bar{E}/\bar{F})}{P(\bar{E}/F)\ .\ P(E/\bar{F})} =$$

$$= \frac{\frac{P(EF)}{P(F)} \cdot \frac{P(\bar{E}\bar{F})}{P(\bar{F})}}{\frac{P(\bar{E}F)}{P(F)} \cdot \frac{P(E\bar{F})}{P(\bar{F})}} = \frac{P(EF) \cdot P(\bar{E}\bar{F})}{P(\bar{E}F) \cdot P(E\bar{F})} =$$

$$= \frac{\frac{P(EF)}{P(E)} \cdot \frac{P(\bar{E}\bar{F})}{P(\bar{E})}}{\frac{P(\bar{E}F)}{P(\bar{E})} \cdot \frac{P(E\bar{F})}{P(E)}} = \frac{P(F/E) \cdot P(\bar{F}/\bar{E})}{P(F/\bar{E}) \cdot P(\bar{F}/E)} \qquad (0.b.17)$$

Um estimador de (0.b.16) é

$$\hat{\Omega} = \frac{A/M_1}{C/M_1} \cdot \frac{D/M_2}{B/M_2} = \frac{AD}{BC} \qquad (0.b.18)$$

e um estimador de (0.b.17) é

$$\hat{\Omega} = \frac{A/N_1}{B/N_1} \cdot \frac{D/N_2}{C/N_2} = \frac{AD}{BC} \qquad (0.b.19)$$

que coincide com (0.b.18)

Logo, Ω pode ser indiferentemente estimado pelos três modelos de estudos epidemiológicos citados. A expressão (0.b.19) é a mesma apresentada por Cornfield (1951),(0.b.13). Assim, usaremos a denominação "*razão dos produtos cruzados*" ("RPC") indistintamente, tanto para citar o risco relativo aproximado (0.b.12), apresentado por Cornfield (1951), quanto para o risco adicional expresso por (0.b.16) ou (0.b.17) apresentado por Goodman e Kruskal (1954, 1959).

Exemplificando, tem-se:

(i) Modelo Naturalístico Descritivo (Seccional)

Pelos dados da TABELA 0.a.A, resulta de(0.b.5)

$\hat{R} = \frac{10 \times 150}{15 \times 50} = 2,0$, para risco relativo estimado, e por (0.b.18)

$\hat{\Omega} = \frac{10 \times 135}{15 \times 40} = 2,25$ para "*razão dos produtos cruzados*" estimada.

(ii) Modelo Quase-Experimental Prospectivo (Coorte)

Pelos dados da TABELA 0.a.B, resulta de (0.b.5)

$\hat{R} = \frac{20 \times 100}{10 \times 100} = 2,0$, para risco relativo estimado, e por (0.b.18)

$\hat{\Omega} = \frac{20 \times 90}{10 \times 80} = 2,25$, para "*razão dos produtos cruzados*" estimada.

(iii) Modelo Quase-Experimental Retrospectivo (Caso-Controle)

Pelos dados da TABELA 0.a.C., resulta de (0.b.19)

$\hat{\Omega} = \frac{40 \times 77}{60 \times 23} = 2,23$, para "*razão dos produtos cruzados*" estimada.

Neste caso não se pode estimar o risco relativo diretamente, pois é impossível estimar P(E/F) e P(E/$\bar{F}$), mas admitindo que a prevalência de IPRN é pequena, o que não é verdade na realidade, e segundo (0.b.13) o risco relativo seria aproximadamente igual a 2,23.

Outras considerações sobre a estimação do risco relativo serão retomadas com maiores detalhes no Capítulo 2.

Essas medidas de risco vêm sendo criticadas por distintos motivos. Por exemplo, Berkson (1958), citado em Fleiss (1970) diz que a razão de taxas como medida de associação não é uma descrição adequada, pois um crescimento de dez vezes sobre uma taxa de um por um milhão e um crescimento de dez vezes sobre uma de um por mil produzem o mesmo risco relativo, porém é evidente que o crescimento na segunda taxa é muito mais delicado e sério que no outro.

Ainda Berkson (1958) sustenta que uma simples diferença entre duas taxas é a medida aproximada da magnitude prática de uma associação em termos de saúde pública e ilustra isso com um exemplo da associação do fumo com câncer de pulmão e com doença coronariana.

Sheps (1958,1961), citada em Fleiss (1970), propos uma modificação apropriada ao índice de Berkson. Ela designa a taxa de mortalidade numa amostra controle por p_c, e por p_f a taxa correspondente na amostra em estudo, supondo que o risco seja elevado. É evidente que $p_c < p_f$. Ela argumenta também que o excesso de risco associado ao grupo em es

tudo, p_e, opera somente sobre aqueles indivíduos que não tenham tido a ocorrência do evento. Assim, pelo modelo citado será:

$$p_f = p_c + p_e(1-p_c)$$

isto é, a taxa do grupo em estudo p_f é a soma da taxa do grupo controle, p_c, e do excesso do risco, p_e, aplicado aqueles que por outro lado não tiveram o evento, $(1-p_c)$. Consequentemente, Sheps sugeriu o uso de:

$$p_e = \frac{p_f - p_c}{1-p_c} \qquad (0.b.20)$$

que chamou de *diferença relativa* ("*relative difference*") como medida do risco em excesso.

Nota-se que, se p_c for pequeno, p_e será próxima da diferença, que define o índice de Berkson.

Como o objetivo deste trabalho é descrever e medir o grau de associação entre fator antecedente e o efeito consequente nos estudos epidemiológicos do tipo caso-controle, a partir deste instante, toda a atenção será dirigida a esse tipo de estudo, sem partir para a linha de análise proposta por Berkson.

O sucesso no uso do estudo caso-controle está em conseguir grupos de casos e controles que difiram apenas quanto à presença ou ausência do fator em estudo, o que, na maioria das vezes, é impossível. Assim, é comum casos e contro-

les diferindo em relação a outras variáveis com efeito potencial na variável consequente e/ou variáveis relacionadas com a variável antecedente. Nesta situação, para se contornar o problema de possíveis associações espúrias, usa-se o procedimento do pareamento ou da estratificação dos casos e controles, segundo estratos das variáveis ditas intervenientes, fazendo-se a análise em cada estrato separadamente e, posteriormente, a global.

As várias colocações acima serão desenvolvidas nos próximos capítulos.

CAPÍTULO 1

TESTES PARA A DETECÇÃO DA ASSOCIAÇÃO

A seguir serão apresentados vários testes de hipóteses para detectar a presença de uma possível associação entre as variáveis antecedentes e consequentes nos estudos caso-controle.

1.0 ASSOCIAÇÃO ENTRE DUAS VARIÁVEIS EM ESTUDO CASO-CONTROLE

Sem dúvida, pelo fato do *risco relativo* e da *"razão dos produtos cruzados"* serem medidas muito usadas, há uma vasta literatura a respeito de como obtê-las, caso realmente haja uma associação, que é detectada por meio de testes de hipóteses adequados. Alguns procedimentos estatísticos a respeito serão apresentados a seguir.

Retome-se a TABELA 0.a.D, onde N_1 representa o número de doentes (casos) de A doentes com o fator em estudo e de B doentes sem o fator, enquanto que o número de elemen

tos sem a doença em estudo (controles) é N_2, formado por C elementos com o fator e D, sem o fator.

Independente de como medir o grau de associação, surgiria a questão de verificar a existência de associação. Serão estudadas várias maneiras de abordagem nos estudos epidemiológicos do tipo caso-controle.

1.a TABELA 2x2

Para o caso em que as informações podem ser sintetizadas em uma tabela 2x2, como a TABELA 0.a.D., utiliza-se o teste de homogeneidade: um qui-quadrado a um grau de liberdade, que serve para verificar se há diferença entre as proporções de indivíduos com o fator em estudo, nos casos e nos controles. Utilizando-se a correção de continuidade, a estatística será:

$$\chi^2 = \frac{(|AD-BC|-T/2)^2 T}{N_1 N_2 M_1 M_2} \qquad (1.a.1)$$

onde $T = N_1+N_2 = M_1+M_2 = A+B+C+D$

1.b VÁRIAS TABELAS 2x2

Conforme já foi mencionado anteriormente, pela dificuldade de encontrar casos e controles homogêneos, com exceção da presença ou ausência do fator em estudo, para evitar o problema de possíveis associações espúrias, será uti-

lizado o procedimento do pareamento ou da estratificação dos casos e controles, segundo as variáveis ditas intervenientes.

Por exemplo, em um estudo sobre a associação entre câncer de pulmão e hábito de fumar, diante de estruturas de idade diferentes entre casos e controles, é aconselhável considerar a distribuição de casos e controles em relação às categorias de fumantes, segundo estratos de idade, pois o fator idade, que poderá estar relacionado tanto com o câncer do pulmão, quanto com o hábito de fumar, poderia alterar o grau e mesmo, o sentido da verdadeira associação. Nestas situações, o termo genérico para representar o j-ésimo sub-grupo (estrato) dos k formados é:

TABELA 1.b.A - Informações do estrato j em um estudo caso-controle

Estrato j	Casos	Controles	Total
Com Fator	A_j	C_j	M_{1j}
Sem Fator	B_j	D_j	M_{2j}
Total	N_{1j}	N_{2j}	T_j

$j = 1,2...,k$

Neste subgrupo, a *"razão dos produtos cruzados"* dos elementos doentes (casos) com o fator, relativamente àqueles ca

sos sem o fator, é estimada por:

$$\hat{\Omega}_j = A_j D_j / B_j C_j$$

$$j = 1,2,\ldots,k \qquad (1.b.1)$$

Sob a hipótese do risco considerado ser unitário, o número esperado de doentes com o fator será:

$$E[A_j] = N_{1j} M_{1j} / T_j$$

$$j = 1,2,\ldots,k \qquad (1.b.2)$$

e a variância:

$$Var[A_j] = N_{1j} N_{2j} M_{1j} M_{2j} / T_j^2 (T_j - 1)$$

$$j = 1,2,\ldots,k \qquad (1.b.3)$$

Assim, a estatística de teste, que tem distribuição quiquadrado a um grau de liberdade, com a correção de continuidade, será:

$$\chi_j^2 = \frac{(|A_j - E[A_j]| - 1/2)^2}{Var[A_j]}$$

$$j = 1,2,\ldots,k$$

o que se reduz a:

$$\chi_j^2 = \frac{(|A_j D_j - B_j C_j| - T_j/2)^2 (T_j - 1)}{N_{1j} N_{2j} M_{1j} M_{2j}}$$

$$j = 1,2,\ldots,k \qquad (1.b.4)$$

Com isso, resolve-se o problema de detectar a presença de associação para cada subclassificação, porém nada se pode concluir sobre a presença de associação entre o fator e a doença no estudo inicialmente considerado. Um teste de hipótese, que englobaria todas as informações das diversas subclassificações, é efetuado através de um quiquadrado, a um grau de liberdade, expresso por:

$$X^2 = \frac{\left(\left|\sum_{j=1}^{k} A_j - \sum_{j=1}^{k} E[A_j]\right| - 1/2\right)^2}{\sum_{j=1}^{k} Var[A_j]} \qquad (1.b.5)$$

onde $E[A_j]$ e $Var[A_j]$ são expressas por (1.b.3) e (1.b.4), respectivamente.

Fleiss (1973) cita um meio de detectar a presença de associação em, pelo menos, um dos subgrupos em estudo. Ele esquematizou cada subgrupo, como na TABELA 1.b.A, e denotou o valor da medida de associação em um dos k subgrupos, digamos o j-ésimo, por Ψ_j e o seu erro padrão por e.p. $[\Psi_j]$. Definiu o peso, w_j, a ser atribuído ao Ψ_j, como sendo o inverso da variância:

$$w_j = 1/(e.p.[\Psi_j])^2$$

$$j = 1,2,\ldots,k \qquad (1.b.6)$$

Note que a expressão (1.b.6) faz sentido, pois, quanto maior for o e.p. $[\Psi_j]$ (isso significa que Ψ_j é pouco preciso) a contribuição deste valor deve ter peso menor e vice-versa. Supor que Ψ_j tenha seu valor igual a zero indica não associação, ou seja, Ψ_j poderia ser por exemplo, diferença entre proporções ou logaritmo da *"razão dos produtos cruzados"*. Sob a hipótese de nulidade, ocorre que

$$X_j = \Psi_j / e.p.[\Psi_j] = \Psi_j \sqrt{wj}$$

$$j = 1,2,\ldots,k$$

tem, aproximadamente, uma distribuição normal padronizada. Segue que

$$X_j^2 = W_j \Psi_j^2$$

$$j = 1,2,\ldots,k$$

tem, aproximadamente, uma distribuição quiquadrado com um grau de liberdade.

O interesse não é num particular grupo e sim na totalidade deles, e para tal, é válido calcular:

$$X_{TOT}^2 = \sum_{j=1}^{k} X_j^2 = \sum_{j=1}^{k} W_j \Psi_j^2$$

Como os k grupos são supostos independentes, X_{TOT}^2, tem aproximadamente, uma distribuição quiquadrado com k graus

de liberdade. Assim, se for detectada uma diferença ao comparar o χ^2_{TOT} com $\chi^2_{(k)}$ tabelado, pode-se concluir que deve haver uma associação em pelo menos um dos k subgrupos. Contudo, não é possível especificar o sub-grupo. Pelo fato de só o χ^2_{TOT} não ser suficiente para fornecer as informações desejadas, separa-se o quiquadrado em dois componentes:

$$\chi^2_{TOT} = \chi^2_{HOMO} + \chi^2_{ASSOC} \qquad (1.b.7)$$

onde

χ^2_{HOMO} avalia o grau de homogeneidade ou igualdade entre as k medidas de associação Ψ_j e χ^2_{ASSOC} avalia a significância do grau de associação.

Tem-se que

$$\chi^2_{ASSOC} = \bar{\Psi}^2 \sum_{j=1}^{k} W_j = \left[\sum_{j=1}^{k} \Psi_j W_j\right]^2 / \sum_{j=1}^{k} W_j \qquad (1.b.8)$$

é distribuído, aproximadamente, como quiquadrado a um grau de liberdade.

Consequentemente,

$$\chi^2_{HOMO} = \chi^2_{TOT} - \chi^2_{ASSOC} = \sum_{j=1}^{k} W_j \bar{\Psi}^2_j - \bar{\Psi}^2 \sum_{j=1}^{k} W_j$$

ou

$$\chi^2_{HOMO} = \sum_{j=1}^{k} W_j (\Psi_j - \overline{\Psi})^2 \qquad (1.b.9)$$

tem aproximadamente, uma distribuição quiquadrado a (k-1) graus de liberdade, sob a hipótese de igualdade nas k medidas de associação.

Do que foi exposto, tiram-se as conclusões:

(i) A homogeneidade da associação pode ser testada através de (1.b.9):

(a) se for significante, deve-se separar o χ^2_{HOMO} em componentes apropriados para detectar individualmente os grupos onde a associação difere dos demais. Esse processo de partição é descrito por Fleiss (1973), com detalhes.

(b) se não for significante, pode-se testar a significância do grau de associação geral através de (1.b.8).

(ii) A melhor medida do grau de associação geral será expressa por:

$$\overline{\Psi} = \sum_{j=1}^{k} W_j \Psi_j \Big/ \sum_{j=1}^{k} W_j$$

e o seu erro padrão será:

$$e.p.[\bar{\Psi}] = \left[\sum_{j=1}^{k} W_j\right]^{-1/2}$$

Ainda Fleiss (1973) ilustra algumas aplicações utilizando-se de:

1. Método de Cochran, através da *"diferença padronizada"* também recomendado por Yates (1955).

2. Combinação do Logaritmo da *"Razão dos Produtos Cruzados"*, abordado por Woolf (1955), Gart (1962) e Sheehe (1966). Naylor (1967) e Gart (1970) mostraram que o estimador citado e o seu erro padrão são viciados quando os tamanhos das amostras são pequenos.

1.c. ESTUDO DE AMOSTRA PAREADA

Como foi dito anteriormente, para os estudos do tipo caso-controle, a fim de evitar associações espúrias entre os fatores de estudo, utiliza-se o pareamento. Este consiste em selecionar, para cada caso (indivíduos com o fator consequente), um controle (indivíduos sem o fator consequente) que seja similar em relação a várias variáveis associadas com o fator antecedente e/ou consequente, capazes de levar a associações espúrias. Ver TABELA 1.c.A, sobre *"Estudo de Amostra Pareada"* na página seguinte.

Os totais marginais A, B, C e D são aqueles da TABELA 0.a.D.

TABELA 1.c.A - Estudo de amostra pareada

Controles	Casos		Total
	Com Fator	Sem Fator	
Com Fator	F	G	C
Sem Fator	H	J	D
Total	A	B	N

Observando-se o esquema montado na TABELA 1.c.A, espera-se que, na ausência de associação entre a doença e o fator em estudo, apareça o mesmo número de indivíduos com o fator tanto entre os doentes como entre os sadios, isto é, espera-se que

$$A = F+H \text{ seja igual a } C = F+G$$

É fácil concluir que isso ocorre somente quando G=H e o teste será, simplesmente, verificar se G difere significantemente de 50% da soma G+H, o que leva a testar G como uma variável binomial com parâmetros 1/2 e G+H. Assim

$$E[G] = (G+H)/2 \text{ e } Var[G] = (G+H)/4$$

Obtém-se o quiquadrado a um grau de liberdade para a variável considerada que, com a correção de continuidade, torna-se

$$\chi^2 = (|G-H|-1)^2/(G+H) \qquad (1.c.1)$$

Note que se fosse abordado o problema como consistindo de N classificações, onde cada $N_{1j} = N_{2j} = 1$, logo $T_j = 2$, e aplicando os procedimentos já vistos, obter-se-ia o mesmo quiquadrado apresentado por (1.c.1), utilizando-se(1.b.5.).

Tendo para F das N classificações:

$$A_j = 1,\ M_{1j} = 2,\ M_{2j} = 0,\ E[A_j] = 1,\ Var[A_j] = 0$$

Para G:

$$A_j = 0,\ M_{1j} = M_{2j} = 1,\ E[A_j] = 1/2,\ Var[A_j] = 1/4$$

Para H:

$$A_j = 1,\ M_{1j} = M_{2j} = 1,\ E[A_j] = 1/2,\ Var[A_j] = 1/4$$

Para J:

$$A_j = 0,\ M_{1j} = 0, M_{2j} = 2,\ E[A_j] = 0,\ Var[A_j] = 0$$

Então:

$$\sum_{j=1}^{N} A_j = F+H,\ \sum_{j=1}^{N} E[A_j] = F+\frac{1}{2}(G+H) \text{ e } \sum_{j=1}^{N} Var[A_j] = \frac{1}{4}(G+H)$$

que resulta na expressão (1.c.1).

1.d. TABELA DE CONTINGÊNCIA 2xk

A discussão agora é estendida aos estudos onde os fatores em observação assumem mais de dois níveis, pois na prática isso se verifica com grande frequência e é de suma importância levar em consideração, na análise, os valores intermediários do fator em vez de somente dois extremos, como até o momento. Exemplificando:

Consideremos um grupo de mulheres não fumantes, e outro de habitualmente fumantes de mais de um maço de cigarros por dia.

Verificou-se que há uma diferença entre os grupos ao se calcular o risco relativo de adquirir uma doença. Agora, o fato de incluir um outro grupo, o de mulheres fumantes de até um maço de cigarros por dia, produzindo um resultado intermediário entre as outras classes, pouco teria contribuído, praticamente em nada, para a decisão estatística. Um teste geral das diferenças entre as três classes consideradas poderia deixar de detectar uma diferença porventura existente. Porém, o fato de ter considerado o terceiro grupo, produzindo risco relativo intermediário, contribuiria para fortalecer a confiança nas conclusões sugeridas anteriormente pelos dois grupos extremos considerados.

Mantel (1963) sugeriu um esquema considerando que o fator sob estudo assumisse ℓ níveis ordenáveis, resultando

uma tabela de contingência 2xℓ, como mostra a TABELA 1.d.A, genérica para o j-ésimo experimento, dentre os k experimentos:

TABELA 1.d.A - Estudo caso-controle para vários níveis do fator

Experimento j j= 1,2,...,k	X	y	Nível do Fator em Estudo 1	2	...	i	...	ℓ	Total
			y_1	y_2	...	y_i	...	y_ℓ	
Casos	1		A_{1j}	A_{2j}	...	A_{ij}	...	$A_{\ell j}$	N_{1j}
Controles	0		B_{1j}	B_{2j}	...	B_{ij}	...	$B_{\ell j}$	N_{2j}
Total	-		M_{1j}	M_{2j}	...	M_{ij}	...	$M_{\ell j}$	T_j

Note-se que foram fixados os valores.

X = 1 para elementos com a doença e

X = 0 para elementos livres da doença em estudo, assim como os escores y_i para níveis de 1 a ℓ.

O resultado da análise, sugerida por Mantel (1963), pode variar acentuadamente, dependendo do tipo de escore utilizado para os vários níveis do fator. Não há uma regra ou um procedimento único para estabelecer estes escores. Há sim, sempre, a indicação dos escores seguindo uma lógica composi

cional dos níveis dos fatores, por meio de uma observação adequada e bom senso em quantificar objetivamente os níveis. Há alguns tipos de escores já consagrados que serão exemplificados abaixo:

(i) Utilizar o escore i para o i-ésimo nível do fator é útil, pois é bem simples, e tem o poder estatístico aumentado quando se está diante de efeitos progressivos do fator. Isto não implica em que os níveis sejam tomados, necessariamente, equiespaçados.

(ii) Especificamente, para casos de fumantes de cigarro, pode-se atribuir como escore, o número médio diário de cigarros, por exemplo.

(iii) Outro comumente utilizado é o de postos. Pode-se utilizar um posicionamento separado para cada tabela de contingência ou para todas as tabelas combinadas. A fim de padronizar o número variado de indivíduos nas diferentes tabelas, os escores-postos serão expressos relativamente aos T_j.

(a) Para tabelas separadas, tem-se:

$$Y_{1j} = \frac{1}{T_j}\left(\frac{M_{1j}+1}{2}\right)$$

$$Y_{2j} = \frac{1}{T_j}\left(M_{1j} + \frac{M_{2j}+1}{2}\right)$$

ou genericamente

$$Y_{rj} = \frac{1}{T_j}\left(\sum_{i=0}^{r-1} M_{ij} + \frac{M_{rj}+1}{2}\right)$$

sendo $r = 1,2,3,4,\ldots,\ell$, com $M_{oj} = 0$ para todos os $j = 1,2,\ldots,k$.

(b) Para todas as tabelas juntas, tem-se:

$$Y_r = \frac{1}{\sum_{j=1}^{k} T_j}\left[\sum_{j=1}^{k}\sum_{i=1}^{r-1} M_{ij} + \frac{\sum_{j=1}^{k} M_{rj}+1}{2}\right]$$

com $M_{oj} = 0$ para todos os $j = 1,2,3,\ldots,k$ e $r = 1,2,3,\ldots,\ell$.

A estatística que Mantel (1963) utilizou foi:

$$\Sigma_j XY = \sum_{i=1}^{\ell}\left[1xA_{ij} + 0xB_{ij}\right] Y_{ij} = \sum_{i=1}^{\ell} A_{ij}Y_{ij}$$

para $j = 1,2,\ldots,k$. (1.d.1)

Note que sob a hipótese

Ho: não há associação entre fator e doença,

(1.d.2)

o valor esperado da estatística é:

$$E\left[\Sigma_j XY\right] = T_j \mu_{jx} \mu_{jy}$$

$$j = 1,2,\ldots,k \qquad (1.d.3)$$

mas como $\mu_{jx} = N_{ij}/T_j$ e $\mu_{jy} = \sum_{i=1}^{\ell} M_{ij}Y_{ij}/T_j$, a expressão (1.d.3) ficará:

$$E\left[\Sigma_j XY\right] = N_{1j} \sum_{i=1}^{\ell} M_{ij}Y_{ij}/T_j$$

$$j = 1,2,\ldots,k$$

e a variância é dada por:

$$Var\left[\Sigma_j XY\right] = N_{1j}N_{2j}\sigma_j^2/(T_j-1)$$

$$j = 1,2,\ldots,k \qquad (1.d.4)$$

onde σ_j^2 é a variância populacional que é expressa por:

$$\sigma_j^2 = \frac{1}{T_j}\left[\sum_{i=1}^{\ell} M_{ij}Y_{ij}^2 - \frac{(\sum_{i=1}^{\ell} M_{ij}Y_{ij})^2}{T_j}\right]$$

e $\frac{N_{2j}}{T_j-i} = \frac{T_j-N_{1j}}{T_j-1}$ é o fator de correção para população finita. Daí, (1.d.4) será:

$$Var\left[\Sigma_j XY\right] = \frac{N_{1j}N_{2j}}{T_j^2(T_j-1)}\left[T_j \sum_{i=1}^{\ell} M_{ij}Y_{ij}^2 - (\sum_{i=1}^{\ell} M_{ij}Y_{ij})^2\right]$$

$$j = 1,2,\ldots,k. \qquad (1.d.5)$$

Por consequência, o quiquadrado com um grau de liberdade, pode ser computado:

(i) para uma tabela de contingência simples:

$$\chi^2 = \frac{\left[\sum_{i=1}^{\ell} A_{ij}Y_{ij} - \frac{N_{1j}}{T_j}\sum_{i=1}^{\ell} M_{ij}Y_{ij}\right]^2}{\mathrm{Var}\left[\sum_{j=1}^{\ell} A_{ij}Y_{ij}\right]}$$

(1.d.6)

(ii) para uma tabela de contingência considerando os k níveis:

$$\chi^2 = \frac{\left\{\sum_{j=1}^{k}\sum_{i=1}^{\ell} A_{ij}Y_{ij} - \sum_{j=1}^{k} E\left[\sum_{i=1}^{\ell} A_{ij}Y_{ij}\right]\right\}^2}{\sum_{j=1}^{k} \mathrm{Var}\left[\sum_{i=1}^{\ell} A_{ij}Y_{ij}\right]}$$

(1.d.7)

desde que os efeitos de quaisquer outros fatores sejam mantidos constantes.

1.e. TABELAS DE CONTINGÊNCIA rxk

Serão agora analisadas as situações nas quais não só o fator está classificado em vários níveis, mas também o efeito, que para o caso particular da epidemiologia é a doença, considerada em seus vários estágios de desenvolvimento.

Considere-se a possibilidade de graduar as doenças, da mesma forma como se fez com o fator.

Tomando X_i como escore do estado de doença e Y_i como escore do nível do fator do i-ésimo indivíduo observado, resulta, para análise dos dados, a estatística:

$$\sum_{i=1}^{T_j} X_{ij} Y_{ij}$$

$$j = 1,2,\ldots,k \qquad (1.e.1)$$

para o j-ésimo conjunto de dados. O desvio de (1.e.1) do seu valor esperado é expresso por:

$$\sum_{i=1}^{T_j} X_{ij} Y_{ij} - \frac{\sum_{i=1}^{T_j} X_{ij} \sum_{i=1}^{T_j} Y_{ij}}{T_j}$$

$$j = 1,2,\ldots,k.$$

A variância de (1.e.1), sob a hipótese de independência entre X_{ij} e Y_{ij}, e utilizando-se do conceito de população finita, é:

$$\text{Var}\left[\sum_{i=1}^{T_j} X_{ij} Y_{ij}\right] = \frac{T_j^2}{T_j - 1} \sigma_j^2{}_x \sigma_j^2{}_y$$

(1.e.2)

onde

$$\sigma^2_j x = \left[\sum_{i=1}^{T_j} X_{ij}^2 - \frac{\left(\sum_{i=1}^{T_j} X_{ij} \right)^2}{T_j} \right] \Big/ T_j$$

$$j = 1, 2, \ldots, k.$$

e

$$\sigma^2_j y = \left[\sum_{i=1}^{T_j} Y_{ij}^2 - \frac{\left(\sum_{i=1}^{T_j} Y_{ij} \right)^2}{T_j} \right] \Big/ T_j$$

Quando há vários conjuntos de dados, pode-se computar o quiquadrado a um grau de liberdade:

$$\chi^2 = \frac{\left\{ \sum_{j=1}^{k} \left[\sum_{i=1}^{T_j} X_{ij} Y_{ij} - \frac{\sum_{i=1}^{T_j} X_{ij} \sum_{i=1}^{T_j} Y_{ij}}{T_j} \right] \right\}^2}{\sum_{j=1}^{k} \left\{ \frac{1}{T_j - 1} \left[\sum_{i=1}^{T_j} X_{ij}^2 - \frac{\left(\sum_{i=1}^{T_j} X_{ij} \right)^2}{T_j} \right] \left[\sum_{i=1}^{T_j} Y_{ij}^2 - \frac{\left(\sum_{i=1}^{T_j} Y_{ij} \right)^2}{T_j} \right] \right\}} \qquad (1.e.3)$$

O uso de (1.e.3) equivale a testar um simples coeficiente de regressão para um único conjunto de dados ou um coeficiente de regressão conjunto com vários conjuntos de dados, mas isso não significa testar a existência de uma linea

ridade dos dados, e sim tentar detectar qualquer associação progressiva que poderá existir.

Já vistos alguns procedimentos de como detectar a existência de associação nos vários enfoques epidemiológicos adequados, trataremos de estudar os meios de medir os graus de associação, assunto esse que será discutido a seguir.

CAPÍTULO 2

ESTIMAÇÃO DO RISCO RELATIVO NOS ESTUDOS CASO-CONTROLE

No presente capítulo, serão apresentados alguns enfoques para a estimação do risco relativo, tanto por intervalo quanto por ponto.

2.a ESTIMATIVA POR INTERVALO

Limites razoáveis para o risco relativo

Um problema de grande interesse é a obtenção de limites razoáveis para o risco relativo quando de fato é diferente da unidade. Seguindo a abordagem de Cornfield (1956), o intervalo proposto é equivalente para os casos em que os limites de confiança incluem ou não, a unidade.

2.a.1 TABELA 2x2

Seja a TABELA 2.a.A que representa o caso mais simples de classificação de dados observados:

TABELA 2.a.A Esquema de um Estudo Caso-Controle.

Característica	Casos	Controles	Total
Com	X (p_1)	Y (p_2)	M
Sem	N_1-X ($1-p_1$)	N_2-Y ($1-p_2$)	T-M
Total	N_1	N_2	T

As variáveis X e Y definem o número de indivíduos com a característica em estudo nas amostras de tamanho n_1 (casos) e n_2 (controles), respectivamente. Considerando-se a proporção p_1 , presença da característica nos casos, e p_2 , presença da característica nos controles, define-se a sua probabilidade conjunta pela expressão:

$$\Pr\{X = x,\ Y = y\} = \binom{n_1}{x} p_1^{\,x}\, q_1^{\,n_1-x} \binom{n_2}{y} p_2^{\,y}\, q_2^{\,n_2-y}$$

fixados os totais marginais, resulta a probabilidade condicional:

$$\Pr\{X=x/X+Y=m\} = \Pr\{X=x/Y=m-x\} = \frac{\Pr\{X=x,\ Y=m-x\}}{\Pr\{Y=m-x\}}$$

ou

$$\Pr\{X=x/X+Y=m\} = \frac{\binom{n_1}{x}\binom{n_2}{m-x}\Omega^x}{\sum_{u=0}^{n_1}\binom{n_1}{u}\binom{n_2}{m-u}\Omega^u} = \mathbb{C}\cdot\binom{n_1}{x}\binom{n_2}{m-x}\Omega^x$$

$$\text{para } x = 0,1,2\ldots,n_1 \qquad (2.a.1)$$

onde

$$\Omega = \left[p_1 \ (1-p_2)\right] \Big/ \left[p_2 (1-p_1)\right]$$

O intervalo de confiança, com um grau de confiança $(1-\alpha)$ para o parâmetro Ω será estimado, através da solução das equações:

$$\sum_{t=0}^{x} \binom{n_1}{t} \binom{n_2}{m-t} \Omega^t = \alpha/2 \qquad (2.a.2)$$

$$\sum_{t=x}^{\hat{n_1}} \binom{n_1}{t} \binom{n_2}{m-t} \Omega^t = \alpha/2 \qquad (2.a.3)$$

De (2.a.2) obtém-se Ω_2 e de (2.a.3), Ω_1 , que são os limites do intervalo.

$$\Omega_1 \leq \Omega \leq \Omega_2 \qquad (2.a.4)$$

Pelo fato da assimetria de (2.a.1), o intervalo (2.a.4) poderá ter seus limites viciados. Isso tem uma analogia com o teste exato de Fisher, de independência numa tabela 2x2. Quando o min (n_1, n_2, m) não for muito pequeno, podem-se utilizar os métodos numéricos para obter soluções das equações (2.a.2) e (2.a.3). Caso contrário, a computação é muito difícil, embora seja válida a tentativa de evitar que leve às investigações de aproximações assintóticas.

A fim de conseguir a distribuição limite para (2.a.1),

que será denotada por P(X), e contornar a dificuldade inicial da obtenção da constante $\mathbb{C}$, e consequentemente, a obtenção das expressões da esperança e da variância de X, procuraremos a distribuição limite da razão: P(X)/P($\tilde{X}$), onde $\tilde{X}$ é a moda da distribuição (2.a.1). Segundo Cornfield (1956), resulta:

$$\frac{\tilde{X}\ (n_2-m+\tilde{X})}{(n_1-\tilde{X}+1)\ (m-\tilde{X}+1)} \leqslant \Omega \leqslant \frac{(\tilde{X}+1)\,(n_2-m+\tilde{X}+1)}{(n_1-\tilde{X})\,(m-\tilde{X})} \qquad (2.a.5)$$

Porém, se a amostra é grande, é suficiente obedecer a igualdade:

$$\frac{\tilde{X}\ (n_2-m+\tilde{X})}{(n_1-\tilde{X})\ (m-\tilde{X})} = \Omega \qquad (2.a.6)$$

Essas considerações são utilizadas para concluir, após várias suposições e cálculos matemáticos, que a distribuição limite desejada será normal de média $\tilde{X}$ e a sua variância definida por:

$$\left[\frac{1}{\tilde{X}} + \frac{1}{n_1-\tilde{X}} + \frac{1}{m-\tilde{X}} + \frac{1}{n_2-m+\tilde{X}}\right]^{-1} \qquad (2.a.7)$$

Chamando de $\tilde{X}_2$ a maior raiz real da equação na variável $\tilde{X}$:

$$(X-x-\frac{1}{2})^2 \left[\frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}}\right] = \chi^2_\alpha \qquad (2.a.8)$$

e $\tilde{X}_1$, a menor raiz real de:

$$(\tilde{X}-x+\frac{1}{2})^2 \left[\frac{1}{\tilde{x}}+\frac{1}{n_1-\tilde{x}}+\frac{1}{m-\tilde{x}}+\frac{1}{n_2-m+\tilde{x}}\right] = \chi^2_\alpha \qquad (2.a.9)$$

onde χ^2_α é o ponto com probabilidade α(superior) da distribuição quiquadrado com um grau de liberdade, 1/2 aparece como fator de correção de continuidade da distribuição (2.a.1) e, x é a observação amostral.

Usando um método iterativo para solução de (2.a.8) e (2.a.9), resultará o intervalo para $\tilde{X}$, com um grau de confiança $(1-\alpha)$.

$$\tilde{X}_1 \leqslant \tilde{X} \leqslant \tilde{X}_2$$

É automática a obtenção do intervalo de confiança para o parâmetro Ω, através de (2.a.6):

$$\Omega_1 \leqslant \Omega \leqslant \Omega_2$$

que é assintoticamente correta, com um grau de confiança $(1-\alpha)$.

EXEMPLO 2.a.1

É dada a TABELA 2.a.B de fumantes e não fumantes em dois grupos, um com câncer de pulmão e outro sem (controles), como foi ilustrada por Cornfield (1956):

TABELA 2.a.B Tabagismo e câncer de pulmão

Tabagismo	Câncer de Pulmão	Controles	Total
Não fumantes	3	11	14
Fumantes	60	32	92
Total	63	43	106

Fonte: Wynder e Cornfield (1953)

Aqui deseja-se estimar o risco relativo dos não fumantes sofrerem de câncer de pulmão em relação aos fumantes, através da *"razão dos produtos cruzados"* ("RPC").

Tem-se que m = 14 e x = 3 e para $1-\alpha=0,95$, determina-se o intervalo para o risco esperado. Calculando $\tilde{X}$ através das equações (2.a.8) e (2.a.9), com $\chi^2_\alpha = 3,84$, pelo método iterativo, obtém-se:

$$\tilde{X}_1 = 0,815 \quad \text{e} \quad \tilde{X}_2 = 6,905$$

Levando a (2.b.6), resulta:

$$0,0296 \leq \Omega \leq 0,6229 \qquad (2.a.10)$$

Conclui-se pois, com um grau de confiança de 95%, que o risco dos não fumantes em relação ao dos fumantes é de 62%, no máximo, e de 3%, no mínimo.

Para $1-\alpha = 0,99$, tendo $\chi^2_\alpha = 6,635$, obtém-se:
$\tilde{X}_1 = 0,59$ e $\tilde{X}_2 = 7,94$, resultando daí:

$$0,02909 \leq \Omega \leq 0,8790 \qquad (2.a.11)$$

Estes valores, (2.a.10) e (2.a.11), levados ao cálculo exato resultam nas probabilidades extremas superiores estimadas, que são:

(i) De 6,2% quando na realidade desejamos 5% e

(ii) De 2,1% quando na realidade queremos 1%.

2.a.II Várias Tabelas 2x2

Neste caso, seria útil abordar dois problemas que surgem:

(i) Estimar a extensão do intervalo para o qual os riscos relativos podem diferir, e

(ii) Obter uma maneira de combinar os resultados para aqueles que não parecem diferir, estatisticamente falando.

A TABELA 2.a.C da página seguinte descreve a situação.

Esquematizando, genericamente, um estudo ou uma subclassificação (j) resulta a TABELA 2.a.D da página seguinte.

A probabilidade não condicional nos k estudos, considerando X_{1j} doentes com fator, com os totais marginais fixados, é expressa por:

$$\prod_{j=1}^{k} \binom{n_{1j}}{x_{1j}} \binom{n_{2j}}{m_j - x_{1j}} \Omega_j^{x_{1j}}$$

, com $m_j = x_{1j} + x_{2j}$ e Ω_j é a "RPC" esperada do j-ésimo estudo.

TABELA 2.a.C K estudos independentes do mesmo fenômeno.

Número do estudo	casos		controles	
	Com fator	total	com fator	total
1	X_{11}	N_{11}	X_{21}	N_{21}
2	X_{12}	N_{12}	X_{22}	N_{22}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
j	X_{1j}	N_{1j}	X_{2j}	N_{2j}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
k	X_{1k}	N_{1k}	X_{2k}	N_{2k}
Total	$X_1=n$	N_1	$X_2=m-n$	N_2

TABELA 2.a.D Estudo j do fenômeno pesquisado

j	Casos	Controles	Total
Com fator	X_{1j} (p_{1j})	X_{2j} (p_{2j})	M_j
Sem fator	$N_{1j}-X_{1j}$ $(1-p_{1j})$	$N_{2j}-X_{2j}$ $(1-p_{2j})$	T_j-M_j
Total	N_{1j}	N_{2j}	T_j

$j = 1,2,3,\ldots,k$

Note que a probabilidade condicional com a restrição de que

$$\sum_{j=1}^{k} X_{1j} = n$$ vale:

$$\mathbb{C} \cdot \begin{pmatrix} n_{1k} \\ \hbar - \sum_{j=1}^{k-1} x_{1j} \end{pmatrix} \begin{pmatrix} n_{2k} \\ m_k - \hbar + \sum_{j=1}^{k-1} x_{1j} \end{pmatrix} \prod_{j=1}^{k-1} \begin{pmatrix} n_{1j} \\ x_{1j} \end{pmatrix} \begin{pmatrix} n_{2j} \\ m_j - x_{1j} \end{pmatrix} \theta_j^{x_{1j}} \tag{2.a.12}$$

sendo $\theta_j = \Omega_j / \Omega_k \qquad j = 1,2,\ldots,k-1$ (2.a.13)

$$\frac{1'}{\mathbb{C}} = \sum_{X_{11}} \sum_{X_{12}} \cdots \sum_{X_{1k-1}} \begin{pmatrix} n_{1k} \\ \hbar - \sum_{j=1}^{k-1} x_{1j} \end{pmatrix} \begin{pmatrix} n_{2k} \\ m_k - \hbar + \sum_{j=1}^{k-1} x_{1j} \end{pmatrix} \prod_{j=1}^{k-1} \begin{pmatrix} n_{1j} \\ x_{1j} \end{pmatrix} \begin{pmatrix} n_{2j} \\ m_j - x_{1j} \end{pmatrix} \theta_j^{x_{1j}}$$

A distribuição (2.a.12) permite obter regiões de confiança dos parâmetros θ_j, que são as razões das "RPC" verdadeiras, como mostra (2.a.13). Analogamente à secção anterior, consideram-se os pontos X_{1j} que tem densidade máxima em $\tilde{X}_{1j}$, o que, numa amostra grande, leva a:

$$\frac{\tilde{X}_{1j}\,(n_{2j}-m_j+\tilde{X}_{1j})(n_{1k}-\hbar+\sum_{j=1}^{k-1}\tilde{X}_{1j})(m_k-\hbar+\sum_{j=1}^{k-1}\tilde{X}_{1j})}{(n_{1j}-\tilde{X}_{1j})(m_j-\tilde{X}_{1j})(\hbar-\sum_{j=1}^{k-1}\tilde{X}_{1j})(n_{2k}-m_k+\hbar-\sum_{j=1}^{k-1}\tilde{X}_{1j})} = \theta_j \tag{2.a.14}$$

A região de confiança da amostra grande para $\tilde{X}_{1j}$ é obtida por

$$\sum_{j=1}^{k} \frac{(X_{1j}-\tilde{X}_{1j})^2}{\tilde{X}_{1j}} = \chi^2\,(k-1);\alpha \tag{2.a.15}$$

onde

$\chi^2(k-1);\alpha$ é o ponto com probabilidade α (superior) da

distribuição quiquadrado com (k-1) graus de liberdade.

EXEMPLO 2.a.2

Utilizam-se os dados de Dorn (1954), citados em Cornfield (1956), dos 14 estudos retrospectivos para estudo de associação entre o hábito de fumar e câncer de pulmão, como ilustra a TABELA 2.a.E. da página seguinte.

Neste caso, a variável X_{1j} representa o número de indivíduos do grupo de casos sem o fator, e X_{2j}, o número de indivíduos do grupo controle sem o fator, no estudo j, onde $j = 1,2,\ldots,14$.

Nota-se que a *"razão dos produtos cruzados"* mínima é 1,21 e a máxima é 39,73, diferindo em mais de trinta vezes.

Para o computo dos limites de confiança a 95% faz-se $X_{1j} = \tilde{X}_{1j}$ para $j \neq 7$ e usando (2.a.15):

$$(\tilde{X}_{1j}-5)^2 \left[\frac{1}{\tilde{X}_{1j}} + \frac{1}{17-\tilde{X}_{1j}} + \ldots + \frac{1}{719+\tilde{X}_{1j}} + \frac{1}{251-\tilde{X}_{1j}} \right] = 22,36$$

para $j = 7$. A menor raiz desta equação é $\tilde{X}_{17} = 0,85$, de modo que, a partir de (2.b.14), estima-se o limite superior de θ_7 que é 1,71. Embora anteriormente parecesse diferir de mais de trinta vezes, nota-se porém que, na realidade, essa diferença é de, no máximo, 71%. Para os estudos 4 e 6, a diferença é superior a 7 vezes quando calculadas as "RPC", porém, ao se obter a menor raiz, $\tilde{X}_{14} = 4,05$ e por extensão o

TABELA 2.a.E Quatorze estudos para a análise da influência do fumo no câncer de pulmão

Número do Estudo	Pacientes com Câncer de Pulmão		Pacientes Controle		"Razão dos Produtos Cruzados"
	Total	Não Fumantes	Total	Não Fumantes	
1	86	3	86	14	5,38
2	93	3	270	43	5,68
3	136	7	100	19	4,32
4	82	12	522	125	1,84
5	444	32	430	131	5,64
6	605	8	780	114	12,77
7	93	5	186	12	1,21
8	1.357	7	1.357	61	9,08
9	63	3	133	27	5,09
10	477	18	615	81	3,87
11	728	4	300	54	39,73
12	518	19	518	56	3,18
13	490	39	2.365	636	4,25
14	265	5	287	28	5,62
Total	5.437	165	7.949	1.401	6,84

Fonte: Dorn (1954)

θ_4=1,08, a um nível α=0,05, a diferença se reduz a apenas 8%, chegando-se à conclusão de que esses estudos praticamente não diferem.

Para os demais, os limites sobre θ_j incluem a unidade, concluindo-se que, apesar do θ_j superior diferir de três vezes do inferior, dez dos quatorze estudos caso-controle estariam fornecendo as mesmas estimativas do risco relativo. Assim, ao combinar estes dez estudos, obtém-se 136 não fumantes entre 3.929 pacientes de câncer pulmonar, ao passo que nos 6.161 pacientes controle, 1.096 não são fumantes.

A estimação do intervalo para o risco relativo a 95% pode ser conseguida utilizando (2.a.8) e (2.a.9):

$$\tilde{X}_1 = 116,4 \quad e \quad \tilde{X}_2 = 158,3$$

e de (2.a.6), obtém-se os limites para o risco de não fumantes terem câncer de pulmão, relativamente a fumantes:

$$0,1381 \leqslant \Omega^* \leqslant 0,1989$$

Consequentemente, tem-se o seguinte intervalo de variação para o risco de câncer de pulmão dos fumantes relativamente aos não fumantes:

$$5{,}03 \leqslant \Omega \leqslant 7{,}24$$

As conclusões acerca dos dados analisados na TABELA 2.a.E , são:

(i) estudos 7 e 11 não são amostras da mesma população;

(ii) estudos 4 e 6 não são amostras da mesma população;

(iii) todos os demais estudos podem ser amostras da mesma população; e

(iv) estes estudos restantes indicam um risco de ter câncer de pulmão entre os fumantes relativamente aos não fumantes de não menos que 5,03 e não mais de 7,24. A probabilidade do resultado conjunto estar correto é de no máximo 0,903, isto é, $(0{,}95)^2$.

2.a.III. Estudos com mais de dois níveis do Fator e da Doença

Em muitas situações, deseja-se classificar os fatores não somente em presença e ausência, e os efeitos em doentes e controles, mas sim, estratificar os fatores em diversos níveis e os doentes em vários graus de gravidade ou tipos de doença. Os dados nesta forma de abordagem podem ser apresentados como na TABELA 2.a.F da página seguinte.

O propósito aqui é obter regiões de confiança para os riscos a serem considerados, com probabilidade 1-α. Para obter esta região, com base na abordagem da TABELA 2.a.F, con

TABELA 2.a.F Estudo do tipo caso-controle com subclassificação do fator e da doença.

Níveis do Fator	Categoria dos doentes						Controles	Total
	1	2	...	j	...	s-1	s	
1	X_{11}	X_{12}	...	X_{1j}	...	$X_{1\ s-1}$	X_{1s}	M_1
2	X_{21}	X_{22}	...	X_{2j}	...	$X_{2\ s-1}$	X_{2s}	M_2
.	.	.		.		.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
i	X_{i1}	X_{i2}	...	X_{ij}	...	$X_{i\ s-1}$	X_{is}	M_i
.	.	.		.		.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
$\hbar$	$X_{\hbar 1}$	$X_{\hbar 2}$	...	$X_{\hbar j}$	...	$X_{\hbar\ s-1}$	$X_{\hbar s}$	$M_{\hbar}$
Total	N_1	N_2	...	N_j	...	N_{s-1}	N_s	T

sidera-se a distribuição de probabilidade não condicional :

$$\frac{\prod_{j=1}^{s} n_j}{\prod_{i=1}^{r} \frac{p_{ij}^{x_{ij}}}{x_{ij}!}}$$

onde

n_j é o tamanho da amostra considerada na população j, com $j = 1,2,\ldots,s$;

x_{ij} é o número de elementos no subgrupo (i,j) com $i = 1,2,\ldots,r$;

p_{ij} é a proporção de elementos naquela subclassificação (i,j);

$$n_j = \sum_{i=1}^{r} x_{ij} \quad e \quad \sum_{i=1}^{r} p_{ij} = 1$$

Consequentemente, a probabilidade condicional das observações para o subconjunto de amostras onde os totais marginais são fixados por:

$$\sum_{j=1}^{s} x_{ij} = m_i \qquad \text{é}$$

$$\mathbb{C} \; \frac{\prod_{j=1}^{s} n_j \,!}{\prod_{i=1}^{r} x_{ij} \,!} \; \prod_{j=1}^{s-1} \prod_{i=1}^{r-1} \Omega_{ij}^{x_{ij}} \qquad (2.a.16)$$

sendo $\Omega_{ij} = p_{ij}\, p_{rs} \,/\, p_{is}\, p_{rj}$

e $\mathbb{C}$ é determinado pela condição da soma de (2.a.16), sobre todas as variações possíveis de X_{ij}, resultando 1.

Como interpretar o valor Ω_{ij}?

Seja o r-ésimo nível do fator em cada amostra tomado como o de não fumantes e, por exemplo, a s-ésima categoria de doença, como controle. Ω_{ij} é o risco dos fumantes do i-ésimo nível terem a j-ésima categoria de doença relativamen

te àqueles do grupo controle e do de não fumantes.

Para estimar Ω_{ij}, considera-se a distribuição limite de (2.a.16), denotada por $P(X_{ij})$ e, como na Secção 2.a.I, denotando os valores de X_{ij} no ponto de máxima densidade por $\tilde{X}_{ij}$, resultará, para grandes amostras:

$$\frac{\tilde{X}_{ij}\left[n_j - \sum_{i=1}^{r-1} m_i + \sum_{i=1}^{r-1}\sum_{j=1}^{s-1} \tilde{X}_{ij}\right]}{\left[m_i - \sum_{j=1}^{s-1} \tilde{X}_{ij}\right]\left[n_j - \sum_{i=1}^{r-1} \tilde{X}_{ij}\right]} = \Omega_{ij} \qquad (2.a.17)$$

Assim, a distribuição limite para (2.b.16) é normal multivariada e a forma quadrática:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(X_{ij}-\tilde{X}_{ij})^2}{\tilde{X}_{ij}} \qquad (2.a.18)$$

tem distribuição quiquadrado com (r-1)(s-1) graus de liberdade. De (2.a.18) tem-se as expressões para determinação da região de confiança:

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(X_{ij}-\tilde{X}_{ij})^2}{\tilde{X}_{ij}} \leq \chi^2_{\alpha} \qquad (2.a.19)$$

e a correspondente região de confiança para Ω_{ij} é obtida através de (2.a.17), pois Ω_{ij} é uma função monotônica de $\tilde{X}_{k\ell}$, para todos i,j,k,ℓ.

EXEMPLO 2.a.3

Cornfield (1956) ilustra esta abordagem através dos dados de Breslow et al. (1954), como mostra a TABELA 2.a.G.

TABELA 2.a.G Hábito de fumar e câncer de pulmão

Hábito de fumar (i= 1,2,3,4)	Com Câncer		Controles	Total
	Adeno-Carcinoma	Carcinoma Epidermóide		
Não fumantes	4	15	56	75
Fumantes				
-cachimbo e charuto somente	2	13	68	83
-cigarros somente	31	298	240	569
-cigarros mais cachimbo e charuto	9	146	154	309
Total	46	472	518	1036

Fonte: Breslow et al. (1954).

Com estes dados, tomando $\alpha = 0,05$ e consequentemente $\chi^2_{(6);0,05} = 12,59$, determinar-se-á a região de confiança por meio de:

$$\frac{(4-\tilde{X}_{11})^2}{\tilde{X}_{11}}+\ldots+\frac{\left[9-(46-\tilde{X}_{11}-\tilde{X}_{21}-\tilde{X}_{31})\right]^2}{46-\tilde{X}_{11}-\tilde{X}_{21}-\tilde{X}_{31}}+\ldots+\frac{\left[56-(75-\tilde{X}_{11}-\tilde{X}_{12})\right]^2}{75-\tilde{X}_{11}-\tilde{X}_{12}}+\ldots+$$

$$+\frac{\left\{154-\left[518-(75-\tilde{X}_{11}-\tilde{X}_{12})-(83-\tilde{X}_{21}-\tilde{X}_{22})-(569-\tilde{X}_{31}-\tilde{X}_{32})\right]\right\}^2}{518-(75-\tilde{X}_{11}-\tilde{X}_{12})-(83-\tilde{X}_{21}-\tilde{X}_{22})-(569-\tilde{X}_{31}-\tilde{X}_{32})} \leq 12,59$$

Por exemplo, para estimar Ω_{12}, primeiro deve-se obter $\tilde{X}_{12}$ tal que resultem as seguintes equações, usando (2.a.19):

$$(15-\tilde{X}_{12})^2\left[\frac{1}{\tilde{X}_{12}}+\frac{1}{71-\tilde{X}_{12}}+\frac{1}{161-\tilde{X}_{12}}+\frac{1}{139+\tilde{X}_{12}}\right] = 12,59$$

e $X_{ij} = \tilde{X}_{ij}$ para $ij \neq 12$, $i = 1,2,3,4$ e $j = 1,2,3$.

A maior e a menor raiz da equação acima são, respectivamente, 29,2 e 6,49 e, por correspondência, os limites para Ω_{12} são 1,33 e 10,2, usando (2.a.17).

Os resultados das onze afirmações foram tabelados por Cornfield (1956) como destaca a TABELA 2.a.H, que será apresentada para fins ilustrativos.

Algumas afirmações acerca da TABELA 2.a.H da página seguinte, poderão ser feitas, sabendo-se que estão sujeitas a uma probabilidade de cometer um erro, no máximo, de 0,05:

(i) Linha 1: os fumantes tem risco maior que os não fumantes

(ii) Linha 4: não há evidência de que os fumantes de cachimbo ou charutos tenham um risco excessivo em relação aos não fumantes.

(iii) Linha 6: os fumantes de cigarro tem um risco maior que os de cachimbo ou charutos.

(iv) Linha 7: os fumantes de cigarro tem um risco ao menos 2,5 vezes e não mais que 11,9 vezes maior que os não fumantes de cigarro.

TABELA 2.a.H - Limites simultâneos de confiança a 95% do

RISCO DE	DESENVOLVER	RELATIVO AO RISCO DE	DESENVOLVER	SÃO
1. Qualquer fumante	Carc. epidermóide	Um não fumante	Carc. epidermóide	1,33 a 10,2
2. Qualquer fumante	Adenocarcinoma	Um não fumante	Adenocarcinoma	0,22 a 7,1
3. Qualquer fumante	Adenocarcinoma	Qualquer fumante	Carc. epidermóide	0,05 a 2,1
4. Fumante de charuto ou cachimbo	Carc.epidermóide	Um não fumante	Carc. epidermóide	0,17 a 2,9
5. Fumante só de cigarro	Carc. epidermóide	Um não fumante	Carc. epidermóide	1,64 a 13,0
6. Fumante só de cigarro	Carc. epidermóide	Fum.de cachimbo ou charuto	Carc. epidermóide	2,19 a 18,0
7. Fumante só de cigarro	Carc. epidermóide	Fum.que não seja de cigarro	Carc. epidermóide	2,49 a 11,9
8. Fumante só de cigarro	Adenocarcinoma	Fum.que não seja de cigarro	Adenocarcinoma	0,59 a 12,1
9. Fumante só de cigarro	Adenocarcinoma	Não fumante (só de cigarro	Adenocarcinoma	0,79 a 7,2
10 Fumante de cigarros e outros	Carc. epidermóide	Fum. somente de cigarro	Carc. epidermóide	0,45 a 1,3
11. Fumante de cigarros e outros	Carc. epidermóide ou adenocarcinoma	Fum.que não seja de cigarro	Carc. epidermóide ou adenocarcinoma	2,39 a 10,4

FONTE: Cornfield (1956).

Com estes enfoques procurou-se dar uma idéia das técnicas que tem sido desenvolvidas para estimar o grau de associação por intervalo, porém a questão não está fechada, visto haver outros esquemas de abordagem do assunto.

A estimação por intervalo, pelo fato de fornecer uma informação aproximada do risco relativo, de amplitude de variação ampla, às vezes, desperta uma curiosidade natural de quanto deverá ser o valor do grau de associação, em termos pontuais. Isso será apresentado no tópico seguinte.

2.b. ESTIMATIVA POR PONTO

Como a literatura acerca da estimação do risco relativo por ponto é bastante vasta, procurar-se-á destacar neste item apenas alguns aspectos.

2.b.1. TABELA 2x2

Em um estudo epidemiológico do tipo caso-controle no qual as informações obtidas podem ser apresentadas numa tabela 2x2 simples, o estimador por ponto do risco relativo é o mesmo citado no capítulo zero (0.b.19), isto é, o estimador da "*razão dos produtos cruzados*":

$$\hat{\Omega} = \frac{AD}{BC}$$

2.b.II. Várias TABELAS 2x2

Se for suposto que os riscos relativos são constantes em todas as subclassificações - uma suposição pouco sustentável na prática - poder-se-ia obter *o risco relativo global* ("*over-all relative risk*") através de ponderações de várias estimativas, de acordo com suas respectivas precisões. Pode-se utilizar outro critério: o de ponderações por "importância". Justifica-se esta afirmação, pois:

(i) um duplo aumento num risco relativo grande é "mais importante" que um duplo aumento num pequeno risco; ou

(ii) um risco aumentado para grandes grupos é "mais importante" que para um pequeno grupo; ou

(iii) um risco aumentado para mais jovens pode ser "mais importante" que para mais velhos com uma menor expectativa de vida, ou qualquer outra suposição deste tipo.

São evidentes as dificuldades que surgirão nestas ponderações, por "importância", pois dependem dos conceitos, dos pontos de vista e objetivos de cada abordagem.

Há várias outras questões que surgem, tais como:

(i) precisam-se ajustar os riscos de doença entre as pessoas com o fator, à distribuição da população sem o fator, ou, vice-versa?

(ii) precisam-se ajustar os riscos para as populações com e sem o fator, a uma população padrão combinada?

Responder ou tentar responder satisfatoriamente a estas questões, levaria a várias expressões de comparações e certamente forneceria diferentes respostas. Na prática, é usual encontrar riscos relativos obtidos não em termos de ponderação por precisão, nem por importância, mas sim em termos de ajuste por subcategorias, como é o caso de taxas de mortalidade ajustadas por idade, utilizando os métodos de padronização direta ou indireta, onde o padrão de referência é a distribuição de freqüência ou ainda as taxas correspondentes à amostra de pessoas doentes, de controles, ou de pessoas doentes e de controles combinados.

Mantel e Haenszel(1959) sugeriram uma fórmula para estimativa do risco relativo global:

$$\hat{R}_o = \sum_{j=1}^{k} \left[A_j D_j / T_j \right] \Big/ \sum_{j=1}^{k} \left[B_j C_j / T_j \right] \qquad (2.b.1)$$

A expressão (2.b.1) é provida de uma ponderação da ordem de $N_{1j}\ N_{2j} / (N_{1j}+N_{2j})$ $j = 1,2,3,\ldots,k$, que se aproxima à ponderação, por precisão, do risco relativo para cada sub - classificação, e também será razoável interpretá-la como uma ponderação por "importância". Nota-se uma propriedade bastante particular: $\hat{R}_o$ é igual à unidade somente quando:

$$\sum_{j=1}^{k} A_j = \sum_{j=1}^{k} E\left[A_j\right] \qquad (2.b.2)$$

isto é, terá o quiquadrado correspondente nulo. Vê-se, pois que $A_j - E\left[A_j\right] = \left[A_j D_j - B_j C_j\right] T_j$ para $j = 1,2,\ldots,k$, e quando (2.b.2) se verifica, resulta que:

$$\sum_{j=1}^{k}\left[A_j D_j\right] = \sum_{j=1}^{k}\left[B_j C_j / T_j\right]$$, e conseqüente -

mente $\chi^2 = 0$, que equivale a tomar $\hat{R}_o = 1$.

Uma outra fórmula sugerida por Mantel-Haenszel(1959) para a estimativa do *risco relativo global* é:

$$\hat{R}_1 = \frac{\sum_{j=1}^{k} A_j \sum_{j=1}^{k} D_j}{\sum_{j=1}^{k} B_j \sum_{j=1}^{k} C_j} \Bigg/ \frac{\sum_{j=1}^{k} E\left[A_j\right] \sum_{j=1}^{k} E\left[D_j\right]}{\sum_{j=1}^{k} E\left[B_j\right] \sum_{j=1}^{k} E\left[C_j\right]} \qquad (2.b.3)$$

onde:

$$\begin{aligned} E\left[A_j\right] &= N_{1j} M_{1j}/T_j \\ E\left[B_j\right] &= N_{1j} M_{2j}/T_j \\ E\left[C_j\right] &= N_{2j} M_{1j}/T_j \\ E\left[D_j\right] &= N_{2j} M_{2j}/T_j \end{aligned} \qquad j = 1,2,\ldots k.$$

Verifica-se que $\hat{R}_1 = 1$ quando $\sum_{j=1}^{k} A_j = \sum_{j=1}^{k} E\left[A_j\right]$, $\sum_{j=1}^{k} B_j = \sum_{j=1}^{k} E\left[B_j\right]$;

$\sum_{j=1}^{k} C_j = \sum_{j=1}^{k} E\left[C_j\right]$ e $\sum_{j=1}^{k} D_j = \sum_{j=1}^{k} E\left[D_j\right]$. Portanto, testa-se por meio da distribuição quiquadrado a um grau de liberdade.

O numerador de (2.b.3) representa o valor bruto do risco relativo, ignorando todas as subclassificações por outros fatores, ao passo que o denominador é o valor bruto para risco relativo que teria resultado de um arranjo dos dados quando todos os riscos relativos dentro de cada subclassificação forem exatamente a unidade. O procedimento acima tem analogia com o método "indireto" de padronização (no caso de taxas de mortalidade, por exemplo).

Haenszel, Shimkin e Mantel (1958) utilizaram um estimador que depois Mantel e Haenszel (1959) formalizaram como $\hat{R}_1$, apesar de ter poder pequeno, para apresentar resultados bem comparáveis com outros estimadores. É o caso ilustrado pelo cálculo do risco relativo para carcinoma pulmonar epidermóide e carcinoma pulmonar indiferenciado, entre mulheres, associado com o hábito de fumar mais de um maço de cigarros por dia quando comparado com não fumantes. Este caso é ilustrado detalhadamente por Mantel e Haenszel (1959). Nestes cálculos, $\hat{R}_1$ baixou de 7,1 quando controlado por idade, para 5,6 quando controlado por idade e consumo de café, ao passo que com $\hat{R}_o$ os valores correspondentes foram 9,7 e 9,9.

Podem-se citar outros estimadores semelhantes apresentados por Mantel e Haenszel (1959), como é o

$$\hat{R}_2 = \frac{\sum_{j=1}^{k} A_j \sum_{j=1}^{k} D_j \frac{N_{1j}}{N_{2j}}}{\sum_{j=1}^{k} B_j \sum_{j=1}^{k} C_j \frac{N_{1j}}{N_{2j}}} \qquad (2.b.4)$$

que surge quando a distribuição de casos é tomada como padrão à qual os controles são ajustados.

Se, porém, a distribuição dos controles é tomada como padrão, o estimador será:

$$\hat{R}_3 = \frac{\sum_{j=1}^{k} A_j \frac{N_{2j}}{N_{1j}} \sum_{j=1}^{k} D_j}{\sum_{j=1}^{k} B_j \frac{N_{2j}}{N_{1j}} \sum_{j=1}^{k} C_j} \qquad (2.b.5)$$

Agora, se a distribuição combinada é tomada como padrão, o estimador torna-se:

$$\hat{R}_4 = \frac{\sum_{j=1}^{k} A_j \frac{T_j}{N_{1j}} \sum_{j=1}^{k} D_j \frac{T_j}{N_{2j}}}{\sum_{j=1}^{k} B_j \frac{T_j}{N_{1j}} \sum_{j=1}^{k} C_j \frac{T_j}{N_{2j}}} \qquad (2.b.6)$$

Algumas restrições imediatas dos estimadores $\hat{R}_2, \hat{R}_3$

e $\hat{R}_4$:

(i) $\hat{R}_2$ não é definido quando algum N_{2j} assumir valor zero,

(ii) $\hat{R}_3$ não é definido quando algum N_{1j} assumir valor zero,

(iii) $\hat{R}_4$ não é definido quando algum N_{1j} e/ou N_{2j} assumirem valor zero.

Assim, quando isso ocorre, essas subclasses devem ser eliminadas.

No $\hat{R}_1$, (2.b.3), não há este tipo de restrição, porém essas subclasses são mantidas às custas da inevitável tendenciosidade à unidade, ao passo que no $\hat{R}_o$, (2.c.1), às subclasses onde isso ocorrer, é atribuído peso zero. Logo, equivale a eliminá-las para o cálculo.

Observando $\hat{R}_4$, verifica-se que é um estimador ajustado de maneira direta à distribuição combinada como padrão, enquanto que $\hat{R}_2$ e $\hat{R}_3$, podem ser ajustados tanto direta como indiretamente.

2.b.III. Estudo de Amostra Pareada

É fácil notar que neste estudo, com o esquema apresentado pela TABELA 1.c.A, os estimadores do risco relativo: $\hat{R}_1, \hat{R}_2, \hat{R}_3$ e $\hat{R}_4$, apresentados por Mantel e Haenszel (1959) serão expressos por

$$\hat{R}_o = H/G$$

e

$$\hat{R}_1 = \hat{R}_2 = \hat{R}_3 = \hat{R}_4 = AD/BC$$

2.b.IV. Tabela de Contingência 2xk

Um estudo caso-controle que considera vários níveis do fator suspeito será abordado agora. Mesmo nas situações amplas, com vários níveis do fator, o risco relativo de um sobre o outro poderá ser calculado através do uso dos dados pertinentes àqueles dois níveis ou do uso dos dados de todos os níveis considerados para análise.

As estimativas do risco relativo, $\hat{R}_o$, $\hat{R}_1$, $\hat{R}_2$, $\hat{R}_3$ e $\hat{R}_4$, poderão ser calculadas para os dois procedimentos citados, bastando para tal simplesmente saber estabelecer os valores de N_{1j} e N_{2j}, respectivamente, para os casos e controles, com $T_j = N_{1j} + N_{2j}$, $j = 1,2,\ldots,k$.

As relações exatas que se verificam nos casos em que se tem risco relativo do primeiro nível sendo o dobro daquele do segundo nível que, por seu turno, será o dobro daquele do terceiro, resultando que o risco relativo do primeiro será o quádruplo do terceiro, leva a não aceitar o estimador $\hat{R}_o$ como estimador adequado. Isto requer uma formulação mais refinada e abrangente para assegurar a validade da propriedade considerada.

Gart (1962) apresentou outros tipos de estimadores do risco relativo, partindo do esquema de se ter k pares de amostras binomiais mutuamente independentes de parâmetro p_{1j} e p_{2j}, onde $j = 1,2,\ldots,k$, tomando como o número de sucessos nas amostras de tamanho n_{1j} e n_{2j}, respectivamente, A_j e C_j.

A TABELA 1.b.A mostra o j-ésimo par de amostras e a "*razão dos produtos cruzados*" é definida por:

$$\Omega_j = \frac{p_{1j}\ (1-p_{2j})}{p_{2j}\ (1-p_{1j})} \qquad \text{para } j = 1,2,\ldots,k$$

cujo estimador de máxima verossimilhança é equivalente a 1. b.1).

O interesse é o de testar $Ho: \Omega_j = \Omega$ para todos os $j = 1,2,\ldots,k$. Desde que isso se verifique, é natural utilizar, segundo a notação da TABELA 1.b.A,

$$\hat{\Omega}_G = \frac{\sum_{j=1}^{k} A_j \sum_{j=1}^{k} D_j}{\sum_{j=1}^{k} B_j \sum_{j=1}^{k} C_j} \qquad (2.b.\text{-})$$

Considerando todas as informações das k tabelas 2x2. A tendenciosidade deste estimador é ilustrada pelo exemplo hipotético da TABELA 2.b.A, para mostrar que $\hat{\Omega}_1 = \hat{\Omega}_2 = 1$, porém $\hat{\Omega}_G \neq 1$.

TABELA 2.b.A Estudos do tipo caso-controle

Estudo j	Com doença		Livres de doença		Total
	com fator	sem fator	com fator	sem fator	
1	10	10	150	150	320
2	250	50	50	10	360
Total	260	60	200	160	680

Fonte: Gart (1962)

$$\hat{\Omega}_1 = \frac{10x150}{150x10} = 1; \quad \hat{\Omega}_2 = \frac{250x10}{50x50} = 1 \quad \text{e} \quad \hat{\Omega}_G = \frac{260x160}{60x200} = 3,47$$

Pelo uso das fórmulas (2.a.8) e (2.a.9)e, posteriormente, passando por (2.a.6) com 95% de confiança, obtém-se os limites inferiores e superiores do intervalo de confiança: 2,4 e 5,0, que não contém a unidade. É mais um enfoque que reforça a inadequabilidade do estimador (2.b.7). Outro reforço na afirmação de que (2.b.7) não é bom estimador é a análise teórica da consistência do estimador, desenvolvida por Gart (1962). (Ver Apêndice III)

Um método que pode ser aqui utilizado para produzir um estimador consistente e eficiente do *risco relativo global* é o de máxima verossimilhança. Parte inicialmente da função de verossimilhança dada por:

$$V = \prod_{j=1}^{k} p_{1j}^{a_j} \, p_{2j}^{b_j} \, (1-p_{1j})^{c_j} \, (1-p_{2j})^{d_j}$$

e que será maximizada em relação a p_{ij}, sujeita às condições de que $\Omega_j = \Omega$, para todos os $j = 1,2,\ldots,k$ e $i = 1,2$.

Assim, o estimador obtido é:

$$\hat{\Omega}_{MVj} = \hat{p}_{1j}\ (1-\hat{p}_{2j}) \Big/ \hat{p}_{2j}\ (1-\hat{p}_{1j}) \qquad (2.b.8)$$

sendo que os $\hat{p}_{ij}$ são estimadores de máxima verossimilhança de p_{ij}. A expressão da variância assintótica de (2.b.8) é:

$$\text{Var}\left[\hat{\Omega}_{MV}\right] = \Omega^2 W^{-1} \qquad (2.b.9)$$

onde $W = \sum_{j=1}^{k} W_j$ com $W_j^{-1} = \dfrac{1}{n_{1j}p_{1j}(1-p_{1j})} + \dfrac{1}{n_{2j}p_{2j}(1-p_{2j})}$

para $j = 1,2,\ldots,k$. (2.b.10)

Como o método de máxima verossimilhança descrito acima requer solução iterativa das equações de verossimilhança, Gart (1962) apresentou estimadores simples que podem ser obtidos não iterativamente e que são assintoticamente eficientes:

(i) estimador através da *média aritmética*:

$$\Omega_a = \sum_{j=1}^{k} \hat{W}_j\ \hat{\Omega}_j / \hat{W} \qquad (2.b.11)$$

onde

$\hat{W}$ e $\hat{W}_j$, são obtidos, substituindo por proporções amostrais correspondentes em (2.b.10);

(ii) estimador através da *média geométrica*:

$$\ell n\ \hat{\Omega}_g = \hat{W}^{-1} \sum_{j=1}^{k} \hat{W}j\ \ell n\ \hat{\Omega}_j \qquad (2.b.12)$$

onde $\hat{W}$ e $\hat{W}_j$ são obtidos da maneira anteriormente descrita;

(iii) estimador através da *média harmônica*:

$$\hat{\Omega}_h = \hat{W} / \sum_{j=1}^{k} \hat{W}_j\ \hat{\Omega}_j^{-1} \qquad (2.b.13)$$

onde $\hat{W}$ e $\hat{W}_j$ são os já citados.

Todos os três estimadores, $\hat{\Omega}_a$, $\hat{\Omega}_g$ e $\hat{\Omega}_h$, tem sua variância assintótica expressa por (2.b.9).

A ilustração acerca do que foi dito até o momento será através dos dados da TABELA 2.a.E, colhidos por Dorn (1954), já utilizados por Cornfield (1956) e agora por Gart (1962), só que este considerou somente dez dos quatorze estudos utilizados por Cornfield (1956) deixando de incluir na análise os estudos 4,6,7 e 11.

EXEMPLO 2.b.1

(i) estimativa de Ω global, utilizando (2.b.7) é:

$$\hat{\Omega}_G = \frac{3793 \times 1096}{136 \times 5065} = 6,03$$

Observa-se que este valor supera 9 das 10 estimativas individuais obtidas na TABELA 2.a.E. Logo, para este ca

so, este estimador ponderado simples não é apropriado.

Na obtenção dos valores dos estimadores eficientes citados: média aritmética (2.b.11), média geométrica (2.b.12) e média harmônica (2.b.13), utilizam-se os cálculos apresentados pela TABELA 2.b.B da página seguinte, e alguns outros adicionais necessários, resultando assim:

(ii) estimativa através da média aritmética:

$$\hat{\Omega}_a = \frac{518,39}{105,50} = 4,91$$

com o erro padrão estimado:

$$\text{e.p. } \hat{\Omega}_a = \hat{\Omega}_a \, \hat{W}^{-1/2} = 4,91 \; (105,50)^{-1/2} = 0,478$$

(iii) estimativa através da média geométrica:

$$\hat{\Omega}_g = \exp \left\{\frac{164,12}{105,50}\right\} = 4,74$$

com o erro padrão estimado:

$$\text{e.p. } \hat{\Omega}_g = \hat{\Omega}_g \, \hat{W}^{-1/2} = 4,74 \; (105,50)^{-1/2} = 0,461.$$

(iv) estimativa através da média harmônica:

$$\hat{\Omega}_h = \frac{105,50}{23,05285} = 4,57$$

com o erro padrão estimado:

$$\text{e.p. } \hat{\Omega}_h = \hat{\Omega}_h \cdot \hat{W}^{-1/2} = 4,57 (105,50)^{-1/2} = 0,445$$

Para quaisquer dos estimadores eficientes, $\hat{\Omega}$, a equação da estimação por intervalo será:

TABELA 2.b.B Cálculo das Estimativas do *Risco Relativo* através da Média *Aritmética, Geométrica e Harmônica*.

NÚMERO DO ESTUDO	"PESO" ESTIMADO	"RPC" (*)	PRODUTO I	*ln* DA "RPC"	PRODUTO II	INVERSO DA "RPC"	PRODUTO III
(i)	$\hat{w}_i$	$\hat{\Omega}_i$	$\hat{w}_i\,\hat{\Omega}$	$ln\,\hat{\Omega}_i$	$\hat{w}i\ ln\hat{\Omega}_i$	$\hat{\Omega}_i^{-1}$	$\hat{w}_i\,\hat{\Omega}_i^{-1}$
1	2,32	5,38	12,48	1,68	3,90	0,1859	0,431288
2	2,69	5,68	15,28	1,74	4,68	0,1761	0,473709
3	4,64	7,32	33,96	1,99	9,24	0,1366	0,633824
5	22,39	5,64	126,28	1,73	38,73	0,1773	3,969747
8	6,22	9,08	56,48	2,21	13,72	0,1101	0,684822
9	2,52	5,09	12,83	1,63	4,10	0,1965	0,495180
10	13,90	3,87	53,79	1,35	18,81	0,2584	3,591760
12	13,39	3,18	42,58	1,16	15,53	0,3145	4,211155
13	33,32	4,25	141,61	1,45	48,31	0,2353	7,840196
14	4,11	5,62	23,10	1,73	7,10	0,1779	0,731169
Total	105,50	-	518,39	-	164,12	-	23,052850

(*) "*Razão dos Produtos Cruzados*"

$$\hat{W}\,\frac{(\hat{\Omega}-\Omega)^2}{\Omega^2} = \chi^2(1);\alpha \qquad (2.b.14)$$

com $\chi^2(1);\alpha$ sendo o ponto de probabilidade superior α da distribuição quiquadrado com um grau de liberdade.

A equação (2.b.14) é quadrática em Ω, sua menor raiz dá o limite inferior.

$$\hat{\Omega}_1 = \hat{\Omega}/(1+\chi_\alpha \cdot \hat{W}^{-1/2}) \qquad (2.b.15)$$

e sua maior raiz dá o limite superior

$$\hat{\Omega}_2 = \hat{\Omega}/(1-\chi_\alpha \cdot \hat{W}^{-1/2}) \qquad (2.b.16)$$

do intervalo de confiança com coeficiente de confiança assintoticamente igual a $1-\alpha$. Particularmente, para o estimador através da média geométrica $\hat{\Omega}_g$, é possível uma abordagem diferente, tomando $\ell n\hat{\Omega}_g$, em vez de $\hat{\Omega}_g$, como normalmente distribuído, o que leva à equação:

$$\hat{W}(\ell n\hat{\Omega}_g - \ell n\Omega_g)^2 = \chi^2_\alpha \qquad (2.b.17)$$

Obtendo soluções para $\ell n\Omega_g$ e calculando o antilogaritmo leva-nos ao limite inferior:

$$\hat{\Omega}_1 = \hat{\Omega}_g \exp\,(-\chi_\alpha \cdot W^{-1/2}) \qquad (2.b.18)$$

o limite superior:

$$\hat{\Omega}_2 = \hat{\Omega}_g \exp\,(\chi_\alpha \cdot \hat{W}^{-1/2}) \qquad (2.b.19)$$

Com os dados das TABELAS 2.a.E e 2.b.B poderemos calcular estes intervalos para fins de ilustração:

(i) estimando através da média aritmética:

$$4,12 \leqslant \Omega \leqslant 6,07 \qquad \text{com } \alpha = 0,05$$

(ii) estimando através da média geométrica:

(a) pelas fórmulas (2.b.15) e (2.b.16)

$$3,98 \leqslant \Omega \leqslant 5,86 \qquad \text{com } \alpha = 0,05$$

(b) pelas fórmulas (2.b.18) e (2.b.19)

$$3,92 \leqslant \Omega \leqslant 5,74 \qquad \text{com } \alpha = 0,05$$

(iii) estimando através da média harmônica:

$$3,84 \leqslant \Omega \leqslant 5,65 \qquad \text{com } \alpha = 0,05$$

Pelo observado, os estimadores assintoticamente consistentes são bem concordantes um com o outro. Somente um intervalo, o da média aritmética, por sinal o de maior amplitude, contém a estimativa de Ω obtida pelo estimador ponderado $\hat{\Omega}_G$, sendo mais uma razão para justificar a sua não adequabilidade.

2.c. O MODELO CONDICIONAL PARA A OBTENÇÃO DOS RISCOS ALTERNATIVOS

Há várias outras maneiras de se medir o risco de uma doença e aqui será apresentada a abordagem de Sheps (1959), que é o caso comum de comparar o risco experimental entre duas populações em observação, através de amostras aleatórias. Ela

se vale de dois estudos para exemplificar o modelo condicional, analisando a influência do fumo no câncer de pulmão e a eficiência da vacinação na poliomielite.

2.c.1 Comparação do Risco por Diferença e por Quociente

Inicialmente, tomam-se as variações dos riscos, definidas por:

(i) $\Delta(p) = p_2 - p_1$ (2.c.1)

e (ii) $\rho(p) = p_2/p_1$ (2.c.2)

onde p_j é o risco experimental por elemento na j=ésima população, sendo j = 1,2, nesta abordagem particular.

A estimativa de p_j é dada pelo quociente $\hat{p}_j = X_j/n_j$, com j=1,2, onde X_j é o número de ocorrências observadas do fator sob investigação em n_j elementos da amostra da população j.

Sob este modelo experimental, qual das comparações, (2.c.1) ou (2.c.2), é a mais indicada?

A fim de visualizar melhor o problema, considera-se o exemplo:

Seja p_j, taxa de mortalidade de uma doença e conseqüentemente q_j taxa de sobrevivência da mesma doença. Nas comparações das taxas, a escolha entre utilizar as de mortalidade ou as de sobrevivência é arbitrária. Percebe-se que, ao utilizar a diferença, (2.c.1), ocorre:

$$\Delta(q) = q_2 - q_1 = (1-p_2) - (1-p_1) = p_1 - p_2 = -\Delta(p)$$

isto é, a única influência que esta medida sofre ao se escolherem as celas é o sinal, sem afetar no valor absoluto, ao passo que ao se utilizar a fórmula da razão (2.c.2) ocorrem situações não prognosticáveis, visto que, se:

$$\rho(p) = p_2/p_1 = \theta \quad , \text{ ou seja, } p_2 = \theta p_1$$

então $\rho(q) = q_2/q_1 = (1-p_2)/(1-p_1) = (1-\theta p_1)/(1-p_1)$, que é dependente de p_1 (ou q_1).

Apesar de (2.c.2) fornecer valores do tipo "dobro", "metade" e outros, terá algumas desvantagens em relação a (2.c.1), conforme pode ser visto pelo exemplo numérico no QUADRO 2.c.A.

QUADRO 2.c.A Quadro ilustrativo das situações não prognosticáveis, quando do uso de $\rho(p)$ e $\rho(q)$.

p_2	0,02	0,20	0,60	0,85	0,25	0,70
p_1	0,01	0,10	0,30	0,95	0,75	0,90
$\rho(p)$	2,00	2,00	2,00	0,90	0,33	0,78
$\rho(q)$	0,99	0,89	0,57	3,00	3,00	3,00

Fonte: Sheps (1959)

Assim, a escolha da cela influi decididamente no valor da razão das duas taxas (2.c.2) utilizadas para medir o "*risco relativo*" entre duas populações observadas.

A diferença entre duas taxas (2.c.1) é facilmente interpretada como uma estimativa do número adicional de elementos afetados e, para se testar essa discrepância ou estimar o intervalo de confiança, é só utilizar o procedimento bastante conhecido da comparação de duas proporções, sob hipótese nula de $p_1 = p_2$. É considerado razoável utilizar o modelo de abordagem descrito acima se os riscos individuais p_j, experimentados por várias populações, são mutuamente independentes, como é o caso de tomarmos populações essencialmente diferentes quanto à incidência de alguma doença nos dois sexos, ou quanto às pessoas que diferem geneticamente, ou submetidas a tratamentos médicos diferentes. Mas, quando se trata de categorias não mutuamente independentes, das quais o hábito de fumar é um exemplo, aconselha-se outra abordagem, tal como a do modelo condicional que descreveremos a seguir.

2.c.11. O Modelo Condicional

Este modelo condicional será desenvolvido através de uma ilustração. Considere-se a população de fumantes que inicialmente eram não fumantes e, pelo fato de terem adquirido o hábito de fumar, presumivelmente tiveram alterado o risco de câncer de pulmão. Assim, sem nunca haver fumado um cigarro, alguns (digamos, p_o) morrerão de câncer de pulmão. Por outro lado, aqueles que teriam escapado do câncer de pulmão, se vierem a ser grandes fumantes, alguns deles (digamos, p_f) deverão ser somados às mortes de câncer de pulmão. Isso signi

fica que, entre n_f fumantes, $p_o n_f$ são esperados morrerem de câncer de pulmão, independentemente do fumo, e p_f dos restantes $(1-p_o)n_f$ também morrerão de câncer de pulmão. Sendo assim, a proporção X_f/n_f de fumantes morrerem de câncer de pulmão é o resultado de duas proporções:

$$E\left[X_f/n_f\right] = p_o + (1-p_o)p_f$$

A TABELA 2.c.A da página seguinte resume os valores esperados de fumantes e não fumantes morrerem de câncer de pulmão.

A chance que não fumantes tem de escapar do câncer de pulmão num ano é q_o, enquanto que, para os fumantes, essa chance fica diminuída de um fator $q_f = 1-p_f$. Sob este modelo, o *"Risco Relativo"*, definido como a *razão de duas taxas de mortalidade* observadas D_f/D_o, terá um valor esperado:

$$E\left[D_f/D_o\right] = (1-q_o q_f)/p_o = (p_o+q_o-q_o q_f)/p_o = 1+\frac{q_o p_f}{p_o} \qquad (2.c.3)$$

que sempre será, no mínimo, igual à unidade e é chamado *razão do tipo A*. Evidentemente, é possível estimar o valor da *razão de taxas de sobrevida*:

$$E\left[\frac{(1-D_f)}{(1-D_o)}\right] = q_o q_f/(1-p_o) = q_o q_f/q_o = q_f \qquad (2.c.4)$$

que sempre será, no máximo, igual à unidade, e é chamado de *razão do tipo B*. Tem-se também o valor esperado da diferença:

TABELA 2.c.A - Taxas de mortalidade esperadas de câncer de pulmão entre fumantes e não fumantes.

	Não Fumantes	Fumantes
Nº de homens que morrerão de câncer de pulmão	$E[X_o] = p_o n_o$	$E[X_f] = \{p_o+(1-p_o)p_f\}n_f = (1-q_o q_f)n_f$
Nº de homens que não morrerão de câncer de pulmão	$E[n_o-X_o] = (1-p_o)n_o = q_o n_o$	$E[n_f-X_f] = (1-p_o)(1-p_f)n_f = q_o q_f n_f$
TOTAL	n_o	n_f
Taxa de mortalidade	$E[D_o] = E[X_o/n_o] = p_o$	$E[D_f] = E[X_f/n_f] = 1-q_o q_f$

Fonte: Sheps (1959)

$$E\left[D_f - D_o\right] = 1 - q_o q_f - p_o = q_o - q_o q_f = q_o p_f \qquad (2.c.5)$$

O reverso desta abordagem, que é identificado pelo processo biológico de imunidade ou proteção, é exemplificado pelo processo de vacinação contra poliomielite.

Tomam-se, a princípio, duas amostras de crianças de uma mesma população, tanto as que foram vacinadas quanto as que receberam placebo.

Neste caso, se uma proporção p_1 das crianças não vacinadas (receberam placebo) contrair poliomielite e, consequentemente, $1-p_1$ não contrair, presume-se que a vacinação não afetará o estado de $1-p_1$ crianças. Ao contrário, a vacinação protegerá uma fração das crianças que teria a doença, na ordem de $1-p_2$, visto que diminuiria o número de casos de poliomielite de um fator p_2.

Nas campanhas de vacinação procura-se determinar o grau de proteção que a vacina confere à população. Isto poderá ser obtido através da *eficiência da vacina:*

$$Ef\ (vac) = \left[1-(I_2/I_1)\right] x100\% \qquad (2.c.6)$$

onde I_1 e I_2 são as taxas de incidência em não vacinados e vacinados, respectivamente. A TABELA 2.c.B, mostrando as composições das taxas esperadas, ilustra esse caso:

TABELA 2.c.B

Crianças	Não vacinadas	Vacinadas
sem poliomielite	$E\left[X_1\right] = p_1 n_1$	$E\left[X_2\right] = p_1 p_2 n_2$
com poliomielite	$E\left[n_1 - X_1\right] = (1-p_1)n_1$	$E\left[n_2 - X_2\right] = \left[(1-p_1)+p_1(1-p_2)\right]n_2 = (1-p_1p_2)n_2$
Total	n_1	n_2
Taxa de incidência	$E\left[I_1\right] = p_1$	$E\left[I_2\right] = p_1 p_2$

Fonte: Sheps (1959)

Note-se que a razão das taxas de incidência dá um valor esperado de p_2, ou seja, $E\left[I_2/I_1\right] = p_2$. Assim Ef(vac) = $(1-p_2)$x100%.

Note-se que se entre os fumantes o hábito de fumar aumenta a incidência do câncer de pulmão, o fato de vacinar aumentaria a proteção, isto é, diminuiria a incidência da poliomielite, o que justifica afirmar-se que uma abordagem é o reverso da outra.

É interessante considerar aqui um fato que se pode verificar nos casos de vacinação:

É possível que a vacinação feita provoque a doença numa proporção, por exemplo, p_3 de $(1-p_1)$ das crianças, resultando que a taxa de incidência da poliomielite entre

crianças vacinadas terá um valor esperado de: $(1-p_1)p_3+p_1p_2$. Porém, em muitos experimentos deste tipo, dificilmente os dados proporcionam condições de estimar essa proporção adicional.

Um outro caso, apresentado por Sheps (1959) para ilustrar o modelo citado, é a apresentação dos dados de um ensaio clínico, descrito em Medical Research Council (1948), sobre o uso da estreptomicina no tratamento da tuberculose. Neste estudo, foram comparados pacientes tratados com este antibiótico com pacientes que ficaram apenas acamados.

Acompanhando a evolução dos pacientes, após seis meses, observou-se uma melhora considerável em $\hat{p}_T = 51\%$ dos tratados, mas também $\hat{p}_o = 8\%$ dos não tratados apresentaram melhoras, levando à conclusão de que o antibiótico exercera efeito somente em uma proporção $\hat{q}_o = 1-\hat{p}_o = 92\%$ dos pacientes. Evidentemente, aqui p_o é definida não como um risco, mas ao contrário, como uma probabilidade de melhora. Assim, o tratamento efetuado tende a aumentar essa probabilidade. A taxa de insucesso é expressa por: $q_I = \frac{q_T}{q_o} \times 100\%$. No caso, $\hat{q}_I = 100(1-0,51)/(1-0,08) = 53\%$, e o seu complemento poderá ser utilizado como medida da eficácia do tratamento de tuberculose com estreptomicina: Ef (TRAT.) $= 1-q_I$.

2. c.III. Obtenção do Estimador pelo Método de Máxima Verossimilhança

Sheps (1959) partiu do modelo descrito pela TABELA

2.c.A, supondo que cada amostra é retirada aleatoriamente de suas populações e fixando os tamanhos de amostras n_o e n_f, respectivamente, para populações sem fator e com fator. Os estimadores de p_o, p_f e q_f foram obtidos utilizando o método de máxima verossimilhança. A função de verossimilhança considerada é:

$$V\left[p_o, p_f, x_o, x_f\right] = Pr\{X_o = x_o\} \cdot Pr\{X_f = x_f\} \qquad (2.c.7)$$

onde X_o é a variável aleatória que descreve o número de doentes em n_o elementos da população sem o fator e X_f a variável aleatória do número de doentes em n_f elementos da população com o fator. Considerando que X_o e X_f tem distribuições binomiais independentes, com parâmetros p_o e $1-q_oq_f$ e tamanho de amostra n_o e n_f, respectivamente, por (2.c.7) tem-se

$$V\left[p_o, p_f, x_o, x_f\right] = \binom{n_o}{x_o} p_o^{x_o} (1-p_o)^{n_o-x_o} \binom{n_f}{x_f} (1-q_oq_f)^{x_f} (q_oq_f)^{n_f-x_f} \qquad (2.c.8)$$

Então, o logaritmo da função de verossimilhança, excetuando as constantes, será:

$$L = x_o \ell n p_o + (n_o - x_o) \ell n (1-p_o) + x_f \ell n (1-q_oq_f) + (n_f - x_f) \ell n q_oq_f \qquad (2.c.9)$$

As derivadas parciais de primeira ordem de (2.d.9) são:

$$\frac{\partial L}{\partial p_o} = \frac{x_o}{p_o} - \frac{n_o - x_o}{q_o} + \frac{x_f(1-p_f)}{1-q_oq_f} - \frac{n_f - x_f}{q_o} \qquad (2.c.10)$$

$$\frac{\partial L}{\partial p_f} = \frac{x_f q_o}{1-q_o q_f} - \frac{n_f - x_f}{q_f} \qquad (2.c.11)$$

Igualando a zero as expressões (2.c.10) e (2.c.11), e resolvendo o sistema para p_o e p_f, obtém-se as estimativas desejadas:

$$\hat{p}_o = X_o/n_o \quad \text{e} \quad \hat{p}_f = \frac{X_f/n_f - X_o/n_o}{1-X_o/n_o} \qquad (2.c.12)$$

e portanto $\hat{q}_f = 1-\hat{p}_f = \frac{n_f - X_f}{n_o - X_o} \cdot \frac{n_o}{n_f}$

É fácil notar que $\hat{p}_o$ e $\hat{p}_f$ são correlacionadas e a matriz de variância-covariância, $\underset{\sim}{\Sigma}_p$, é obtida invertendo a matriz de informação, $\underset{\sim}{I}_p$, onde seus elementos são obtidos por:

$$-E\left[\partial^2 L \Big/ \partial p_i \, \partial p_j\right]_{i,j=1,2,\ldots,k} \qquad (2.c.13)$$

Daí, a variância e a covariância dos estimadores serão (Ver APÊNDICE I):

$$\text{Var}\left[\hat{p}_o\right] = p_o q_o/n_o \quad ; \quad \text{Cov}\left[\hat{p}_o \hat{p}_f\right] = -p_o q_f/n_o$$
$$\text{Var}\left[p_f\right] = \frac{q_f}{q_0}\left(\frac{1-q_o q_f}{n_f} + \frac{p_o q_f}{n_o}\right) \qquad (2.c.14)$$

e, das três últimas expressões acima, pode-se calcular o coe

ficiente de correlação:

$$p\hat{p}_o\hat{p}_f = Cov\left[\hat{p}_o\hat{p}_f\right] \Big/ \sqrt{Var\left[\hat{p}_o\right] \cdot Var\left[\hat{p}_f\right]} \qquad (2.c.15)$$

EXEMPLO 2.c.1

Sheps (1959) exemplificou o uso dos estimadores de máxima verossimilhança selecionando alguns dos dados citados por Dawber et al. (1957) no estudo epidemiológico de doenças do coração em Framingham, USA , acompanhado durante 4 anos. Foram observados novos casos de doença arteriosclerótica do coração ("ASHD") em 4,46% dos 717 pacientes nas duas categorias de peso (PRF) mais baixo e, em 11,36% dos 176 pacientes, nas duas categorias de peso mais elevado, como mostra a TABELA 2.c.C. da página seguinte.

Assim, presume-se que a obesidade atuaria somente sobre pessoas que estão na categoria de "PRF" mais elevado.

Estimam-se:

$$\text{(i)}\ \widehat{e.p.}\left[\hat{p}_o\right] = \left(\frac{\hat{p}_o\hat{q}_o}{n_o}\right)^{1/2} = \left(\frac{0,0446 \times 0,9554}{717}\right)^{1/2} = 0,0077 \text{ ou } 0,77\%$$

(ii) Risco associado com obesidade $\hat{p}_f = 1-\hat{q}_f = 7,22\%$

(iii) Erro padrão da taxa associada com obesidade:

$$\widehat{e.p.}\left[\hat{p}_f\right] = \left[\frac{\hat{q}_f}{\hat{q}_o}\left(\frac{1-\hat{q}_o\hat{q}_f}{n_f} + \frac{\hat{p}_o\hat{q}_f}{n_0}\right)\right]^{1/2} \times 100 = 2,61\%$$

TABELA 2.c.C Distribuição da população a risco segundo "peso relativo de Framingham" (PRF) em relação à incidência de "ASHD" em quatro anos de acompanhamento (Homens de 45 a 62 anos).

"PESO RELATIVO DE FRAMINGHAM" (PRF)*	POPULAÇÃO A RISCO	DOENTES NOVOS	TAXA (o/oo)
MENOS DE 100	397	16	40
100 — 112	320	16	50
113 — 119	95	10	105
120 ou +	81	10	123
DESCONHECIDO	5	-	-
TOTAL	898	52	58

Fonte: Dawber et al. (1957)

* "Peso Relativo de Framingham (PRF):

É um índice calculado como a razão (multiplicada por 100) do peso observado de um indivíduo pelo peso mediano para o seu grupo de sexo-altura. (Maiores detalhes - ver referência citada).

(iv) Coeficiente de correlação para os dois estimadores:

$$r_{(\hat{p}_o;\hat{p}_f)} = \frac{\hat{\sigma}_{\hat{p}_o\hat{p}_f}}{\hat{\sigma}_{\hat{p}_o}\hat{\sigma}_{\hat{p}_f}} = -0,289$$

(v) O "*risco relativo*" dos pacientes das duas categorias de "PRF" mais elevado relativamente ao daqueles das categorias de "PRF" mais baixo:

$$\hat{R} = \left[\hat{p}_o + (1-\hat{p}_o)\hat{p}_f\right]/\hat{p}_o = 11,36/4,46 = 2,55$$

(vi) A diferença entre taxas de incidência:

$$\Delta I = \hat{p}_f(1-\hat{p}_o) = 6,90\%$$

Retornando à TABELA 2.c.B para obter as estimativas de um efeito benéfico ou protetor, faz-se a troca adequada de p por q, e resultam os estimadores de máxima verossimilhança:

$$\hat{p}_1 = {}^{X_1}/_{n_1} \qquad \hat{p}_2 = \left({}^{X_2}/_{n_2}\right) \Big/ \left({}^{X_1}/_{n_1}\right) \qquad (2.c.16)$$

$$\text{Var}\left[\hat{p}_1\right] = p_1q_1/n_1 \qquad \text{Cov}\left[\hat{p}_1\hat{p}_2\right] = -q_1p_2/n_1$$

$$\text{Var}\left[\hat{p}_2\right] = \frac{p_2}{p_1}\left(\frac{1-p_1p_2}{n_2} + \frac{q_1p_2}{n_1}\right) \qquad (2.c.17)$$

Pode-se generalizar um pouco o que foi abordado nos tópicos anteriores, mas a escolha de um modelo adequado está intimamente ligada com os efeitos esperados dos fatores

que forem considerados. Descreveremos e analisaremos algumas situações possíveis de serem observadas:

(i) grupo não tratado e outros submetidos a tratamentos diferentes.

Suponha inicialmente que a taxa de mortalidade verdadeira seja p_1 para o grupo de pacientes não tratados e para outros (k-1) grupos com diferentes tratamentos. Suponha também que cada tratamento exerce, por si, efeitos modificantes p_j, $j = 2,\ldots,k$. Assim, as taxas de mortalidade esperadas seriam:

(a) entre pacientes não tratados: p_1

(b) entre pacientes com tratamento i : $p_1 p_i$ $i = 2,3,\ldots k$

(c) entre pacientes com tratamento j : $p_1 p_j$ $j = 2,3,\ldots k$

Note que há uma analogia com o modelo condicional considerado anteriormente, no que se refere às estimativas e variâncias, com uma característica adicional de que os estimadores $\hat{p}_i$ e $\hat{p}_j$ $(i,j \neq 1)$ são correlacionados, sendo:

$$\mathrm{Cov}\left[\hat{p}_i;\hat{p}_j\right] = q_1 p_i p_j / p_1 n_1 \qquad (2.c.18)$$

$$i \neq j = 2,\ldots,k.$$

(ii) várias amostras expostas a graduações progressivas quantificadas do mesmo fator.

Um dos exemplos de ilustração e interpretação fácil é o da categorização do grau de fumantes como não fumantes, fumantes de menos de um maço diário ou de um ou mais maços de

cigarros diários. É razoável assumir sucessivo decréscimo nas taxas de sobrevida conforme grau progressivo de tabagismo. A TABELA 2.c.D ilustra bem o que foi considerado agora.

TABELA 2.c.D Taxas de mortalidade e de sobrevida entre várias categorias de fumantes.

	Não Fumantes	Fumantes Moderados	Grandes Fumantes
Mortalidade taxa anual Doll e Hill (1956)	0,01325 $\hat{p}_{.1}$	0,01492 $\hat{p}_{.2}$	0,01884 $\hat{p}_{.3}$
Taxa de Mortalidade esperada	p_1	$p_1+q_1p_2 = 1-q_1q_2$	$p_1+q_1p_2+q_1q_2p_3 = 1-q_1q_2q_3$
Taxa de sobrevida esperada	q_1	q_1q_2	$q_1q_2q_3$

Fonte: Sheps (1959)

Os estimadores de máxima verossimilhança para os q_j são:

$$\hat{q}_1 = 1-\hat{p}_{.1} \quad ; \quad \hat{q}_2 = (1-\hat{p}_{.2})/\hat{q}_1 \quad ; \quad \hat{q}_3 = (1-\hat{p}_{.3})/\hat{q}_1\hat{q}_2$$

ou, genericamente:

$$\hat{q}_j = (1-\hat{p}_{.j}) \Big/ \prod_{i=o}^{j-1} \hat{q}_i \qquad j = 1,2,\ldots,r$$

onde $\hat{q}_o = 1$ e

$\hat{p}_{.j} = \frac{X_j}{n_j}$ é a taxa de mortalidade total anual da categoria j entre os fumantes.

Valem também: (Ver APÊNDICE II)

$$\text{Var}\left[\hat{q}_1\right] = (1-q_1)q_1/n_1 \quad ; \quad \text{Var}\left[\hat{q}_2\right] = \frac{q_2}{q_1}\left[\frac{q_2(1-q_1)}{n_1} + \frac{1-q_1q_2}{n_2}\right]$$

$$\text{Var}\left[\hat{q}_3\right] = \frac{q_3}{q_1q_2} \quad ; \quad \text{Cov}\left[\hat{q}_1;\hat{q}_3\right] = 0 \qquad (2.c.20)$$

$$\text{Cov}\left[\hat{q}_1;\hat{q}_2\right] = -\frac{(1-q_1)q_2}{n_1} \quad ; \quad \text{Cov}\left[\hat{q}_2;\hat{q}_3\right] = -\frac{(1-q_1q_2)q_3}{n_2q_2}$$

Da TABELA 2.c.D, estimam-se os valores:

$\hat{q}_1 = 0,98675$, $\hat{q}_2 = 0,99831$ e $\hat{q}_3 = 0,99602$, mas seria útil dar o significado dos valores q_j.

Se entre os não fumantes a taxa de sobrevida foi $\hat{q}_1 = 0,98675$, entre os fumantes moderados esta taxa foi reduzida de um fator $\hat{q}_2 = 0,99831$, resultando que a taxa de sobrevida de um elemento desta categoria será de 0,98508.

Se, teoricamente, supusermos que o efeito de fumar é constante, poder-se-ia afirmar que a taxa de sobrevida de um fumante moderado daqui a 20 anos será $(0,99831)^{20}$ da sobrevida de um não fumante, ou seja, será de 96,7%, resultando numa redução de 3,3%. Comparando, nos mesmos termos, a taxa de um grande fumante com os demais seria de $(0,99602)^{20}$

ou 92,3% de um fumante moderado e $(0,99602)^{20} \times (0,99831)^{20}$ ou 89,3% de um não fumante, reduzindo-se aproximadamente de 10,7%.

(iii) um outro modelo, este com um enfoque um pouco diferente, será apresentado aqui para comparação entre grupos não fumantes, fumantes e antigos fumantes, isto é, aqueles que, já tendo fumado uma vez, pararam de fumar.

Definindo as duas primeiras amostras como na TABELA 2.c.A, ao considerarmos a terceira amostra, pode ser esperado que diminua o risco de morrer para os $q_o\ p_f$ antigos fumantes, por um fator, digamos p_a.

Os resultados esperados constam da TABELA 2.c.E.

TABELA 2.c.E Taxas de mortalidade e de sobrevida esperadas entre não fumantes, fumantes e antigos fumantes.

	Não Fumantes	Fumantes	Antigos Fumantes
Taxa de Mortalidade	p_o	$p_o+q_op_f = 1-q_op_f$	$p_o+q_op_fp_a$
Taxa de Sobrevida	q_o	q_op_f	$1-p_oq_op_fp_a = q_o(1-p_fp_a)$

Fonte: Sheps (1959)

Aqui, para estimar p_a e q_a, será preciso obter q_o e q_f pelos estimadores já apresentados e daí obter $\hat{q}_a$:

$$\hat{q}_a = \left[\frac{X_f}{n_f} - \frac{X_a}{n_a}\right] / \hat{q}_o\hat{p}_f \quad \text{onde } \frac{X_a}{n_a} \text{ é o estimador da taxa de mortalidade entre antigos fumantes.}$$

A variância e as covariâncias complementares são:

$$\text{Var}\left[\hat{p}_a\right] = \frac{1}{q_o q_f^2}\left[\frac{(p_o+q_o p_f p_a)(1-p_f p_a)}{n_a} + \frac{q_f(1-q_o q_f)p_a^2}{n_f} + \frac{p_o q_a^2}{n_o}\right] \quad (2.c.21)$$

$$\text{Cov}\left[\hat{p}_o;\hat{p}_a\right] = \frac{-p_o q_a}{p_f n_o} \quad \text{e} \quad \text{Cov}\left[\hat{p}_f;\hat{p}_a\right] = \frac{q_f}{q_o p_f}\left[\frac{p_o q_a}{n_o} - \frac{p_a(1-q_o q_f)}{n_f}\right]$$

Pode-se facilmente encontrar situações diversas destas apresentadas aqui; porém, parece-nos que pela abordagem aqui feita foi possível ter uma visão bem ampla para analisar situações semelhantes, com algumas alterações no modelo, levando em conta a existência das correlações citadas.

É importante destacar que o controle experimental das variáveis intervenientes através do pareamento implica geralmente dificuldades operacionais, em virtude da impossibilidade de se encontrar controles que satisfaçam as exigências do pareamento. Além disso, esse tipo de procedimento não garante um aumento de eficiência, segundo Cochran (1950) e Worcester (1964) citado em Fleiss (1973).

A alternativa de se controlar variáveis intervenientes através da estratificação, por sua vez, não permite estimar o efeito dessas variáveis, além das dificuldades computacionais que surgem quando se cogita do controle de mais de uma variável interveniente.

No capítulo seguinte, discutiremos o modelo *log-li-*

near. Este nos parece mais adequado para o estudo de associações causais na vigência de possíveis associações espúrias, já que não sofre as restrições acima apontadas.

CAPÍTULO 3

ANÁLISE DOS RISCOS RELATIVOS EM k TABELAS 2x2 POR MEIO DO MODELO LOG-LINEAR

O objeto deste capítulo continua sendo o Risco Relativo, definido por Cornfield (1951), que será estimado por meio da "Razão dos Produtos Cruzados" ("RPC"), nos estudos caso-controle. Já se sabe que este é um bom estimador quando a prevalência da doença é baixa e que, ao estimar, para cada subgrupo resultante da estratificação, o seu risco relativo, pode-se também obter uma estimativa do *risco relativo global* ("*over-all relative risk*") através das fórmulas apresentadas por Mantel e Haenszel (1959), caso a hipótese da igualdade destes riscos relativos individuais não seja rejeitada.

Neste capítulo também será descrito um outro tipo de enfoque, mais abrangente, devido a Zelen (1971), que se utiliza das distribuições condicionais e os analisa por meio do Modelo Log-Linear, modelo este que será definido posterior-

mente.

Halperin et al. (1977) fazem comentários acerca de uma estatística proposta por Zelen (1971), discordando da conclusão deste, através de exemplos e desenvolvimentos teóricos, e propõem uma outra estatística em substituição à de Zelen. Essa discussão também está incorporada ao capítulo.

3.0 - O MODELO LOG-LINEAR

Seja a "Razão dos Produtos Cruzados" definida por:

$$\Omega = \frac{p_2(1-p_1)}{p_1(1-p_2)} \qquad (3.0.1)$$

onde

p_2: é a probabilidade de presença do fator em estudo nos casos; e

p_1: é a probabilidade de presença do mesmo fator no grupo controle.

Sobre os p_i $(i=1,2)$, que são probabilidades de presença do fator em estudo no grupo de estudo i, pode-se estudar a transformação logística: [ver Cox (1970)]

$$\lambda_i = \ell n\left(\frac{p_i}{1-p_i}\right) \qquad i = 1,2 \qquad (3.0.2)$$

Logo,

$$p_i = \frac{e^{\lambda_i}}{1+e^{\lambda_i}} \qquad (3.0.3)$$

e

$$i = 1,2$$

$$1-p_i = \frac{1}{1+e^{\lambda_i}}$$

Assim, poderíamos desejar descrever (3.0.2) por meio de um modelo linear, tal que

$$\lambda_i = \ell n\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^{h} a_{ij}\beta_j = \underset{\sim}{a}_i \underset{\sim}{\beta} \qquad (3.0.4)$$

onde $\underset{\sim}{a}_i$ é um vetor-linha de constantes conhecidas e $\underset{\sim}{\beta}$ é um vetor-coluna de parâmetros desconhecidos.

Resulta que o vetor genérico, considerando (3.0.4), é

$$\underset{\sim}{\lambda}_{2x1} = \underset{\sim}{a}_{2xh}\underset{\sim}{\beta}_{hx1} \qquad (3.0.5)$$

Efetuando a transformação sobre a "Razão dos Produtos Cruzados" resulta:

$$\ell n\Omega = \ell n\left(\frac{p_2}{1-p_2}\right) - \ell n\left(\frac{p_1}{1-p_1}\right) = \lambda_2 - \lambda_1 \qquad (3.0.6)$$

Conclui-se, assim, que o risco relativo poderá ser expresso por um modelo linear, quando for utilizada uma transformação logística nos $p_i (i = 1, 2)$. Este é o motivo pelo qual dizemos que a "Razão dos Produtos Cruzados" pode ser expressa por um modelo Logístico Linear, ou simplesmente, por um modelo Log-Linear.

Uma abordagem bem ampla, não somente para dados binários mas para dados categorizados, de um modo geral, foi apresentada por Plackett (1974). O autor analisa a situação em que se considera um fator F com vários níveis, podendo acarretar respostas não somente dicotômicas mas em várias categorias, ou, ainda, o caso de um fator detectado em várias populações, levando a respostas policotômicas. Assim, a probabilidade de obter uma resposta i no j-ésimo nível do fator F é expressa por:

$$Pr\{R = i,\ F = j\} = p_{ij}$$

onde $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, s$, com $\sum_{i=1}^{r} p_{ij} = 1$ para qualquer j.

Definiu-se:

$$\lambda_a = \ell n (p_{as} / p_{rs}),$$

$$\Omega_{ab} = \frac{p_{ab} \cdot p_{rs}}{p_{as} \cdot p_{rb}} \qquad (3.0.7)$$

e

$$\lambda_{ab} = \ell n(p_{ab}p_{rs}/p_{as}p_{rb})$$

onde $a = 1,2,\ldots,r-1$
$b = 1,2,\ldots,s-1$

Para o caso de $r=s=2$, particulariza-se uma tabela 2x2 de respostas dicotômicas. Assim, como as probabilidades p_{ij} $(i,j=1,2)$ são tais que:

$$p_{11}+p_{21} = 1 \text{ e } p_{12}+p_{22} = 1 \quad ,$$

então

$$\Omega = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{22}/(1-p_{22})}{p_{21}/(1-p_{21})} \qquad (3.0.8)$$

A expressão (3.0.8) coincide com (3.0.1). Note-se que a abordagem de Plackett (1974) é mais abrangente que a de Cox (1970), pois dados binários formam um caso particular de dados categorizados.

Para a análise de dados binários que resultam em respostas de duas categorias, por exemplo, "Sucesso" e "Insucesso", utilizam-se as considerações de Cox (1970). Para dar idéia acerca da transformação logística empregada por Cox, redefine-se (3.0.2) como:

$$\lambda_1 = \alpha \text{ e } \lambda_2 = \alpha+\Delta \qquad (3.0.9)$$

Assim, da expressão (3.0.6) resulta que:

$$\ell n\Omega = \lambda_2 - \lambda_1 = \Delta \quad , \qquad (3.0.10)$$

conseqüentemente,

$$\Omega = e^{\Delta} \qquad (3.0.11)$$

Ao analisarmos k tabelas independentes 2x2, comparando dois grupos (1 e 2), poder-se-ia ter um modelo geral, considerando:

$$\lambda_{1j} = \alpha_j \quad e \quad \lambda_{2j} = \alpha_j + \Delta_j \qquad (3.0.12)$$

ou, genericamente:

$$\underset{\sim}{\lambda}_{2xk} = \underset{\sim}{a}_{2x2}\underset{\sim}{\beta}_{2xk}$$

onde

$$\underset{\sim}{\lambda}_{2xk} = \begin{bmatrix} \lambda_{11}\lambda_{12}\cdots\lambda_{1k} \\ \lambda_{21}\lambda_{22}\cdots\lambda_{2k} \end{bmatrix} \; ; \; \underset{\sim}{a}_{2x2} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

e

$$\underset{\sim}{\beta}_{2xk} = \begin{bmatrix} \alpha_1\alpha_2\cdots\alpha_k \\ \Delta_1\Delta_2\cdots\Delta_k \end{bmatrix}$$

Os parâmetros α_j são chamados de parâmetros *de distúrbio* ("*nuisance*") e os Δ_j são os parâmetros de interesse, sen

do que a inferência acerca de Δ_j se obtém através de distribuição condicionada.

Zelen (1971) aborda esse assunto detalhadamente, considerando

$$\alpha_j = \mu + \beta_j$$

e

$$\Delta_j = \alpha + \gamma_j \qquad j = 1,2,\ldots,k,$$

como pode ser visto na secção que se segue.

3.a - DESCRIÇÃO DA ABORDAGEM DE ZELEN PARA k TABELAS 2x2

Há conveniências quanto a considerar a análise de k tabelas $2x2$ como oriundas de resultados das observações de experimentos independentes de Bernoulli, embutidos numa classificação fatorial $2xk$, sendo 2 e k, respectivamente, o número de níveis dos fatores A e B. Assim, cada um dos $2k$ pares é indexado por

$$(i;j) \text{ onde } i = 1,2 \text{ e } j = 1,2,3,\ldots,k.$$

Para a $(i;j)$-ésima combinação há a_{ij} experimentos de Bernoulli que serão denotados por

$$\{y_{ijh}\} \text{ onde } i=1,2\ ,\ j=1,2,3,\ldots,k \text{ e } h=1,2,\ldots a_{ij}$$

sendo

$$y_{ijh} = \begin{cases} 0 \text{ Associado ao "insucesso"} \\ 1 \text{ Associado ao "sucesso"} \end{cases}$$

Além disso, são consideradas:

$m_j = a_{1j}$: número de experimentos a $(1;j)$

$n_j = a_{2j}$: número de experimentos a $(2;j)$

$r_j = \sum_{h=1}^{m_j} y_{1jh}$: número de "sucessos" (por exemplo, doentes no caso de estudos epidemiológicos) a $(1;j)$

$s_j = \sum_{h=1}^{n_j} y_{2jh}$: número de "sucessos" a $(2;j)$

$t_j = r_j + s_j$: número total de "sucessos" ao nível j do fator B.

$t = \sum_{j=1}^{k} t_j$: número total de "sucessos" em $2k$ combinações

$N_j = m_j + n_j$: número total de experimentos ao nível j do fator B.

Percebe-se facilmente que a j-ésima tabela 2x2, correspondente ao nível j do fator B, é como ilustra a TABELA 3.a.A. na página seguinte..

Para a $(i;j)$-ésima combinação do tratamento, definam-se:

(i) A probabilidade de "sucesso":

$$p_{ij} = P_r\{y_{ijh} = 1\} \qquad (3.a.1)$$

TABELA 3.a.A Resultado de estudo tipo Caso-Controle quando considera o nível j do fator B.

NÍVEL j DO FATOR B	CONTROLES (1)	CASOS (2)	TOTAL
EXPOSIÇÃO	r_j	s_j	t_j
NÃO EXPOSIÇÃO	$m_j - r_j$	$n_j - s_j$	$N_j - t_j$
TOTAL	m_j	n_j	N_j

$j = 1, 2, \ldots, k.$

(ii) O logaritmo da razão das probabilidades de "sucesso" e "insucesso":

$$\lambda_{ij} = \ell n \left[p_{ij} / (1 - p_{ij}) \right] \qquad (3.a.2)$$

Se os λ_{ij} assumem os modelos (reparametrizando):

$$\lambda_{1j} = \mu + \beta_j \qquad (3.a.3)$$

para $j = 1, 2, \ldots, k$

$$\lambda_{2j} = \mu + \alpha + \beta_j + \gamma_j \qquad (3.a.4)$$

a partir de (3.a.3) e (3.a.4) será possível descrever a "RPC" como:

$$\Omega_j = \frac{p_{2j}(1-p_{1j})}{p_{1j}(1-p_{2j})} = exp\ (\lambda_{2j}-\lambda_{1j}) = exp\ (\alpha+\gamma_j) \qquad (3.a.5)$$

O Ω_j será constante para cada uma das k tabelas de contingência $2x2$, se $\gamma_j=0$ para todos os $j=1,2,\ldots,k$.

Inicialmente, o problema era de $2k$ parâmetros, isto é, k pares p_{ij} e p_{2j}. Agora, porém, com a reparametrização introduzida em (3.a.3) e (3.a.4), tem-se $2k+2$ parâmetros.

Afirma Zelen (1971) que, quando $\gamma_j=0$ para todos os j, a estatística apropriada para fazer inferência sobre α é

$$S = S_1 + S_2 + \ldots + S_k \qquad (3.a.6)$$

onde S_j é o número de doentes expostos ao nível j do fator B.

Tomar a distribuição de S_j condicionada a $T_j = t_j$ $(j=1,2,\ldots,k)$ é o suficiente para eliminar os outros parâmetros.

Cochran (1954) e Mantel e Haenszel (1959) recomendaram utilizar a distribuição assintótica de S_j, condicionada a $T_j = t_j$ $(j=1,2,\ldots,k)$, quando

$$p_{1j} = p_{2j} \text{ e } \Omega_j = 1 \quad \text{para todos os } j = 1,2,3,\ldots,k.$$

A fim de obter a distribuição citada acima, considera-se inicialmente a distribuição conjunta de R_j e S_j para a j-ésima tabela $2x2$.

$$P_{r}\{R_j=r_j, S_j=s_j\} = \binom{m_j}{r_j}\binom{n_j}{s_j} p_{1j}^{r_j} (1-p_{1j})^{m_j-r_j} p_{2j}^{s_j} (1-p_{2j})^{n_j-s_j}$$

que, com o uso de (3.a.3) e (3.a.4), será equivalente a

$$P_{r}\{R_j=r_j; S_j=s_j\} = C(s_j,t_j) \frac{\exp[r_j(\mu+\beta_j)]}{\{1+\exp(\mu+\beta_j)\}^{m_j}} \frac{\exp[s_j(\mu+\alpha+\beta_j+\gamma_j)]}{\{1+\exp(\mu+\alpha+\beta_j+\gamma_j)\}^{n_j}}$$

ou melhor, (3.a.8)

$$P_{r}\{R_j=r_j; S_j=s_j\} = C(s_j,t_j) \frac{\exp[(\mu+\beta_j)t_j+(\alpha+\gamma_j)s_j]}{\{1+\exp(\mu+\beta_j)\}^{m_j}\{1+\exp(\mu+\alpha+\beta_j+\gamma_j\}^{n_j}}$$

onde (3.a.9)

$$t_j = r_j+s_j \quad e$$

$$C(s_j,t_j) = \binom{m_j}{r_j}\binom{n_j}{s_j} = \binom{m_j}{t_j-s_j}\binom{n_j}{s_j}$$

Para $j=1,2,\ldots,k$. (3.a.10)

De (3.a.9) pode-se obter a distribuição conjunta de

$$\underset{\sim}{R} = (R_1,R_2,\ldots R_k)' \quad e \quad \underset{\sim}{S} = (S_1,S_2,\ldots,S_k)'$$

$$P_{\underset{\sim}{R}}\{\underset{\sim}{R}=\underset{\sim}{r},\underset{\sim}{S}=\underset{\sim}{s}\} \propto \frac{\exp\left[\mu\sum_{j=1}^{k} t_j+\sum_{j=1}^{k}\beta_j t_j+\alpha\sum_{j=1}^{k} s_j+\sum_{j=1}^{k}\gamma_j s_j\right]}{\prod_{j=1}^{k}\left\{\{1+\exp(\mu+\beta_j)\}^{m_j}\{1+\exp(\mu+\alpha+\beta_j+\gamma_j)\}^{n_j}\right\}}$$

(3.a.11)

Conclui-se de (3.a.11) que a estatística suficiente de $(\mu,\{\beta_j\},\alpha,\{\gamma_j\})$ é $(t,\{t_j\},s,\{s_j\})$. Porém, pela definição de $t=\sum_{j=1}^{k} t_j$ e $s=\sum_{j=1}^{k} s_j$, a dimensão do vetor da estatística suficiente é 2k, retornando à dimensão inicial, ou seja, ao número de parâmetros linearmente independentes.

Assim, para o nosso objetivo, considera-se a distribuição das estatísticas suficientes

$$g(\underset{\sim}{s},\underset{\sim}{t}) = P_{\underset{\sim}{R}}\{\underset{\sim}{S}=\underset{\sim}{s},\underset{\sim}{T}=\underset{\sim}{t}\} = \frac{C(\underset{\sim}{s},\underset{\sim}{t})\exp\left[\mu t+\underset{\sim}{\beta}'\underset{\sim}{t}+\alpha s+\underset{\sim}{\gamma}'\underset{\sim}{s}\right]}{\prod_{j=1}^{k}\left\{\left[1+\exp(\mu+\beta_j)\right]^{m_j}\left[1+\exp(\mu+\alpha+\beta_j+\gamma_j)\right]^{n_j}\right\}}$$

(3.a.12)

onde

$$\underset{\sim}{t} = (t_1, t_2, \ldots, t_k)', \quad \underset{\sim}{s} = (s_1, s_2, \ldots, s_k)' \text{ e } C(\underset{\sim}{s}, \underset{\sim}{t}) = \prod_{j=1}^{k} C(s_j, t_j) \tag{3.a.13}$$

3.b - INFERÊNCIA ACERCA DA CONSTÂNCIA DA "RAZÃO DOS PRODUTOS CRUZADOS"

Inicialmente, deseja-se testar se a "Razão dos Produtos Cruzados" é constante para cada uma das k tabelas 2x2 e, assim, com esta suposição, partir para outras inferências, como a da "RPC" ser unitária ou a da probabilidade de sucesso constante para cada nível do fator A, como por exemplo, nos doentes e nos controles, que serão considerados posteriormente.

Como ponto de referência inicial utiliza-se (3.a.12) reescrevendo em termos de distribuições condicionadas

$$f(\underset{\sim}{s}, \underset{\sim}{t}) \equiv f(\underset{\sim}{s}, \underset{\sim}{t}/\mu, \alpha, \underset{\sim}{\beta}, \underset{\sim}{\gamma}) = f(\underset{\sim}{s}/\sum_{j=1}^{k} s_j = s, \underset{\sim}{t}, \underset{\sim}{\gamma})\, f(s, \underset{\sim}{t}/\sum_{j=1}^{k} t_j = t, \alpha, \underset{\sim}{\beta}, \underset{\sim}{\gamma})\, f(t/\mu, \alpha, \underset{\sim}{\beta}, \underset{\sim}{\gamma}) \tag{3.b.1}$$

onde, por exemplo,

$$f(\underset{\sim}{s}/\sum_{j=1}^{k} s_j = s, \underset{\sim}{t}, \underset{\sim}{\gamma}) = P_r\{\underset{\sim}{S} = \underset{\sim}{s}/\sum_{j=1}^{k} s_j = s, \underset{\sim}{T} = \underset{\sim}{t}, \underset{\sim}{\gamma}\} \tag{3.b.2}$$

Similaridade para as outras distribuições condicionadas de (3.b.1):

$$f(s, \underset{\sim}{t}/\sum_{j=1}^{k} t_j = t, \alpha, \underset{\sim}{\beta}, \underset{\sim}{\gamma}) \text{ e } f(t/\mu, \alpha, \underset{\sim}{\beta}, \underset{\sim}{\gamma}).$$

Como (3.b.2) depende só do vetor de parâmetros $\underset{\sim}{\gamma}$, ela será

utilizada para verificar a constância das "RPC" em cada uma das k tabelas; isto é, a hipótese a ser testada será:

$$H_o: \ \gamma_1 = \gamma_2 = \ldots = \gamma_k = 0 \ \ ou \ \ \underset{\sim}{\gamma} = \underset{\sim}{0} \qquad (3.b.3)$$

equivalente a

$$H_o: \ \Omega_j = e^{\alpha} \text{ para qualquer } j = 1,2,\ldots,k. \qquad (3.b.4)$$

Reescrevendo (3.b.2) com maiores detalhes, teremos:

$$f\left(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t}, \underset{\sim}{\gamma}\right) = \frac{f(\underset{\sim}{s}, \underset{\sim}{t})}{\sum_{T} f(\underset{\sim}{s}, \underset{\sim}{t})} = \frac{C(\underset{\sim}{s}, \underset{\sim}{t}) \exp\left[\sum_{j=1}^{k} s_j \gamma_j\right]}{\sum_{T} C(\underset{\sim}{s}, \underset{\sim}{t}) \exp\left[\sum_{j=1}^{k} s_j \gamma_j\right]} \qquad (3.b.5)$$

onde

$$T = \left\{\underset{\sim}{s}: \sum_{j=1}^{k} s_j = s; \ 0 < s_j < \min(t_j, n_j)\right\} \qquad (3.b.6)$$

Agora sob H_o, (3.b.3), a expressão (3.b.5) torna-se

$$f_o\left(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t}\right) = f\left(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t}, \underset{\sim}{\gamma} = \underset{\sim}{0}\right) = \frac{C(\underset{\sim}{s}, \underset{\sim}{t})}{\sum_{T} C(\underset{\sim}{s}, \underset{\sim}{t})} \qquad (3.b.7)$$

Daí, a probabilidade de ocorrência da região extrema associada ao teste de hipótese quando a hipótese alternativa é:

$$H_a: \ \gamma_j \neq 0 \text{ para algum } j = 1,2,\ldots,k.$$

será dada pela expressão

$$P = \sum_{z \in \Theta} \delta_0 \left(\underset{\sim}{z} / \sum_{j=1}^{k} z_j = s, \underset{\sim}{t} \right) \qquad (3.b.8)$$

com

$$\Theta = \left\{ \underset{\sim}{w} = (w_1, w_2, \ldots w_k)' : \delta_0 \left(\underset{\sim}{w} / \sum_{j=1}^{k} w_j = s, \underset{\sim}{t} \right) \leq \delta_0 \left(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t} \right) \right\}$$

(3.b.9)

Zelen (1971) afirma que a distribuição (3.b.5) sendo H_o, (3.b.3) verdadeira, pode ser escrita como

$$\delta_0 \left(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t} \right) = \frac{\prod_{j=1}^{k} \left[C(s_j, t_j) / \binom{N_j}{t_j} \right]}{\sum_{T} \left[C(\underset{\sim}{s}, \underset{\sim}{t}) / \prod_{j=1}^{k} \binom{N_j}{t_j} \right]} = \frac{\delta_0 (\underset{\sim}{s} / \underset{\sim}{t})}{\delta_0 \left(\sum_{j=1}^{k} s_j = s / \underset{\sim}{t} \right)}$$

onde

(3.b.10)

$$\delta_0 (\underset{\sim}{s} / \underset{\sim}{t}) = \prod_{j=1}^{k} \delta_0 (s_j / t_j)$$

e sendo $\delta_0 (s_j / t_j)$ a distribuição de S_j condicionada a $T_j = t_j$

Assumindo que a "RPC" para a j-ésima tabela seja igual a 1 para um número N_j grande, a distribuição $\delta_0 (s_j / t_j)$ será assintoticamente normal com média μ_j e variância σ_j^2, que se obtêm como:

$$\mu_j = E[S_j / T_j = t_j] = t_j n_j / N_j$$

$$\sigma^2_j = Var\left[S_j/T_j=t_j\right] = t_j n_j m_j (N_j - t_j) \Big/ \left[N^2_j (N_j - 1)\right]$$

(3.b.12)

Representando a função densidade de probabilidade da distribuição normal com os parâmetros $(\mu_j;\sigma^2_j)$ por

$$\Phi(s_j/\mu_j,\sigma^2_j),$$

$$P_R\{S_j = s_j/T_j = t_j\} \sim \Phi(s_j/\mu_j,\sigma^2_j) \qquad (3.b.14)$$

Resultará que (3.b.10) pode ser expressa como

$$f_0\left(\underset{\sim}{s}/\sum_{j=1}^{k} s_j = s, \underset{\sim}{t}\right) \sim \frac{\prod_{j=1}^{k} \Phi(s_j/\mu_j,\sigma^2_j)}{\Phi(s/\mu,\sigma^2)} \qquad (3.b.15)$$

onde o denominador representa a função densidade de probabilidade da distribuição normal com

média $$\mu = \sum_{j=1}^{k} \mu_j \qquad (3.b.16)$$

e variância $$\sigma^2 = \sum_{j=1}^{k} \sigma^2_j \qquad (3.b.17)$$

Assim, Zelen (1971) conclui, explicitamente, que a distribuição assintótica é

$$f_0\left(\underset{\sim}{s}/\sum_{j=1}^{k} s_j = s, \underset{\sim}{t}\right) \sim \frac{\prod_{j=1}^{k} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\{-\frac{1}{2}(s_j-\mu_j)^2/\sigma^2_j\}}{\frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(s-\mu)^2/\sigma^2\}} \propto e^{-\frac{1}{2}Q}$$

(3.b.18)

Sendo que:

$$Q = \sum_{j=1}^{k} (s_j - \mu_j)^2/\sigma^2_j - (s-\mu)^2/\sigma^2 = (\underset{\sim}{s}-\underset{\sim}{\mu})' \left[V^{-1} - J/(\underset{\sim}{1}'V\underset{\sim}{1})\right](\underset{\sim}{s}-\underset{\sim}{\mu}) \tag{3.b.19}$$

com

$$\underset{\sim}{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)'$$

$$V = \text{diag}\ (\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_k)$$

$$\underset{\sim}{1} = (1, 1, \ldots, 1)'$$

$$J = \underset{\sim}{1}.\underset{\sim}{1}'$$

Resulta daí que o vetor das médias e a matriz de variância-covariância de $\underset{\sim}{S}$ serão:

$$E\left[\underset{\sim}{S} / \sum_{j=1}^{k} S_j = s, \underset{\sim}{T} = \underset{\sim}{t}\right] = \underset{\sim}{\mu} + V\underset{\sim}{1}\left(\frac{s-\mu}{\sigma^2}\right) = \underset{\sim}{\tau} \tag{3.b.20}$$

e

$$\text{Var}\left|\underset{\sim}{S} / \sum_{j=1}^{k} S_j = s, \underset{\sim}{T} = \underset{\sim}{t}\right| = V - \frac{VJV}{\sigma^2} = \underset{\sim}{\Sigma}$$

Zelen (1971) afirma que Q tem distribuição quiquadrado com $(k-1)$ $g.\ell.$

Este tratamento assintótico apresentado por Zelen

(1971) é questionado por Halperin et al. (1977) que justificam a discordância através da apresentação de dois exemplos hipotéticos, os quais sugerem que o "teste de Zelen" poderá ser incorretamente generalizado. Desenvolvem ainda uma justificativa teórica que demonstra ser a versão assintótica correta da densidade amostral condicional diferente daquela forma sugerida por Zelen, e que a estatística, geralmente, não terá uma distribuição quiquadrado, sendo que o teste associado é viciado e inconsistente.

Acompanhando o raciocínio de Halperin et al. (1977) serão apresentados os exemplos através de tabelas 2x2x2.

EXEMPLO 3.b.1

TABELA 3.b.A (Dados hipotéticos)

TABELA "1"	DOENTES	CONTROLES	TOTAL
EXPOSTOS	200	30	230
NÃO EXPOSTOS	10	40	50
TOTAL	210	70	280

$$\hat{\Omega}_1 = \frac{200 \times 40}{10 \times 30} = 26,67$$

TABELA "2"	DOENTES	CONTROLES	TOTAL
EXPOSTOS	10	30	40
NÃO EXPOSTOS	10	800	810
TOTAL	20	830	850

$$\hat{\Omega}_2 = \frac{10 \times 800}{10 \times 30} = 26,67$$

Fonte: Halperin et al. (1977)

Note-se que $\hat{\Omega}_1 = \hat{\Omega}_2 = 26,67$

Considerando a hipótese de homogeneidade das "RPC" nos dois estudos analisados, pode-se afirmar que, neste caso particular, não há evidências para rejeitá-la.

Um teste condicional exato produzirá uma probabilidade p=1,0 de não rejeitar a hipótese nula, enquanto que a estatística apresentada por Zelen proporciona o valor Q=36,1 que, comparado ao $\chi^2_{(1)0,05}=3,84$, leva à rejeição da hipótese nula.

Observa-se, desta forma, que há uma contradição entre as conclusões obtidas pela abordagem de Zelen e o teste condicional exato.

Vejamos agora o segundo exemplo ilustrativo.

EXEMPLO 3.b.2

TABELA 3.b.B (Dados Hipotéticos)

TABELA "1"	DOENTES	CONTROLES	TOTAL
EXPOSTOS	190	810	1000
NÃO EXPOSTOS	10	990	1000
TOTAL	200	1800	2000

$$\hat{\Omega}_1 = \frac{190 \times 990}{10 \times 810} = 23,2$$

TABELA "2"	DOENTES	CONTROLES	TOTAL
EXPOSTOS	750	250	1000
NÃO EXPOSTOS	250	750	1000
TOTAL	1000	1000	2000

$$\hat{\Omega}_2 = \frac{750 \times 750}{250 \times 250} = 9,0$$

Fonte: Halperin et al. (1977)

Neste caso, as "RPC" são diferentes (23,2 e 9,0).

Estas tabelas conduzirão à probabilidade p= 0,003 do teste condicional exato, enquanto que a estatística de Zelen fornece o valor ZERO, ou seja, Q=0. Isto leva a um resultado também contraditório.

Pelos dois exemplos apresentados, e pela possibilidade de generalizar estes dados para grandes amostras, é fácil ver que o teste assintótico apresentado por Zelen é inconsistente.

Considerando a TABELA 3.a.A novamente, serão desenvolvidos os resultados assintóticos que serão comparados com os de Zelen.

Sejam ainda:

$$s = \sum_{j=1}^{k} s_j$$

$$\underset{\sim}{s} = (s_1, s_2, \ldots, s_k)'$$

$$\underset{\sim}{t} = (t_1, t_2, \ldots, t_k)'$$

$$\underset{\sim}{r} = (r_1, r_2, \ldots, r_k)'$$

e a probabilidade condicional $f(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t})$. Sob a hipótese nula de não interação e condicionada aos totais marginais, f_0 é livre do parâmetro e é exatamente (3.b.7):

$$f_0(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t}) = \prod_{j=1}^{k} C(s_j, t_j) / \sum_{T} \prod_{j=1}^{k} C(s_j, t_j) \qquad (3.b.21)$$

onde T é o conjunto definido por (3.b.6), e $C(s_j, t_j)$ é como foi definida anteriormente em (3.a.10).

Note-se que esta probabilidade pode ser escrita como uma função exclusivamente de $(k-1)$ variáveis aleatórias (v.a.) $S_1, S_2, \ldots S_{k-1}$ e dos totais marginais.

Utilizando-se da transformação definida por

$$Z_j = \frac{S_j - E_j}{\sqrt{s}} \quad com \; j = 1, 2, \ldots, k-1 \qquad (3.b.22)$$

$$E_k = s - \sum_{j=1}^{k-1} E_j \qquad (3.b.23)$$

sendo E_j correspondentes às estimativas não condicionais de máxima verossimilhança das freqüências esperadas nas celas sob a hipótese de não interação e que satisfazem

$$\frac{E_j(m_j - t_j + E_j)}{(t_j - E_j)(n_j - E_j)} = \frac{E_k(m_k - t_k + E_k)}{(t_k - E_k)(n_k - E_k)} \qquad (3.b.24)$$

para todos os $j=1,2,\ldots k-1$,

pode-se obter uma representação assintótica para a distribuição de probabilidade $f_0(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t})$.

As equações (3.b.24) formam um sistema de $(k-1)$ equações cúbicas citadas por Norton (1945). Segundo as afir-

mações de Gart (1970), as razões expressas em (3.b.24) são iguais à estimativa de máxima verossimilhança da "razão dos produtos cruzados" global, Ω.

Tomando as E_j, soluções do sistema de equações citadas, a (3.b.21) terá a representação assintótica

$$\delta_0(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t}) \approx e^{-\frac{1}{2}Q} / \int_{\Xi} e^{-\frac{1}{2}Q} dz \qquad (3.b.25)$$

onde Ξ é o conjunto que define o campo de variação da variável Z

e

$$Q = s\ \underset{\sim}{Z}'\ \underset{\sim}{V}\ \underset{\sim}{Z} \qquad (3.b.26)$$

que é uma forma quadrática de dimensão $(k-1)$. Com dimensão $(k-1)x(k-1)$ temos

$$\underset{\sim}{V} = \underset{\sim}{D} + \underset{\sim}{1}\ \underset{\sim}{1}'\ d_k \qquad (3.b.27)$$

onde

$$\underset{\sim}{1} = (1,1,\ldots,1)'$$

e $\underset{\sim}{D}$ é uma matriz diagonal de elementos d_j, $j = 1,2,3,\ldots,k-1$, definidos como $j=1,2,3,\ldots,k-1$

$$d_j = \left(\frac{1}{E_j} + \frac{1}{t_j - E_j} + \frac{1}{n_j - E_j} + \frac{1}{m_j - t_j + E_j}\right) \qquad (3.b.28)$$

para $j=1,2,\ldots,k$.

Integrando sobre Ξ, resultará

$$f_0(\underset{\sim}{s} / \sum_{j=1}^{k} s_j = s, \underset{\sim}{t}) \sim \frac{\sqrt{s} \cdot e^{-\frac{1}{2}Q}}{(2\pi)^{\frac{k-1}{2}} |V^{-1}|^{\frac{1}{2}}} \quad . \qquad (3.b.29)$$

que é a representação assintótica para probabilidade amostral condicional.

Este resultado já tinha sido obtido por Bartlett (1935) para o caso particular de $k=2$, isto é, caso 2x2x2 e apresentado por Plackett (1974).

Assim, os resultados apresentados acima entram em contradição com as conclusões de Zelen (1971), onde Q foi definido como em (3.b.19):

$$Q = \sum_{j=1}^{k} \left(\frac{s_j - \mu_j}{\sigma_j} \right)^2 - \left(\frac{s - \mu}{\sigma} \right)^2$$

O valor Q expresso por (3.b.26), que é notoriamente diferente da estatística sugerida por Zelen, terá uma distribuição assintótica quiquadrado com $(k-1)$ g.ℓ. sob a hipótese de não interação.

Halperin et al. (1977) mostram em detalhes, através de um exemplo para o caso 2x2x2, que a estatística de Zelen não terá, de modo geral, uma distribuição assintótica quiquadrado com $(k-1)$ g.ℓ. e, por sinal, rejeita a hipótese de não interação com probabilidade assintótica igual a um, sempre que a hipótese é verdadeira, exceto em casos muito

especiais.

A estatística Q de Zelen, para o caso 2x2x2, seria

$$Q = \left(\frac{s_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{s_2-\mu_2}{\sigma_2}\right)^2 - \frac{(s-\mu_1-\mu_2)^2}{\sigma^2_1+\sigma^2_2} \qquad (3.b.30)$$

Se

$$\tau^2_j = \left(\frac{1}{E_j} + \frac{1}{t_j-E_j} + \frac{1}{n_j-E_j} + \frac{1}{m_j-t_j-E_j}\right)^{-1} \qquad (3.b.31)$$

para $j=1,2$.

então

$$U = (s_1-E_1)\left(\frac{1}{\tau^2_1} + \frac{1}{\tau^2_2}\right)^{1/2} \qquad (3.b.32)$$

tem distribuição assintoticamente normal sob a hipótese de não interação.

Como neste caso particular, $s_2 = s-s_1$ e $E_2 = s-E_1$, a (3.b.30) poderá ser expressa como:

$$Q = \Phi^2 \cdot (U+\beta)^2 \qquad (3.b.33)$$

onde

$$\Phi^2 = \left(\frac{1}{\sigma^2_1} + \frac{1}{\sigma^2_2}\right) \Big/ \left(\frac{1}{\tau^2_1} + \frac{1}{\tau^2_2}\right) \qquad (3.b.34)$$

$$\beta = \frac{(E_1-\mu_1)/\sigma^2_1 - (E_2-\mu_2)/\sigma^2_2}{1/\sigma^2_1 + 1/\sigma^2_2} \cdot \sqrt{\frac{1}{\tau^2_1} + \frac{1}{\tau^2_2}} \qquad (3.b.35)$$

Conclui-se que há dois casos a considerar:

O primeiro será assumir os totais marginais tendendo a infinito, a uma mesma ordem, embora as razões permaneçam fixas. A estatística Q de Zelen expressa por (3.b.33) terá uma distribuição assintótica quiquadrado com 1 g.ℓ. (k-1, com k=2), somente quando

$$\beta = o \quad e \quad \Phi^2 = 1$$

pois nestas circunstâncias

$$Q = U^2 = (\delta_1-E_1)^2 \cdot \left(\frac{1}{\tau^2_1} + \frac{1}{\tau^2_2}\right) : \chi^2_{(1)} \qquad (3.b.36)$$

Isso se verifica quando $\hat{\Omega}$, estimador de máxima verossimilhança da "RPC" global, é 1 para todos os tamanhos de amostra tendendo ao infinito. Mesmo que $\hat{\Omega}\neq 1$, pode ocorrer de $\beta=0$, ou seja

$$(E_1-\mu_1)/\sigma^2_1 - (E_2-\mu_2)/\sigma^2_2 = 0$$

o que leva a concluir que a estatística Q tem distribuição

$$\Phi^2 \cdot \chi^2_{(1)}$$

Uma consideração alternativa será quando somente os totais das linhas e colunas são fixados. Neste caso, a distribuição assintótica de Q, expressa por (3.b.33), dependerá de Ω e assumirá a distribuição quiquadrado somente quando $\Omega=1$.

Trataremos agora dos casos, freqüentemente encontrados na prática, em que as tabelas 2x2 descrevem resultados associados com os valores de uma variável interveniente quantitativa, tais como: idade, peso, dose de uma droga, etc.

Para estas situações, um modelo log-linear para a "RPC" que descreveria de um modo adequado os acontecimentos seria:

$$\ell n\Omega_j = \alpha+\gamma \cdot x_j \qquad (3.b.37)$$

para a j-ésima tabela, onde x_j representa o valor da variável interveniente numa certa escala tal que $\sum_{j=1}^{k} x_j = 0$ e γ é um parâmetro desconhecido.

Com essa modificação, (3.b.5) fica sendo

$$f(\underset{\sim}{s}/\sum_{j=1}^{k} s_j=s,\underset{\sim}{t},\gamma) = \frac{C(\underset{\sim}{s},\underset{\sim}{t})\ \exp\left(\gamma \sum_{j=1}^{k} x_j s_j\right)}{\sum_{T} C(\underset{\sim}{s},\underset{\sim}{t})\ \exp\left(\gamma \sum_{j=1}^{k} x_j s_j\right)} \qquad (3.b.38)$$

onde T é o conjunto descrito por (3.b.6).

A distribuição condicional de $Y = \sum_{j=1}^{k} x_j S_j$ será, então,

$$\delta(y/\sum_{j=1}^{k} s_j = s, \underset{\sim}{t}, \gamma) = \frac{C(y, \underset{\sim}{t})\ e^{\gamma y}}{\sum_{y} C(y, \underset{\sim}{t})\ e^{\gamma y}} \qquad (3.b.39)$$

onde

$$C(y, \underset{\sim}{t}) = \sum_{*} C(\underset{\sim}{s}, \underset{\sim}{t}) \qquad (3.b.40)$$

$$* : x_1 s_1 + \ldots + x_k s_k = y$$

Sob a suposição de não interação, isto é, quando $\gamma=0$

$$\delta_0(y/\sum_{j=1}^{k} s_j = s, \underset{\sim}{t}) = C(y, \underset{\sim}{t}) / \sum_{y} C(y, \underset{\sim}{t}) \qquad (3.b.41)$$

que poderá ser utilizada para um teste de hipótese exato.

Zelen (1971) sugere uma distribuição assintótica de

$$Y = \sum_{j=1}^{k} x_j S_j, \quad \text{condicional a} \quad \sum_{j=1}^{k} S_j = s \ \text{ e } \ \underset{\sim}{T} = \underset{\sim}{t}$$

A seguir, são apresentadas as médias, variâncias e covariância de Y e $S = \sum_{j=1}^{k} S_j$ condicionadas a $\underset{\sim}{T} = \underset{\sim}{t}$, que são:

$$E[y/\underset{\sim}{t}] = \sum_{j=1}^{k} x_j \mu_j \cdot Var[Y/\underset{\sim}{t}] = \sum_{j=1}^{k} x_j^2 \sigma_j^2$$

$$E[S/\underset{\sim}{t}] = \sum_{j=1}^{k} \mu_j = \mu \quad Var[S/\underset{\sim}{t}] = \sum_{j=1}^{k} \sigma_j^2 = \sigma^2$$

e finalmente

$$Cov[Y, S/\underset{\sim}{t}] = \sum_{j=1}^{k} x_j \sigma_j^2 \qquad (3.b.42)$$

A distribuição conjunta de Y e S, condicionadas a $\underset{\sim}{T}=\underset{\sim}{t}$ é assintoticamente normal bivariada e, por conseqüência, a distribuição de Y, condicional a $\underset{\sim}{T} = \underset{\sim}{t}$ e $S = \delta$, é também assintoticamente normal com

$$E[Y/S = \delta, \underset{\sim}{T} = \underset{\sim}{t}] = \sum_{j=1}^{k} x_j [\mu_j + \sigma_j^2 (\delta - \mu)/\sigma^2] = M_y \qquad (3.b.43)$$

$$Var[Y/S = \delta, \underset{\sim}{T} = \underset{\sim}{t}] = \sum_{j=1}^{k} \sigma_j^2 \left[x_j - \sum_{j=1}^{k} x_j \sigma_j^2/\sigma^2\right]^2 = V_y \qquad (3.b.44)$$

Note-se que a distribuição da estatística

$$Y = \sum_{j=1}^{k} x_j S_j$$

é tomada condicionalmente, não somente sobre os totais marginais de cada tabela 2x2, mas também sobre uma condição adicional e limitante que é

$$\sum_{j=1}^{k} S_j = \delta$$

Assim, é possível através de uma transformação linear sobre a variável Y, utilizar a distribuição normal padronizada com média zero e variância um, isto é,

$$Z = \frac{Y - M_y}{\sqrt{V_y}}$$

tem distribuição N (0;1).

3.c - INFERÊNCIA ACERCA DA "RAZÃO DOS PRODUTOS CRUZADOS" SOB A CONDIÇÃO DE CONSTÂNCIA EM TODAS AS TABELAS DE CONTINGÊNCIA

Considerando os resultados do tópico anterior, podem-se estudar dois tipos de inferência acerca da "RPC", sujeitos ao conhecimento, "a priori", de que é constante em todas as tabelas de contingência 2x2, por meio de um teste sobre a hipótese:

$$H_o : \gamma_j = 0 \qquad (3.c.1)$$

Para todos os $j = 1, 2, \ldots, k$.

O primeiro seria relativamente à hipótese de que a "RPC" é igual à unidade para todas as tabelas, sob a suposição de "RPC" constante, e o outro seria relativo à hipótese de que as probabilidades de sucesso , na presença do fator

em estudo, sejam iguais em todos os níveis j e simultaneamente, que as probabilidades de sucesso , na ausência do fator, sejam também, iguais em todos os níveis j, onde $j = 1,2,\ldots,k$ sob a suposição de (3.c.1).

3.c.1. Teste Relativo à Unidade da "Razão dos Produtos Cruzados"

Para testar a hipótese de unidade da "RPC", isto é;

$$H_o : \alpha = 0 \text{ ou } \Omega = 1 \qquad (3.c.2)$$

sob a suposição de (3.c.1) verdadeira, deve-se partir da distribuição conjunta de $S = \sum_{j=1}^{k} S_j$ e $\underset{\sim}{T}$, condicionada a $\sum_{j=1}^{k} T_j = t$, expressa por:

$$f(s,t/\sum_{j=1}^{k} t_j = t) = \frac{C(s,\underset{\sim}{t})\ exp\ (\alpha s+\underset{\sim}{\beta}'\underset{\sim}{t})}{\Sigma\ \Sigma\ C(s,\underset{\sim}{t})\ exp(\alpha s+\underset{\sim}{\beta}'\underset{\sim}{t})} \qquad (3.c.3)$$

onde

$$C\ (s,\underset{\sim}{t}) = \sum_{\$} \{\prod_{j=1}^{k} C(s_j;t_j)\} \qquad (3.c.4)$$

e

$$\$ = \{ \sum_{j=1}^{k} s_j = s, 0 \leq s_j < min\ (n_j, t_j), j=1,2,\ldots k\}$$

(3.c.5)

Decompondo (3.c.3) em duas outras funções densidade de probabilidade:

$$f\ (s, \underset{\sim}{t} \Big/ \sum_{j=1}^{k} t_j = t, \mu, \alpha, \underset{\sim}{\beta}) = f(s/\underset{\sim}{t}, \alpha) \cdot f(\underset{\sim}{t}/\sum_{j=1}^{k} t_j = t, \alpha, \underset{\sim}{\beta})$$

onde

$$f(s, \underset{\sim}{t}, \alpha) = C(s, \underset{\sim}{t})\ e^{\alpha s} / \sum_{\$} C(s, \underset{\sim}{t}) \cdot e^{\alpha s} \qquad (3.c.6)$$

A expressão acima, (3.c.6), é o que permite testar (3.c.2), sob a suposição de (3.c.1) verdadeira. Assim o teste exato é efetuado, utilizando-se de

$$p = \sum_{\Theta} C(v, \underset{\sim}{t}) \Big/ \sum_{\$} C(s, \underset{\sim}{t}) \qquad (3.c.7)$$

onde

$$\Theta = \{v : f_0(v/\underset{\sim}{t}) \leq f_0(s/\underset{\sim}{t})\} \qquad (3.c.8)$$

Uma forma assintótica também foi derivada por Zelen (1971) que afirmou ser uma aproximação normal para a distribuição condicional $S = \sum_{j=1}^{k} S_j$, quando $\alpha = 0$. Para desenvolvê-la, será conveniente considerar S como uma variável aleatória com distribuição aproximadamente normal com

$$E[S] = \sum_{j=1}^{k} \mu_j = \mu \qquad (3.c.9)$$

e

$$Var[S] = \sum_{j=1}^{k} \sigma_j^2 = \sigma^2 \qquad (3.c.10)$$

3.c.II. Teste Relativo à Constância da Probabilidade de Sucesso na Presença e na Ausência do Fator em todas as Tabelas de Contingência.

Aqui, parte-se também da suposição (3.c.1) verdadeira e vai-se em busca do instrumental que permite testar:

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \qquad (3.c.11)$$

que é equivalente a

$$H_o : p_{1j} = p_1 \text{ e } p_{2j} = p_2 \qquad (3.c.12)$$

para todos $j = 1,2,\ldots,k$.

A fim de torná-lo exequível, considera-se a distribuição:

$$f(s, \underset{\sim}{t}/\mu, \alpha, \underset{\sim}{\beta}) = f(\underset{\sim}{t}/\sum_{j=1}^{k} t_j = t, s, \underset{\sim}{\beta}) \cdot f(t, s/\mu, \alpha, \underset{\sim}{\beta}) \qquad (3.c.13)$$

onde a função decomposta:

$$\delta(\underset{\sim}{t} / \sum_{j=1}^{k} t_j = t, s, \underset{\sim}{\beta}) = \frac{C(s, \underset{\sim}{t})\, e^{\underset{\sim}{\beta}' \underset{\sim}{t}}}{\sum_{\Gamma} C(s, \underset{\sim}{t})\, e^{\underset{\sim}{\beta}' \underset{\sim}{t}}} \qquad (3.c.14)$$

$$C(s, \underset{\sim}{t}) = \sum_{T} \prod_{j=1}^{k} C(s_j, t_j) \qquad (3.c.15)$$

$$\Gamma = \{\underset{\sim}{t} : \sum_{j=1}^{k} t_j = t,\ 0 \leq t_j \leq \min(N_j, t)\} \qquad (3.c.16)$$

Quando a hipótese nula (3.c.11) for verdadeira, (3.c.14) simplifica-se como

$$\delta_0(\underset{\sim}{t} / \sum_{j=1}^{k} t_j = t, s) = \frac{C(s, \underset{\sim}{t})}{\sum_{\Gamma} C(s, \underset{\sim}{t})} \qquad (3.c.17)$$

a qual permite fazer inferências acerca de (3.c.11), ou simplesmente: $H_o : \underset{\sim}{\beta} = \underset{\sim}{0}$, lembrando que é novamente condicionada a $\underset{\sim}{\gamma} = \underset{\sim}{0}$.

A probabilidade P, para o teste exato de (3.c.11) contra

$$H_a : \underset{\sim}{\beta} \neq \underset{\sim}{0}$$

é expressa por:

$$P = \sum_{\Lambda} \delta_0(\underset{\sim}{u}, / t, s) \qquad (3.c.18)$$

onde

$$\Lambda = \{\underset{\sim}{u} = (u_1, u_2, \ldots, u_k) : \delta_0(\underset{\sim}{u} / \sum_{j=1}^{k} u_j = t, s) \leq \delta_0(\underset{\sim}{t} / \sum_{j=1}^{k} t_j = t, s)\}$$

O procedimento assintótico geral para testar hipóteses deste tipo, para situações onde poderia haver mais de dois tratamentos, foi desenvolvido por Armitage (1966).

3.d. INFLUÊNCIA DAS VARIÁVEIS INTERVENIENTES NA "RAZÃO DOS PRODUTOS CRUZADOS"

Nos estudos do tipo caso-controle, a amostra considerada pode e, geralmente, é aconselhável ser estratificada em vários níveis do fator controle em estudo relacionado com a doença e/ou exposição. [Mantel e Haenszel (1959)]

Nestas circunstâncias, as "RPC" podem ser estimadas para cada nível do fator em estudo e, para obter a "RPC" global, há várias maneiras e métodos propostos por Cochran, por Mantel e Haenszel [citados em Fleiss (1970)], por Gart (1971), entre outros.

Porém, pouca atenção tem sido dada para o problema de influência das variáveis utilizadas para a formação de estratos nos cálculos das "RPC". Estes tipos de variáveis são comumente chamadas de *variáveis intervenientes ou explanatórias* ("*explanatory variables*").

O uso do modelo de regressão log-linear, segundo Zelen (1971), permite considerar na análise essa influência,

e ainda, segundo Prentice (1976), possibilita detectar se a variável é do tipo *efeito-modificante* ("*effect modifying variable*") e/ou *interferente*.

Os conceitos destes fatores, de efeito-modificante edo interferente, atualmente utilizados com grande assiduidade por epidemiologistas, foram recentemente discutidos por Miettinem (1974) e Fisher e Patil (1974).

Smith et al. (1975) em um estudo caso-controle, onde procuram detectar a relação entre exposição a estrógenos exógenos e câncer do endométrio na pós-menopausa, são citados por Prentice (1976) para fins de ilustração destes fatores.

São apresentados os resultados das observações de 243 casos de câncer e um mesmo número de controles, estratificados segundo pacientes do tipo A e B, como mostra a TABELA 3.d.A, onde paciente do tipo A é paciente clínica comum e paciente do tipo B é aquela possuidora de características que a classificam como paciente "padrão de referência" ("Referral Pattern").

TABELA 3.d.A Exposição a estrógeno para casos de câncer do endométrio e controles estratificados segundo pacientes do tipo A e B.

PACIENTES DO TIPO A

	CASOS	CONTROLES	TOTAL
EXPOSTOS	80	28	108
NÃO EXPOSTOS	37	60	97
TOTAL	117	88	205

PACIENTES DO TIPO B

	CASOS	CONTROLES	TOTAL
EXPOSTOS	56	18	74
NÃO EXPOSTOS	70	137	207
TOTAL	126	155	281

Fonte: Smith et al. (1975)

Da tabela citada extraímos que:

(i) Das pacientes do tipo A, 53% foram expostas e das pacientes do tipo B, somente 26%.

Aqui a exposição nas pacientes do tipo A é mais do dobro que nas pacientes do tipo B.

(ii) Dos casos 48% e dos controles 36% são pacientes do tipo A. Conseqüentemente, dos casos 52% e dos controles 64% são pacientes do tipo B.

Por (i) e (ii) pode-se verificar que, apesar da proporção de expostos nas pacientes do tipo A ser mais do dobro que nas pacientes do tipo B, a proporção de casos é menor nas pacientes do tipo A quando comparadas com as pacientes do tipo B. Isso leva a concluir que possivelmente o conjunto de características que levam a classificar as pacientes como sendo do tipo B será um fator interferente ("Confounding factor").

Agora, como a "RPC" nas pacientes do tipo B é maior que nas pacientes do tipo A, possivelmente poder-se-ia con-

cluir novamente que o mesmo conjunto de características definidoras da paciente "padrão de referência" será um fator efeito-modificante ("effect modifying factor"), se a diferença não for devida a variação aleatória.

A fim de determinar se a "RPC" é influenciada por cada variável e se estas podem ter importantes implicações para a natureza do processo da doença, Breslow (1976) reanalisou os dados apresentados por Stewart e Kneale (1970). Estes buscavam saber se a distribuição por idade de câncer na infância causado pela ação do Raio X, "in-utero", diferia da distribuição por idade, dos casos de câncer na infância nas crianças que não sofreram a ação do raio X. Os dados de Gart (1971) também serão apresentados aqui.

Os dados de Kneale (1971) tem como propósito mostrar que a "RPC" de câncer na infância varia de acordo com a idade dos casos e controles, descritos acima. Entretanto, esta conclusão foi questionada por Gehan (1972) que, por sinal, sugeriu o uso do modelo de regressão logístico de Cox (1970), dizendo que este poderia analisar melhor os dados. Porém, Breslow (1976) utilizou o modelo de regressão linear para o logarítmo da "RPC", proposto por Zelen (1971) em vez de seguir a sugestão de Gehan.

Considera-se que os dados obtidos serão distribuídos numa tabela 2x2, como a TABELA 3.a.A, que representa o resultado do j-ésimo estrato do fator B.

Para cada uma das k-tabelas, um vetor de ordem p .:

$$\underset{\sim}{Z}_j = (Z_{j1}, Z_{j2}, \ldots, Z_{jp}) \qquad (3.d.1)$$

$$j = 1, 2, \ldots, k,$$

de variáveis intervenientes é útil de se considerar.

Como o objetivo principal é o enfoque sobre a "RPC", é apropriado tomar a distribuição condicional em cada uma das tabelas, com os valores dos totais marginais, t_j, n_j e m_j, fixados:

$$Pr\{S_j = s_j / t_j, n_j, m_j\} = \frac{C(s_j, t_j)\ \Omega_j^{s_j}}{\sum_{v=0}^{t_j} C(v, t_j)\Omega_j^{v}} \qquad (3.d.2)$$

onde

$$C(s_j, t_j) = \binom{n_j}{s_j}\binom{m_j}{r_j} = \binom{n_j}{s_j}\binom{m_j}{t_j - s_j} \qquad (3.d.3)$$

Note-se que (3.d.2) é válida sob quaisquer tipos de esquemas conhecidos de amostragem, apresentados pela TABELA 3.a.A, sendo que Ω_j é a "RPC" do estrato j, com $j = 1, 2, \ldots, k$.

Zelen (1971) propôs o modelo de regressão do tipo:

$$\Omega_j(\underset{\sim}{\beta}) = \exp(\underset{\sim}{\beta}' \underset{\sim}{Z}_j) \qquad (3.d.4)$$

O logaritmo neperiano de $\Omega_j(\underset{\sim}{\beta})$ é linearmente relacionado às variáveis intervenientes Z_{ji} ($j = 1, 2, \ldots, k$; $i = 1, 2, \ldots, p$) através do vetor de parâmetros de ordem p,

$\underset{\sim}{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$.

Como o objetivo é obter o estimador de $\underset{\sim}{\beta}$, utilizando os métodos convencionais de inferência baseada nas distribuições condicionais (3.d.2), a introdução do modelo (3.d.4), proposto por Zelen (1971), torna bastante simplificado o cálculo do logarítmo da função de verossimilhança e suas derivadas, haja visto que o termo $\underset{\sim}{\beta}$ ficará bem explícito na densidade condicional (3.d.2). Ao reescrever essa probabilidade, tem-se:

$$Pr\{S_j = s_j / t_j, m_j, n_j\} = \frac{C(s_j, t_j) \exp(s_j \underset{\sim}{\beta}' \underset{\sim}{Z}_j)}{\sum_{v=0}^{t_j} C(v, t_j) \exp(v \underset{\sim}{\beta}' \underset{\sim}{Z}_j)} \quad (3.d.5)$$

para $j = 1, 2, 3, \ldots, k$.

Assim, as equações de verossimilhança são escritas como:

$$\sum_{j=1}^{k} s_j Z_{jh} = \sum_{j=1}^{k} \mu_j(\underset{\sim}{\beta}) Z_{jh} \quad (3.d.6)$$

com $h = 1, 2, \ldots, p$.

onde $\mu_j(\underset{\sim}{\beta})$ é a média da distribuição condicional discreta (3.d.5)

A matriz de informação $I(\underset{\sim}{\beta})$ tem componentes:

$$-I_{h\ell}(\underset{\sim}{\beta}) = \sum_{j=1}^{k} Z_{jh} Z_{j\ell} \delta_j^2(\underset{\sim}{\beta}) \qquad (3.d.7)$$

com $h, \ell = 1, 2, \ldots, p$,

de tal sorte que a inversa, $I^{-1}(\underset{\sim}{\beta})$, fornece a matriz de variância - covariância das estatísticas "estimadoras" dos parâmetros componentes do vetor $\underset{\sim}{\beta}$.

As equações (3.d.6) podem ser resolvidas por métodos iterativos como por exemplo de Newton-Raphson, partindo de um valor inicial conveniente, por exemplo $\underset{\sim}{\beta} = \underset{\sim}{0}$. Pode-se, porém, fazer inferências utilizando a própria função de probabilidade (3.d.5).

Se ocorrer que os estratos individuais sejam grandes, encontraremos dificuldades e, até, será impossível computar os valores exatos dos momentos condicionais. Assim, nestas circunstâncias, a forma de contorná-las é substituí-los por momentos aproximados, levando a obter uma distribuição assintótica de (3.d.5) que, segundo Hannan e Harkness (1963), será aproximadamente normal de média assintótica $\mu_j(\underset{\sim}{\beta})$ e variância assintótica $\sigma_j^2(\underset{\sim}{\beta})$ que serão obtidas de:

$$\frac{\tilde{\mu}_j(\underset{\sim}{\beta})\left[m_j - x_j + \tilde{\mu}_j(\underset{\sim}{\beta})\right]}{\left[x_j - \tilde{\mu}_j(\underset{\sim}{\beta})\right] \cdot \left[n_j - \tilde{\mu}_j(\underset{\sim}{\beta})\right]} = \Omega_j(\underset{\sim}{\beta}) = \exp(\underset{\sim}{\beta}' \underset{\sim}{Z}_j) \qquad (3.d.8)$$

para $j = 1, 2, \ldots, k$.

e

$$\sigma_j^2(\underset{\sim}{\beta}) = \left[\frac{1}{\tilde{\mu}_j(\underset{\sim}{\beta})} + \frac{1}{m_j - t_j + \tilde{\mu}_j(\underset{\sim}{\beta})} + \frac{1}{t_j - \tilde{\mu}_j(\underset{\sim}{\beta})} + \frac{1}{n_j - \tilde{\mu}_j(\underset{\sim}{\beta})}\right]^{-1} \tag{3.d.9}$$

Vê-se facilmente que considerando $\Omega_j(\underset{\sim}{\beta}) = 1$, verdadeira,

$$\tilde{\mu}_j(\underset{\sim}{\beta}) = t_j n_j / N_j \qquad j = 1, 2, \ldots k.$$

Levando os resultados de (3.d.8) e (3.d.9) para (3.d.6) e (3.d.7), proporcionam-se aproximações para as equações de verossimilhança e matriz de informação desejáveis para se aplicar quando os estratos são grandes.

Para efeito de ilustração são apresentados os dados de Gart (1971) sobre os efeitos carcinogênicos de um certo fungicida em ratos, classificados segundo raça, sexo e tratamento, como mostra a TABELA 3.d.B., apresentada por Breslow (1976).

Vê-se, neste caso, que foram consideradas três variáveis intervenientes: $\underset{\sim}{Z}_j = (Z_{j1}, Z_{j2}, Z_{j3})$, usadas para expressar o logaritmo da "RPC" relacionando tratamento e tumor por meio de uma constante, mais a contribuição aditiva de raça e sexo.

A TABELA 3.d.C apresenta resultados de ajustamento do modelo pelos métodos exato e assintótico, utilizando os valores da TABELA 3.d.B.

TABELA 3.d.B Número de ratos classificados por raça, sexo e tratamento

G R U P O	TUMOR (1)	NÃO	TUMOR (2)	NÃO	TUMOR (3)	NÃO	TUMOR (4)	NÃO
CONTROLES	5	74	3	84	10	80	3	79
TRATADOS	4	12	2	14	4	14	1	14
VARIÁVEIS INTERVENIENTES								
Z_1 (Constante)	1		1		1		1	
Z_2 (Raça)	-1		-1		1		1	
Z_3 (Sexo)	-1		1		-1		1	
"RAZÃO DOS PRODUTOS CRUZADOS" $\hat{\Omega}_j$	4,93		4,00		2,29		1,88	

FONTE: Breslow (1976)

TABELA 3.d.C Resultados do ajustamento do modelo de regressão do log da "RPC" para os dados da TABELA 3.d.B

MÉTODO	TESTE DE R.V. DE $\underset{\sim}{\beta} = \underset{\sim}{0}$ $(-2\ \ell n\ \Lambda)$	COEFICIENTE DE REGRESSÃO $\pm$ DESVIO-PADRÃO		
		β_1(CONSTANTE)	β_2(RAÇA)	β_3(SEXO)
EXATO	7,720	$1,094 \pm 0,444$	$-0,375 \pm 0,412$	$-0,100 \pm 0,448$
ASSINTÓTICO	8,065	$1,095 \pm 0,447$	$-0,382 \pm 0,416$	$-0,102 \pm 0,452$

FONTE: Breslow (1976)

Com estes resultados, exatos e assintóticos, poderemos obter as "razões dos produtos cruzados" estimadas segundo um modelo log-linear considerando a influência das variáveis intervenientes.

Sabemos que a "razão dos produtos cruzados" no modelo log-linear (3.d.4), é dada por:

$$\Omega_j(\underset{\sim}{\beta}) = \exp\left[\underset{\sim}{\beta}'\underset{\sim}{Z}_j\right]$$

Para o nosso caso onde k=4 e p=3, o vetor dos parâmetros é:

$$\underset{\sim}{\hat{\beta}} = \begin{cases} (1,094;\ -0,375;\ -0,100)', \text{ para o exato} \\ (1,095;\ -0,382;\ -0,102)', \text{ para o assintótico} \end{cases}$$

e os vetores das variáveis intervenientes são:

$\underset{\sim}{Z}_1 = (1,-1,-1)'$; $\underset{\sim}{Z}_2 = (1,-1,1)'$; $\underset{\sim}{Z}_3 = (1,1,-1)'$ e $\underset{\sim}{Z}_4 = (1,1,1)'$

resultando a TABELA 3.d.D, das "RPC" ajustadas pelo modelo (3.d.4):

TABELA 3.d.D Estimação das "RPC" ajustadas pelo método exato e assintótico.

MÉTODO	$\hat{\Omega}_1$	$\hat{\Omega}_2$	$\hat{\Omega}_3$	$\hat{\Omega}_4$
EXATO	4,802	3,931	2,268	1,857
ASSINTÓTICO	4,850	3,955	2,259	1,842

Mesmo com o número pequeno de observações que aparecem em cada cela, o modelo de três covariáveis aqui utilizado parece adequar-se quase que perfeitamente a estes dados particulares, pois tanto o método exato quanto o assintótico fornecem "RPC" estimadas inteiramente comparáveis.

É interessante notar que, ao observar os dados da TABELA 3.d.B, e se os reclassificar quanto ao sexo e quanto à raça, há uma ligeira diferença entre as relações. Porém, nenhuma rejeita a hipótese de igualdade, embora o efeito global do tratamento sobre o tumor certamente rejeitará, como se reflete pelo coeficiente do termo constante.

Vejamos, pois, em primeiro lugar a variável raça:

$$Z_2 = \begin{cases} -1 \text{ para 1a. raça} \\ 1 \text{ para 2a. raça} \end{cases}$$

TABELA 3.d.E Distribuição dos dados quando estratificados segundo a raça dos ratos

GRUPO	1a. RAÇA		2a. RAÇA	
	TUMOR	NÃO	TUMOR	NÃO
TRATADOS	6	26	5	28
CONTROLES	8	158	13	159

"RPC" (1a. raça) = $\hat{\Omega}_{R1} = 4,56$ e "RPC" (2a. raça) = $\hat{\Omega}_{R2} = 2,18$

Portanto, é menos pronunciada para a 2a. raça.

Para o caso de sexo, consideremos:

$$Z_3 = \begin{cases} 1, & \text{fêmea} \\ -1, & \text{macho} \end{cases}$$

TABELA 3.d.F Distribuição dos dados quando estratificados segundo o sexo dos ratos

GRUPOS	FÊMEAS		MACHOS	
	TUMOR	NÃO	TUMOR	NÃO
TRATADOS	3	28	8	26
CONTROLES	6	163	15	154

"RPC" (fêmea) = $\hat{\Omega}_F = 2,91$ e "RPC" (macho) = $\hat{\Omega}_M = 3,16$

Observa-se que a relação entre tratamento e tumor é ligeiramente menor nas fêmeas.

Outro exemplo citado por Breslow (1976), a fim de aplicar o modelo de Zelen (1971) para descrever a influência das variáveis concomitantes sobre a "RPC", é a reanálise que efetuou em 5.926 casos de mortes de crianças por doenças malignas no período de 1954 a 1964, dentre os 6.347 casos apresentados por Kneale (1971), para análise da relação entre as mortes de câncer na infância e a exposição das mães ao raio-X durante a gestação (*). O autor cita que as mães foram procuradas e concordaram em ser entrevistadas. Os resultados desta pesquisa, estratificados segundo o ano de nascimento e idade ao morrer, resultaram em 120 tabelas 2x2, com as entradas:

Casos: Mortes por câncer infantil

Controles: Mortes por outras doenças malignas

e

Expostos: Mães com irradiação obstétrica ("in utero")

Não Expostos: Mães sem irradiação obstétrica ("in útero").

Utilizando variáveis intervenientes como o ano de nascimento (coorte de nascimento), a idade ao morrer e as variáveis que são funções destas citadas, efetuou-se o cálculo dos parâmetros que melhor descrevessem as "RPC", como

---

(*) Por conveniência utilizaremos as expressões: Câncer infantil e irradiação obstétrica ("in utero").

mostra a TABELA 3.d.G, onde são especificadas quantas e quais variáveis que foram consideradas. (Ver Tabela 3.d.G. na página seguinte).

A TABELA 3.d.G mostra evidência de que há uma relação significante entre exposição e doença, através da comparação das duas primeiras linhas.

Crianças com irradiação obstétrica tem um risco relativo global de câncer estimado por: $\hat{\Omega} = \exp(0,5102) = 1,67$ vezes maior quando comparadas com as crianças sem irradiação obstétrica ("in utero").

O fato de introduzir a variável, ano de nascimento, acarreta um decréscimo linear significante do logaritmo da "RPC" com

$$\chi^2_o = 118,75 - 112,53 = 6,22$$

e $$-2\ell n\ \Lambda = -2\left[-221,22 - (217,66)\right] = 7,12$$

comparado com o valor tabelado $\chi^2_{(1);0,10} = 2,706$ (3.d.10)

Agora, adicionando um termo quadrático em ano de nascimento, melhorará ainda mais o ajustamento, pois resulta:

$$\chi^2_o = 112,53 - 108,74 = 3,79$$

$$-2\ell n\ \Lambda = -2\left[214,89 - (-217,66)\right] = 5,54$$

comparado com (3.d.10).

Verifica-se, ao comparar a 5a. linha com a 3a., que houve pouca melhora ao introduzir um termo linear de idade

TABELA 3.d.G Resultado do ajustamento de vários modelos de regressão do log das "RPC" pelo método da máxima verossimilhança assintótica para os dados apresentados por Kneale (1971)

NÚMERO DE VARIÁVEIS INCLUÍDAS	LOGARITMO DA VEROSSIMILHANÇA	QUI-QUADRADO (ADEQUABILIDADE DO AJUSTAMENTO)	COEFICIENTE DE REGRESSÃO ± DESVIO PADRÃO					
			CONSTANTE	z_1	$z_1^2 - 22$	z_2	$z_2^2 - 33$	$z_1 \times z_2$
0	-262,83	196,74						
1	-221,22	118,75	0,5102±0,0564					
2	-217,66	112,53	0,5218±0,0567	-0,0390±0,0145				
3	-214,89	108,74	0,5707±0,0611	-0,0450±0,0150	0,0068±0,0030			
3	-217,41	112,25	0,5297±0,0576	-0,0312±0,0176		0,0105±0,0133		
6	-214,11	107,32	0,4738±0,1308	-0,0411±0,0182	0,0029±0,0057	0,0069±0,0154	-0,0025±0,0028	-0,0054±0,006

FONTE: Breslow (1976)

ao morrer e na comparação da 6a. linha com a 4a., nota-se que, apesar de ter incluído termos lineares e quadráticos de idade e, ainda, uma interação linear de ano de nascimento e idade ao morrer, praticamente em nada melhorou.

Breslow (1976) comenta ainda que a inclusão de termos quadrático, cúbico e quártico de idade ao morrer não produziu aperfeiçoamento sobre a 5a. linha da TABELA 3.d.G.

A Figura 3.d.1 citada em Breslow (1976) ilustra o delineamento das "RPC" ajustadas por idade versus coorte de nascimento, e, além disso, as linhas de ajuste, tanto linear quanto quadrático, para uma tentativa de interpretação. Pode-se concluir que os valores dos riscos relativos tem diminuído desde meados da década de 40, mas parecem ter aumentado um pouco, novamente, no início da década de 60.

Isso parece contradizer a situação descrita anteriormente, dos ajustes por ano de nascimento. Assim, esta análise confirma as dúvidas de Gehan (1972) sobre se a distribuição por idade de casos radiogênicos e idiopáticos diferem verdadeiramente.

Enquanto os resultados de Kneale (1971) são aparentemente contrariados, é bom ressaltar que Breslow (1976) considerou câncer ocorrendo em todas as localizações e usou a idade ao morrer em vez de idade ao início da doença que parece mais relevante. No entanto, é provável que a dependência entre log da "RPC" e idade de início da doença se veri-

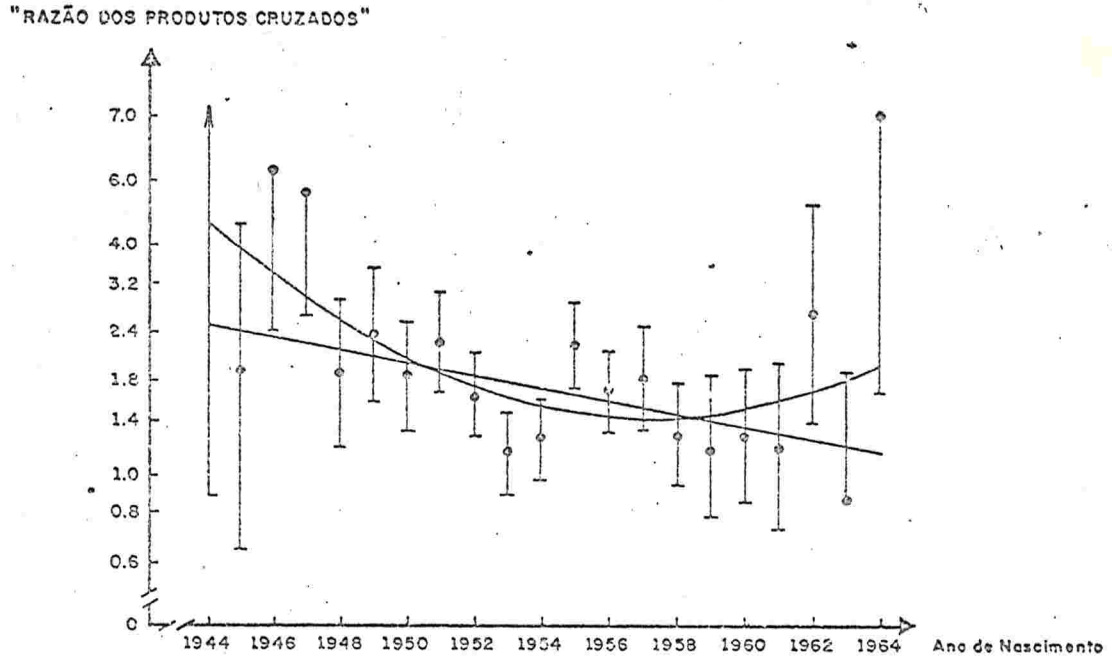

Figura 3.d.1

DELINEAMENTO DA "RAZÃO DOS PRODUTOS CRUZADOS" DE IRRADIAÇÃO "IN UTERO", AJUSTADA POR IDADE. INTERVALO DE CONFIANÇA, COM GRAU DE CONFIANÇA APROXIMADO DE 80%.

FONTE: BRESLOW (1976)

fique para alguns tipos particulares de câncer na infância, como é o caso da leucemia linfocítica aguda, por exemplo.

Um outro procedimento que completa as idéias de Zelen (1971), levando a uma extensão do seu procedimento, portanto permitindo uma aplicabilidade mais ampla, foi descrito por Prentice (1976).

Relembramos aqui que o modelo exposto por Zelen (1971), onde os dados eram distribuídos convenientemente em k tabelas 2x2, permite testar se a "RPC" é constante nestas tabelas e, ainda mais, é útil para testar a igualdade de probabilidade de exposição sobre as k tabelas, condicionada ao fato de que a "RPC" é constante.

Na terminologia empregada por Prentice (1976), o primeiro teste é o que permite detectar se a variável interveniente é uma variável indicadora do fator como não sendo efeito-modificante, ao passo que o segundo teste serve para verificar a não presença de fator interferente na "RPC" estimada.

O método que Prentice (1976) descreveu, permite uma extensão no sentido de que é possível também o emprego de variáveis contínuas como variáveis interferentes ou efeito-modificantes, permitindo um estudo quantitativo da influência destes fatores na "RPC" calculada. A extensão vai mais além, pois permite estudar estes fatores embora permaneçam mantidos os pareamentos, contrastando com o método de Mantel e

Haenszel (1959) e seus refinamentos, onde era necessário abandonar o pareamento se os fatores não incluídos no pareamento fossem usados para formar os estratos.

Ele construiu o modelo tomando um vetor de variáveis intervenientes para um indivíduo sob estudo:

$$\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_p)' \qquad (3.d.11)$$

Para contribuir na descrição da "RPC" relativa à exposição a um fator suspeito (F: 1,presente e 0, ausente) e doença em estudo (D: 1, presente e 0, ausente).

No caso do exemplo citado na TABELA 3.d.A, o vetor $\underset{\sim}{Z}$ poderia ser tomado como a variável indicadora de se o paciente é do tipo A (Z=0) ou se o paciente é do tipo B (Z=1).

A não variação da "razão dos produtos cruzados", para quaisquer $\underset{\sim}{Z}$ fixados, será descrita através de:

$$\frac{Pr\{D=1/F=1,\underset{\sim}{Z}\}/Pr\{D=1/F=0,\underset{\sim}{Z}\}}{Pr\{D=0/F=1,\underset{\sim}{Z}\}/Pr\{D=0/F=0,\underset{\sim}{Z}\}} = \frac{Pr\{F=1/D=1,\underset{\sim}{Z}\}/Pr\{F=0/D=1,\underset{\sim}{Z}\}}{Pr\{F=1/D=0,\underset{\sim}{Z}\}/Pr\{F=0/D=0,\underset{\sim}{Z}\}} \qquad (3.d.12)$$

Analisando (3.d.12), verificamos que:

(i) O numerador do termo do primeiro membro é o risco relativo da doença para aquele valor de $\underset{\sim}{Z}$ fixado;

(ii) O denominador da mesma fração, no caso de doenças de baixa incidência, tenderá à unidade;

(iii) O termo do segundo membro pode ser facilmente estimado em um estudo caso-controle ao tomar a probabilidade de exposição descrita através do modelo de regressão logística binária de Cox (1970):

$$Pr\{F/D,Z\} = exp\left[(\delta+\alpha D+\underset{\sim}{Z}'\underset{\sim}{\beta})F\right] \Big/ \left[1+exp(\delta+\alpha D+\underset{\sim}{Z}'\underset{\sim}{\beta})\right] \qquad (3.d.13)$$

onde, por exemplo:

δ descreve a frequência de exposição ao estrógeno nos controles;

α descreve a "exposição adicional" nos casos e

$\underset{\sim}{\beta}$ no nosso caso particular é o vetor dos coeficientes que permitem descrever a dependência das variáveis intervenientes expressas por $\underset{\sim}{Z}$ na exposição ao estrógeno como fator suspeito na ocorrência de câncer do endométrio.

Note-se que no caso particular, $\underset{\sim}{Z}$ possui apenas um componente.

Por aplicação da expressão (3.d.13) sobre (3.d.12), resulta no segundo membro uma "RPC", como é fácil verificar:

$$\Omega = \exp(\alpha) \qquad (3.d.14)$$

Sendo a "RPC" e^{α} constante, concluímos que uma análise baseada em (3.d.13) leva a inferência acerca da "RPC", assumindo que a influência das variáveis intervenientes $\underset{\sim}{Z}$

na probabilidade de exposição é comum tanto aos casos quanto aos controles.

Modificando (3.d.13), consegue-se atenuar esta restrição:

$$Pr\{F/D,Z\} = exp\ [(\delta+\alpha D+\underset{\sim}{Z}'\underset{\sim}{\beta}+D\underset{\sim}{Z}'\underset{\sim}{\gamma})F]\ /\ [1+exp(\delta+\alpha D+\underset{\sim}{Z}'\underset{\sim}{\beta}+D\underset{\sim}{Z}'\underset{\sim}{\gamma})] \qquad (3.d.15)$$

onde $\underset{\sim}{\gamma}$ descreve uma "dependência adicional" a $\underset{\sim}{Z}$ na probabilidade de exposição dos casos, quando comparada com a dos controles.

A substituição da expressão (3.d.15) em (3.d.12), como foi feito com (3.d.13), resulta na "RPC"

$$\Omega = \exp\ (\alpha + \underset{\sim}{Z}'\underset{\sim}{\gamma}) \qquad (3.d.16)$$

As conclusões possíveis de serem obtidas acerca do vetor $\underset{\sim}{Z}$ serão delineadas abaixo:

(i) Os componentes de $\underset{\sim}{Z}$ que tiverem seus correspondentes componentes de $\underset{\sim}{\beta}$, quando rejeitada a hipótese: $\underset{\sim}{\beta} = \underset{\sim}{0}$, serão potencialmente variáveis indicadoras de fatores interferentes

(ii) Os componentes de $\underset{\sim}{Z}$, com os correspondentes componentes de $\underset{\sim}{\gamma}$, quando rejeitada a hipótese $\underset{\sim}{\gamma} = 0$, serão considerados variáveis efeito-modificantes.

O que se pode extrair daqui é que um fator poderá ser interferente sem ser efeito-modificante, ou vice-versa, como assinalam Fisher e Patil (1974). Assim, seria útil tentar analisar separadamente, em vez de descrever por um único vetor aleatório como foi feito com a aplicação de (3.d.15).

Designar de $\underset{\sim}{Z}$ e $\underset{\sim}{W}$, os vetores que incluam as variáveis indicadores de fatores interferentes e efeito-modificantes, respectivamente, foi a medida sugerida por Prentice (1976).

A probabilidade (3.d.15) fica sendo:

$$Pr\{F/D, \underset{\sim}{Z}, \underset{\sim}{W}\} = exp\left[(\delta+\alpha D+\underset{\sim}{Z}'\underset{\sim}{\beta}+D\underset{\sim}{W}'\underset{\sim}{\gamma})F\right] / \left[1+exp(\delta+\alpha D+\underset{\sim}{Z}'\underset{\sim}{\beta}+D\underset{\sim}{W}'\underset{\sim}{\gamma})\right] \quad (3.d.17)$$

Levando esta expressão em (3.d.12), resulta: $\Omega = \exp(\alpha+\underset{\sim}{W}'\underset{\sim}{\gamma})$ (3.d.18)

Com o método de Mantel e Haenszel (1959), os dados poderiam ser agrupados em estratos para controlar os fatores interferentes.

Então, o vetor $\underset{\sim}{Z}$ pode consistir de variáveis indicadoras dos estratos, embora $\underset{\sim}{W}$ inclua as variáveis quantitativas reais utilizadas na formação destes estratos.

A extensão do método será tal que aqueles fatores comuns a um par serão considerados pelo vetor de variáveis intervenientes $\underset{\sim}{Z}$ e/ou $\underset{\sim}{W}$. Efetua-se a análise utilizando a expressão (3.d.17); porém, se for considerado que δ pode variar, de par a par, o modelo será:

$$P_i\{F/D,\underset{\sim}{Z},\underset{\sim}{W}\} = exp\left[(\delta_i + \alpha D + \underset{\sim}{Z}'\underset{\sim}{\beta} + D\underset{\sim}{W}'\underset{\sim}{\gamma})F\right] \Big/ \left[1+exp(\delta_i + \alpha D + \underset{\sim}{Z}'\underset{\sim}{\beta} + D\underset{\sim}{W}'\underset{\sim}{\gamma})\right]$$

$$\text{com } i = 1,2,\ldots,n \qquad (3.d.19)$$

mas a "RPC" não se altera, isto é, será como antes:

$$\Omega = exp\ (\alpha + \underset{\sim}{W}'\gamma)$$

Há um outro refinamento no modelo. É o caso em que se considera mais de um tipo de controle, como, por exemplo, pacientes com doenças outras que não aquelas sob estudo.

Nestas circunstâncias, a "RPC" pode ser analisada em relação a um tipo de controle desejado. Assim, isto consiste em alterar αD para:

$$\underset{\sim}{\alpha}'\underset{\sim}{D} = \sum_{j=1}^{\ell} \alpha_j D_j \qquad (3.d.20)$$

onde $D_j = \begin{cases} 1 \text{ para o } j\text{-ésimo tipo de controle} \\ 0 \text{ para os outros.} \end{cases}$

A transformação (3.d.20) pode ser aplicada tanto em (3.d.17) quanto em (3.d.19), resultando uma mesma "RPC" como era de se esperar:

$$\Omega_j = \exp\ (\alpha_j + \underset{\sim}{W}'\underset{\sim}{\gamma}) \qquad (3.d.21)$$

$$j = 1,2,3,\ldots,\ell$$

É possível, e em algumas vezes útil, fazer generalizações sobre $\underset{\sim}{\beta}$ e $\underset{\sim}{\gamma}$, também, dependendo do tipo de controle empregado.

A inferência acerca dos parâmetros da expressão (3.d.17), surge diretamente dos procedimentos de máxima verossimilhança e razão de verossimilhança descritos por Cox (1970). O mesmo processo pode ser aplicado a (3.d.19) para eliminar os parâmetros de pareamento δ_i.

O método de máxima verossimilhança, descrito por Cox(1970), proporciona-nos os estimadores dos parâmetros de regressão, os erros-padrão assintóticos destes estimadores e, portanto, seus intervalos de confiança.

Partindo da probabilidade de um esquema Bernoulli, temos:

$$Pr\{Y_i = y_i\} = p_{Y_i}(y_i) = \left[\frac{e^{\underset{\sim}{a}_i\,\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{a}_i\underset{\sim}{\beta}}}\right]^{y_i}\left[\frac{1}{1+e^{\underset{\sim}{a}_i\underset{\sim}{\beta}}}\right]^{1-y_i}$$

$$com\ i = 1,2,\ldots,n$$

$$e\quad y_i = 0,1$$

Conseqüentemente, a função de verossimilhança é:

$$V[\underset{\sim}{\beta}] = \prod_{i=1}^{n} p_{Y_i}(y_i) = \frac{\exp\left[\sum_{i=1}^{n} \underset{\sim}{a}_i \underset{\sim}{\beta}\, y_i\right]}{\prod_{i=1}^{n}\left[1+e^{\underset{\sim}{a}_i \underset{\sim}{\beta}}\right]}$$

onde $\underset{\sim}{a}_i \underset{\sim}{\beta} = \sum_{s=1}^{p} a_{is}\beta_s$ e portanto com $\sum_{i=1}^{n} a_{is}y_i = t_s$

resulta que $\sum_{i=1}^{n} a_{is}\underset{\sim}{\beta}\, y_i = \sum_{i=1}^{n}\sum_{s=1}^{p} a_{is}\beta_s y_i = \sum_{s=1}^{p} \beta_s t_s$

A função de verossimilhança então será:

$$V[\beta] = \exp\left[\sum_{s=1}^{p} \beta_s t_s\right] \Big/ \prod_{i=1}^{n}\left[1+\exp\left(\underset{\sim}{a}_i \underset{\sim}{\beta}\right)\right]$$

Para obter os estimadores e matriz de variância-covariância devem-se calcular as derivadas parciais de primeira e segunda ordens da função de verossimilhança. Resultado equivalente se obtém, utilizando-se o logaritmo da função de verossimilhança em vez da própria função de verossimilhança. Faz-se essa transformação por facilidade operacional na estimação dos parâmetros desejados. Assim, o logaritmo da função de verossimilhança será:

$$L(\underset{\sim}{\beta}) = \sum_{s=1}^{p} \beta_s t_s - \sum_{i=1}^{n} \log\left[1+\exp\left(\underset{\sim}{a}_i \underset{\sim}{\beta}\right)\right] \quad (3.d.22)$$

Resultando que:

$$\frac{\partial L\ (\underset{\sim}{\beta})}{\partial \beta_s} = t_s - \sum_{i=1}^{n} \frac{a_{is}\ \exp\ (\underset{\sim}{a}_i\ \underset{\sim}{\beta})}{1 + \exp\ (\underset{\sim}{a}_i\ \underset{\sim}{\beta})} \qquad (3.d.23)$$

com $s = 1,2,\ldots,p$

E os elementos $I_{s_1 s_2}\ (\underset{\sim}{\beta})$ da matriz de informação cuja inversa é a matriz de variância-covariância, são:

$$I_{s_1 s_2}(\underset{\sim}{\beta}) \equiv . E\{- \frac{\partial^2 L\ (\underset{\sim}{\beta})}{\partial \beta_{s_1}\ \partial \beta_{s_2}}\} = \sum_{i=1}^{n} \frac{a_{is_1}\ a_{is_2}\ \exp(\underset{\sim}{a}_i \underset{\sim}{\beta})}{\left[1+\exp\ (\underset{\sim}{a}_i\ \underset{\sim}{\beta})\right]^2} \qquad (3.d.24)$$

Os estimadores de máxima verossimilhança de $\underset{\sim}{\beta}$ que serão representados por $\hat{\underset{\sim}{\beta}}$ satisfazem as equações:

$$\left[\frac{\partial L\ (\underset{\sim}{\beta})}{\partial \beta_s}\right]_{\underset{\sim}{\beta} = \hat{\underset{\sim}{\beta}}} = 0 \qquad (3.d.25)$$

Há muitos meios de resolver o sistema (3.d.25): segundo Beale (1967), há a programação não linear e, segundo Draper e Smith (1966), há o procedimento de *regressão"passo a passo"* (*"step-wise"*). Há também um procedimento bastante utilizado que é o método iterativo de Newton-Raphson.

Exemplificaremos supondo duas situações hipotéticas:

(I) todas as probabilidades de sucesso (p_i) são pequenas;

(II) nenhuma das probabilidades é próxima de zero ou um.

SITUAÇÃO I:

Supondo que:

(i) a probabilidade de sucesso, $p_i = e^{\lambda_i}/(1+e^{\lambda_i})$ é pequena;

(ii) a variância nos λ_i é pequena;

(iii) o modelo considerado contém um termo constante, ou seja:

$$\lambda_i = \alpha + \beta\ (x_i - \bar{x})$$

Assim:

$$P_i = \frac{e^{\lambda i}}{1+e^{\lambda i}} = \frac{\exp\left[\alpha+\beta\ (x_i-\bar{x})\right]}{1+\exp\left[\alpha+\beta(x_i-\bar{x})\ \right]}$$

e

$$P_{Y_i}(y_i) = \left\{\frac{\exp\left[\alpha+\beta(x_i-\bar{x})\right]}{1+\exp\left[\alpha+\beta(x_i-\bar{x})\right]}\right\}^{y_i}\left\{\frac{1}{1+\exp\left[\alpha+\beta(x_i-\bar{x})\right]^2}\right\}^{1-y_i}$$

onde $i = 1,2,\ldots,n$ e com $y_i = 0,1$.

Vê-se que $\underset{\sim}{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ e a função de verossimilhança é

$$V\left[\underset{\sim}{\beta}\right] = \frac{\exp\left[\alpha \sum_{i=1}^{n} y_i + \beta \sum_{i=1}^{n} (x_i-\bar{x})y_i\right]}{\prod_{i=1}^{n}\{\ 1+\exp\left[\alpha+\beta\ (x_i - x)\right]\}}$$

ou seu logaritmo:

$$L\ (\underset{\sim}{\beta}) = \alpha \sum_{i=1}^{n} y_i + \beta \sum_{i=1}^{n} (x_i-\bar{x})y_i - \sum_{i=1}^{n} \ell n\{1+\exp\left[\alpha+\beta\ (x_i-\bar{x})\right]\} \qquad (3.d.26)$$

Por (i)

$$p_i = e^{\lambda i} / \left[1+e^{\lambda i}\right] \text{ é pequena, isto é,}$$

$$e^{\lambda i} = \exp[\alpha+\beta(x_i-\bar{x})] \text{ é pequena.}$$

Resulta que, $1+\exp\left[\alpha+\beta(x_i-\bar{x})\right] \simeq 1$ ou $\ell n\{1+\exp\left[\alpha+\beta(x_i-\bar{x})\right]$ é pequeno, ou seja, poderemos considerar

$$\sum_{i=1}^{n} \ell n\{1+\exp\left[\alpha+\beta(x_i-\bar{x})\right]\} = e^{\alpha} \sum_{i=1}^{n} \exp\left[\beta(x_i-\bar{x})\right]$$

A expressão (3.d.26) tornar-se-á

$$L(\underset{\sim}{\beta}) = \alpha \sum_{i=1}^{n} y_i + \beta \sum_{i=1}^{n} (x_i-\bar{x})y_i - e^{\alpha} \sum_{i=1}^{n} \exp\left[\beta(x_i-\bar{x})\right] \quad (3.d.27)$$

Mas desenvolvendo a última parcela de (3.d.27) por uma série de potência particular

$$e^{\theta} = \sum_{x=0}^{\infty} \frac{\theta^x}{x!} \quad \text{resulta que}$$

$$\exp\left[\beta(x_i-\bar{x})\right] = 1+\beta(x_i-\bar{x})+\frac{1}{2}\beta^2(x_i-\bar{x})^2+\frac{1}{6}\beta^3(x_i-\bar{x})^3 \ldots$$ e tomando as três primeiras parcelas, teremos:

$$e^{\alpha} \sum_{i=1}^{n} \exp\left[\beta(x_i-\bar{x})\right] \simeq e^{\alpha}\left[n+\beta \sum_{i=1}^{n} (x_i-\bar{x})+\frac{1}{2}\beta^2(x_i-\bar{x})^2\right]$$

notando que $\sum_{i=1}^{n} (x_i-\bar{x})=0$, substituindo em (3.d.26), obtém-se:

$$L(\beta) \simeq \alpha \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} (x_i-\bar{x})yi - e^{\alpha}\left[n+\frac{1}{2}\beta^2 \sum_{i=1}^{n} (x_i-x)^2\right]$$

$$\frac{\partial L(\underset{\sim}{\beta})}{\partial \alpha} = \sum_{i=1}^{n} y_i - e^{\alpha} n \qquad (3.d.28)$$

$$\frac{\partial L(\underset{\sim}{\beta})}{\partial \beta} = \sum_{i=1}^{n} (x_i - \bar{x})\, y_i - e^{\alpha} \beta \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (3.d.29)$$

Igualando a zero estas duas expressões, (3.d.28) e (3.d.29), forma-se o sistema de equações para estimadores de máxima verossimilhança, cujas raízes são:

$$\hat{\alpha} = \sum_{i=1}^{n} y_i / n$$

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n} y_i} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

e

$$\text{Var}\,[\hat{\alpha}] = e^{-\alpha}/n$$

$$\text{Cov}\,[\hat{\alpha}, \hat{\beta}] = 0$$

$$\text{Var}\,[\hat{\beta}] = e^{-\alpha} / \sum_{i=1}^{n} (x_i - \bar{x})^2$$

pois os elementos da matriz de informação, obtidos por (3.d.24), são:

$$I_{11}\,(\underset{\sim}{\beta}) = E\left\{ - \frac{\partial^2 L(\underset{\sim}{\beta})}{[\partial \alpha]^2} \right\} \quad E\left[n\, e^{\alpha} \right] = n\, e^{\alpha}$$

$$I_{12}(\underset{\sim}{\beta}) = I_{21}(\underset{\sim}{\beta}) = E\left\{\frac{-\partial^2 L(\underset{\sim}{\beta})}{\partial\beta\ \partial\alpha}\right\} = E\{0\} = 0$$

$$I_{22}(\underset{\sim}{\beta}) = E\left\{-\frac{\partial^2 L(\underset{\sim}{\beta})}{[\partial\ \beta]^2}\right\} = E\left\{e^{\alpha}\sum_{i=1}^{n}(x_i-\bar{x})^2 = e^{\alpha}\sum_{i=1}^{n}(x_i-\bar{x})^2\right.$$

O sistema de equações de estimadores de máxima verossimilhança (3.d.25) poderá ser resolvido pelo método bem conhecido de Newton-Raphson, quando o número de parâmetros é reduzido, resultando em soluções razoavelmente eficientes.

SITUAÇÃO II:

Quando, porém, nenhuma das probabilidades é próxima de zero ou um, é possível fazer uma aproximação do tipo:

$$\frac{e^t}{1+e^t} = \begin{cases} 1 & \text{se } t > 3 \\ \frac{1}{2} + \frac{1}{6}t & \text{se } |t| \leq 3 \\ 0 & \text{se } t < -3 \end{cases} \qquad (3.d.30)$$

tendo como erro máximo 0,07. Ao se aplicar (3.d.30) às equações de verossimilhança (3.d.25) quando $|a_i\ \hat{\beta}| < 3$ para todos os i, pode-se reescrever (3.d.23) como:

$$\frac{\partial\ L(\underset{\sim}{\beta})}{\partial\ \beta_s} = \sum_{i=1}^{n} a_{is}\ y_i - \sum_{i=1}^{n} a_{is}\left(\frac{1}{2} + \frac{1}{6}\ \underset{\sim}{a_i}\ \underset{\sim}{\beta}\right)$$

igualando a zero, resulta:

$$\frac{1}{6} \sum_{i=1}^{n} a_{is} a_{it} \hat{\beta}_t \simeq \sum_{i=1}^{n} a_{is} (y_i - \frac{1}{2})$$

No caso particular do modelo, para a situação I $\lambda_i = \alpha + \beta(x_i - \bar{x})$, tem-se os estimadores:

$$\hat{\alpha} \simeq \frac{6}{n} \sum_{i=1}^{n} (y_i - \frac{1}{2})$$

e

$$\hat{\beta} \simeq 6 \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Estimados os parâmetros e sua matriz de variância-covariância, podem-se obter os intervalos de confiança para cada parâmetro β_s, a um nível de confiança $(1-\alpha)$. Então, esse intervalo é expresso por:

$$(\hat{\beta}_s - Z_{\frac{\alpha}{2}} \hat{\sigma}_s ; \hat{\beta}_s + Z_{\frac{\alpha}{2}} \hat{\sigma}_s) \qquad (3.d.31)$$

onde

$Z_{\frac{\alpha}{2}}$ é o ponto limite da distribuição normal, $\Phi(-Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

e $\hat{\sigma}_s$ é o desvio padrão estimado da variável $\hat{\beta}_s$.

Uma maneira alternativa de encontrar região de confiança aproximada para alguns parâmetros particulares β_s é

tomar um subespaço q-dimensional B*, onde somente uns q parâmetros especificados são não zero. Assim, o logaritmo da função de verossimilhança maximizada é:

$$L\ (\hat{\underset{\sim}{\beta}}) - L\ (\hat{\underset{\sim}{\beta}}^*\ ;\ \hat{\underset{\sim}{\beta}}^* \in B^*) \qquad (3.d.32)$$

onde o segundo elemento é o logaritmo da função de verossimilhança maximizada sujeita à restrição anterior. Sob a condição da hipótese citada ser verdadeira, (3.d.32) é distribuído assintoticamente como $\frac{1}{2}\ \chi^2$ com (p-q) graus de liberdade.

Exemplifica-se para um particular parâmetro β_s, sob a hipótese Ho: $\beta_s = \beta_s^{(o)}$

isto é, maximizando somente com relação aos outros (p-1) componentes de $\underset{\sim}{\beta}$. Assim, a região de confiança é o conjunto definido por:

$$C = \{\beta_s^{(o)}\ ;\ L\ (\hat{\underset{\sim}{\beta}}) - L\ (\hat{\underset{\sim}{\beta}}^*\ ;\ \beta_s^* = \beta_s^{(o)}) \leq \frac{1}{2}\ \chi^2_{(1);\frac{\alpha}{2}}\} \} \qquad (3.d.33)$$

onde $\chi^2_{(1);\frac{\alpha}{2}}$ é o ponto limite superior para $\frac{\alpha}{2}$, da distribuição qui-quadrado com 1 grau de liberdade.

As regiões (3.d.31) e (3.d.33) são assintoticamente equivalentes, porém há uma certa vantagem em utilizar a segunda, visto que independe de uma particular parametrização adotada.

Foram citados aqui alguns processos de estimação dos parâmetros de regressão. Porém, na prática o problema prin-

cipal é o de maximização de (3.d.22). Desde que $\hat{\underset{\sim}{\beta}}$ seja encontrado, os passos seguintes para a estimação da região de confiança, tornam-se uma seqüência de rotina.

Em um par especificado de elementos em observação, utilizando os subscritos 1 para os casos e 2 para os controles, podem-se obter as probabilidades condicionais ao número de expostos no par, isto é, $F_1+F_2 = f$, $f = 0,1\ 2$.

Para o caso de pares discordantes: $F_1 + F_2 = 1$, tem-se:

$$\Pr\{F_1=f_1/F_1+F_2=1,\underset{\sim}{Z}_1,W_1,Z_2,W_2\} = \frac{\exp\{[\alpha+(\underset{\sim}{Z}_1'-Z_2')\underset{\sim}{\beta}+W_1'\gamma]f_1\}}{\{1+\exp[\alpha+(Z_1'-Z_2')\underset{\sim}{\beta}+W_1'\gamma]\}}$$

com $f_1 = 0$ ou 1. (3.d.34)

Para o caso de pares concordantes, $F_1+F_2 = 0$ ou $F_1+F_2 = 2$, a probabilidade condicional é 1 ou zero, isto é:

$$\Pr\{F_1 = f_1 \ / \ F_1 + F_2 = 0, \underset{\sim}{Z}_1, \underset{\sim}{W}_1, \underset{\sim}{Z}_2, \underset{\sim}{W}_2\} = \begin{cases} 0 & \text{se } f_1 = 1 \\ 1 & \text{se } f_1 = 0 \end{cases}$$

$$\Pr\{F_1 = f_1 \ / F_1+F_2 = 2, \underset{\sim}{Z}_1,\underset{\sim}{W}_1,\underset{\sim}{Z}_2,\underset{\sim}{W}_2\} = \begin{cases} 0 & \text{se } f_1 = 0 \\ 1 & \text{se } f_1 = 1 \end{cases}$$

Assim a função de verossimilhança condicional, para casos expostos, dado o número de expostos em cada par, é o produto das expressões (3.d.34) nos pares expostos discordantes ($F_1 + F_2 = 1$). Para eliminar o *parâmetro de distúrbio* ("nuisance") δ_i, utiliza-se a distribuição condicionada ao número de expostos em cada par, o que resulta em uma probabilidade lo

gística binária (3.d.34), análoga a (3.d.17), de cada par exposto discordante. E, para a inferência de (3.d.19), pode-se utilizar o método de máxima verossimilhança descrito por Cox (1970).

Note-se que, como $\underset{\sim}{Z}_1$ e $\underset{\sim}{Z}_2$ em (3.d.34) aparecem sob a forma de $(\underset{\sim}{Z}_1' - \underset{\sim}{Z}_2')$, as variáveis usadas no pareamento não poderão ser estudadas como fatores interferentes, na análise por pares. Contudo, as variáveis de pareamento poderão ser estudadas como fatores efeito-modificantes, pois $\underset{\sim}{W}_1$ e $\underset{\sim}{W}_2$ são expressas somente através de $\underset{\sim}{W}_1'$.

Prentice (1976), a fim de ilustrar o seu método, apresentou o estudo da influência da exposição a estrógenos (F) no câncer do endométrio (D) na pós-menopausa. Os casos foram pareados, segundo ano de diagnóstico e idade no ano do diagnóstico, e os controles eram mulheres com outros tipos de câncer ginecológico. As variáveis intervenientes utilizadas, além das pacientes do tipo A (0) e das do tipo B (1), foram a hipertensão (presença = 1 , ausência = 0) e a obesidade (sim =1, não = 0).

O modelo expresso por (3.d.17) foi aplicado aos 486 elementos, sem guardar o pareamento, pois os membros pareados não eram intrinsecamente similares com respeito a outras variáveis que não aquelas indicadas como variáveis de pareamento. Considerou-se o vetor $\underset{\sim}{Z}$ consistindo de 5 componentes:

(i) idade no ano do diagnóstico

(ii) ano do diagnóstico

(iii) pacientes do tipo A ou do tipo B

(iv) hipertensão

(v) obesidade

Por conveniência computacional e para obter ortogonalidade no vetor $\underset{\sim}{Z}$, 64 foi subtraído da idade e 1967 foi subtraído do ano do diagnóstico.

As estimativas de máxima verossimilhança com seus respectivos desvios-padrão assintóticos, correspondendo ao modelo (3.d.17) ou, mais explicitamente, (3.d.13) são apresentadas na TABELA 3.d.H, na página seguinte. Desta, extrai-se a "razão dos produtos cruzados" estimada, que é:

$$\hat{\Omega} = \exp(\hat{\alpha}) = \exp(2,166) \tilde{=} 8,72$$

e um teste para o log da "razão dos produtos cruzados" (ou $\Omega = 1$) é significante.

Ano do diagnóstico, idade no ano do diagnóstico, paciente do tipo B e obesidade são todos incluídos na obtenção da "RPC" e suas exclusões do modelo podem "viciar" a estimativa. O logaritmo da verossimilhança maximizada tem valor -221,51.

Tomando todas as cinco variáveis intervenientes, também como efeito-modificantes, isto é, considerando $\underset{\sim}{Z} = \underset{\sim}{W}$ em (3.d.17) obtém-se a (3.d.15), o que leva ao logaritmo da

verossimilhança maximizada -212,73.

TABELA 3.d.H Ajustamento do modelo (3.d.13) aos dados de câncer do endométrio por máxima verossimilhança

VARIÁVEL	$\hat{\alpha}$	Desvio Padrão($\hat{\alpha}$)
DOENÇA (CASOS-CONTROLES)	2,166	0,262
VARIÁVEL (Z_j)	**$\hat{\beta}_j$**	**Desvio-padrão($\hat{\beta}_j$)**
ANO DO DIAGNÓSTICO (Z_1)	0,228	0,038
IDADE AO DIAGNÓSTICO (Z_2)	-0,085	0,016
PACIENTE DO TIPO B (Z_3)	-1,104	0,237
HIPERTENSÃO (Z_4)	-0,337	0,307
OBESIDADE (Z_5)	-0,851	0,253

FONTE: Prentice (1976)

Com os dois valores obtidos do log da verossimilhança maximizada, pode-se utilizar a razão de verossimilhança para testar a hipótese de inexistência do fator efeito-modificante:

$$\chi^2_o = -2\ (-221,51 + 212,73) = 17,56$$

que, comparado com $\chi^2_{(5);0,05} = 11,07$, conclui-se ser significante, dando uma evidência da existência do fator efeito-modificante.

Os estimadores de máxima-verossimilhança de (3.d.15) com seus respectivos desvios-padrão são apresentados na TABELA 3.d.I.

TABELA 3.d.I Ajustamento do modelo (3.d.15) aos dados de câncer do endométrio por máxima verossimilhança

VARIÁVEL	$\hat{\alpha}$		DESVIO PADRÃO $(\hat{\alpha})$	
DOENÇA (CASOS — CONTROLES)	2,292		0,464	
VARIÁVEL (Z_j)	$\hat{\beta}_j$	DESVIO PADRÃO $(\hat{\beta}_j)$	$\hat{\gamma}_j$	DESVIO PADRÃO $(\hat{\gamma}_j)$
ANO DO DIAGNÓSTICO (Z_1)	0,132	0,057	0,194	0,080
IDADE AO DIAGNÓSTICO (Z_2)	-0,085	0,024	-0,012	0,033
PACIENTE DO TIPO B (Z_3)	-1,413	0,361	0,609	0,490
HIPERTENSÃO (Z_4)	0,564	0,456	-1,549	0,611
OBESIDADE (Z_5)	-0,471	0,392	-0,771	0,522

FONTE: Prentice (1976)

A "razão dos produtos cruzados", $\Omega = \exp(\alpha + \underset{\sim}{Z}'\underset{\sim}{\gamma})$, depende, aparentemente, do ano do diagnóstico, hipertensão e obesidade, pois $\hat{\gamma}_1 = 0,194$, $\hat{\gamma}_4 = 1,549$ e $\hat{\gamma}_5 = 0,771$, são significantes sob a hipótese nula, Ho: $\gamma_1 = \gamma_4 = \gamma_5 = 0$.

Assim a estimativa obtida da "RPC" pela TABELA 3.d.I

$$\hat{\Omega} = \exp(\hat{\alpha}) = \exp(2,292) = 9,89$$

é a estimativa para pessoas não hipertensas, não obesas em 1967. Deste resultado, partir para outros valores de variável interveniente é imediato. Assim, a "RPC" estimada de uma mulher obesa, não hipertensa em 1967 é:

$\hat{\Omega}_o = \exp(\hat{\alpha}+\hat{\gamma}_5) = 9,89 \times \exp(-0,771) = 4,57$, associada com exposição a estrógeno.

Os riscos são crescentes com o tempo, pois o coeficiente $\hat{\gamma}_1 = 0,194$ é positivo. Ao mesmo tempo, pode-se concluir que a "RPC" é elevada numa mulher não possuidora de fatores como hipertensão e obesidade, fatores estes normalmente considerados de risco quando associados com incidência de câncer do endométrio.

Hipertensão é um fator efeito-modificante, pelos resultados da TABELA 3.d.I, mas parece não ser fator interferente, ao passo que idade é associada com exposição a estrógenos, sem ser efeito-modificante. Nestas situações, Miettinen (1974) cita que realmente a idade pode afetar a "RPC" porém às vezes pode não aparecer diretamente, desde que outras variáveis que se relacionem com idade sejam incluídas simultaneamente no modelo como fatores efeito-modificantes.

Para justificar isso, o modelo (3.d.17) foi aplicado como na TABELA 3.d.I, excluindo hipertensão e obesidade, que são variáveis correlacionadas com a idade, do vetor $\underline{W}$ denotador do fator efeito-modificante. O estimador de máxima verossimilhança, $\hat{\gamma}_2$, coeficiente para idade, resultou em -0,012 com desvio-padrão 0,033, o que leva a concluir que idade não parece ser um fator efeito-modificante. Isso se verifica mesmo que não esteja destituída de variáveis a ela corre

lacionadas.

Foi ilustrado também o caso em que se considera vários tipos de controles, isto é, substituir αD por $\sum_{j=1}^{\ell} \alpha_j D_j$ em (3.d.17). Para o caso, verificou-se que das 243 pacientes controles, 153 tinham câncer cervical, 77 tinham câncer ovariano e 13 câncer vulvar. Assim, αD foi trocado por $\alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3$.

Utilizando as mesmas variáveis, efeito-modificante e interferente, como na TABELA 3.d.I, obteve-se:

$$\hat{\alpha}_1 = 2,624;\ \hat{\alpha}_2 = 2,149 \text{ e } \hat{\alpha}_3 = 1,320,$$

com os desvios-padrão 0,513, 0,527 e 0,727 correspondentes e onde os subscritos 1, 2 e 3 referem-se a controles cervical, ovariano e vulvar, respectivamente.

Aqui, o logaritmo da verossimilhança maximizada resulta em -210,77 e, comparando com o anterior, -212,73, obtém-se um $\chi^2_o = -2\ (-212,73+210,77) = 3,92$.

Interessante aqui é que o teste de hipótese de igualdade da "RPC" global sobre os vários tipos de controles não detecta diferença ao nível de significância de 10%, $\chi^2_{(2);0,10} = 4,605$, embora o valor de $\hat{\alpha}_3$ baseado na comparação dos casos ao controle câncer vulvar seja completamente diferente dos outros.

Este aspecto da análise sugere estudos adicionais de

uma possível associação entre exposição a estrógenos e câncer vulvar.

Como este estudo utilizou um planejamento pareado, ilustrou-se a aplicação do modelo (3.d.19), considerando ainda as mesmas variáveis efeito-modificante e interferentes como na Tabela 3.d.I. As variáveis utilizadas para o pareamento, idade no ano do diagnóstico e ano do diagnóstico, foram excluídas dos componentes do vetor indicador do fator interferente , $\underline{Z}$. É fácil perceber que o modelo (3.d.19) é mais geral que (3.d.17) e, portanto, permite fazer uma melhor inferência, apesar de que há uma perda considerável da eficiência pelo fato de eliminar os parâmetros δi, $i = 1,2,...n$.

Essa análise baseia-se em 486 observações onde 182 são elementos expostos ao fator F (denotados por 1) e 304 são não expostos (denotados por 0). Agora, ao tomarmos para análise os casos de pares discordantes como em (3.d.22) este número chega a 110, dos quais 10 são controles(0) expostos e 100 são casos(1) expostos. , o que leva a uma estimativa da "RPC"

$\hat{\Omega}$=100/10=10, que foi ajustada por ano do diagnóstico e idade no ano do diagnóstico.

A análise de dados pareados, apesar de partir de um conjunto razoavelmente grande, fica restrita a 10 observações somente, mas estas fornecem as informações adequadas para a obtenção de resultados significativos do estudo, como a"RPC"

consideravelmente grande, maior que um e que cresce com os anos. Há, porém, uma restrição imposta para um estudo preciso do efeito-modificante e interferente, como se pode observar comparando os desvios-padrão apresentados na Tabela 3.d.J, da página seguinte, que são consideravelmente maiores que aqueles da Tabela 3.d.I.

Prentice (1976) faz ainda alguns comentários acerca dos modelos expostos e analisados, tais como a generalização dos modelos expressos por (3.d.17) e (3.d.19):

(i) uma generalização é no sentido de que (3.d.19) pode também ser aplicado mesmo quando vários casos e controles estão agrupados em um estrato, sem necessitar um pareamento um a um, embora os cálculos para estimação sejam mais complicados;

(ii) outra, é que o modelo logístico de Mantel (1966), para respostas multinomiais, pode ser usado na generalização de (3.d.17) e (3.d.19), quando a exposição é em vários níveis.

3.e. CONSIDERAÇÕES FINAIS

Com esse trabalho, procuramos descrever um meio abrangente de analisar e estimar o Risco Relativo nos estudos epidemiológicos do tipo caso-controle, quer pela inclusão de fatores utilizados no pareamento ou na subclassificação, quer pela abordagem incluindo o conceito de fator interveni

TABELA 3.d.J Ajustamento do modelo (3.d.17) aos dados de câncer do endométrio por máxima verossimilhança.

Variável		$\hat{\alpha}$		Desvio-padrão ($\hat{\alpha}$)	
Doenças (Casos – Controles)		1,824		0,757	
Variável (Z_j)	$\hat{\beta}_j$	Desvio-padrão($\hat{\beta}_j$)	$\hat{\gamma}_j$	Desvio-padrão($\hat{\gamma}_j$)	
Ano do diagnóstico (Z_1)			0,226	0,120	
Idade ao diagnóstico (Z_2)			-0,079	0,052	
Grupo do tipo B (Z_3)	-0,097	0,448	0,254	1,197	
Hipertensão (Z_4)	-0,279	0,994	0,146	1,360	
Obesidade (Z_5)	-0,858	0,814	0,276	1,050	

Fonte: Prentice (1976)

ente.

Os resultados desta metodologia apresentados aqui estão baseados, essencialmente, nos trabalhos de Mantel e Haenszel (1959), Cox (1970), Zelen (1971) e Fleiss (1973), com aplicações de Breslow (1976) e Prentice (1976). Embora essas técnicas de estimação do Risco Relativo em estudos caso-controle por meio do modelo log-linear proposto por Zelen (1971) tenham sido utilizadas, estão sujeitas a críticas, conforme citam Gehan (1972) e Halperin et al. (1977). Assim, sentimos que deverão ser objeto de pesquisas detalhadas os seguintes tópicos:

(i) como contornar o problema da distribuição condicional do tipo apresentado por (3.b.5), sob a hipótese de (3.d.3) apresentado por Zelen (1971);

(ii) abordagem da problemática através de procedimentos alternativos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] - Armitage,P. (1966), The chi-square test for heterogeneity of proportions after adjustment for stratification, *J.Roy.Statist. Soc., B. 28*: 150-163.

[2] - Bartlett,M.S. (1935), Contingency table interations, *J. Roy. Statist. Soc., B. 11*: 248-252.

[3] - Beale,E.M.L. (1967), "Numerical Methods" in nonlinear programming, *J. Abadie - N.Holland*, Amsterdam, (§6.4).

[4] - Berkson,J. (1958), Smoking and lung cancer. Some observations on two recent reports, *J.Am.Statist.Ass., 53*: 28-38.

[5] - Breslow,N. (1976), Regression analysis of the log odds ratio: A method for retrospective studies, *Biometrics, 32*: 409-416.

[6] - Breslow,L., Hoaglin,L., Rasmussen,G. and Abrams,H.K. (1954), Occupations and cigarrette smoking as factors in lung cancer, *Amer.J.Pub.Health, 44*: 171-181.

[7] - Cochran, W.G. (1950), The comparison of percentages in matched samples, *Biometrika, 37*: 256-266.

[8] - Cochran, W.G. (1954), Some of strengthening the comon χ^2 tests, *Biometrics, 10*: 417-451.

[9] - Cornfield, J. (1951), A method of estimating comparative rates from clinical data. Applications to cancer of the lung,breast and cervix, *J.Nat.Cancer Inst., 11*: 1269-1275.

[10] - Cornfield,J. (1956), A statistical problem arising from retrospective studies, *Proc. Third Berkeley Symp. - Math. Statist. Prob., 4*: 135-148.

[11] - Cornfield,J. and Haenszel,W. (1960), Some Aspects of retrospective studies, *J.Chron.Dis., 11*: 523-534.

[12] - Cox, D.R. (1966), A simple example of a comparison involving quantal data, *Biometrika*, *53*: 215-220.

[13] - Cox, D.R. (1970), The analysis of binary data, *Methuen*, London.

[14] - Cramér, H. (1946), Mathematical methods of statistics, *Princeton University Press*.

[15] - Dawber, T.R., Moore, F.E. and Mann, G.V. (1957), Coronary heart disease in the Framingham study, *Amer.J.Pub.Health*, *47(II)*: 4-24.

[16] - Doll, R. and Hill, A.B. (1952) - A study of the etiology of carcinoma of the lung, *Brit.Med.J.*, *2*: 1271-1286.

[17] - Dorn, H.F. (1954), The relationship of cancer of the lung and the use of tabacco, *Amer. Statistician*, *8*: 7-13.

[18] - Draper, N.R. and Smith, H. (1966), Applied Regression analysis, *Wiley*, New York, (§6.4).

[19] - Fisher, L. and Patil, R. (1974), Mathching and unrelatedness, *Amer. J.Epid.*, *100(5)*: 347-349.

[20] - Fleiss, J.L. (1970), On the asserted invariance of the odds ratio, *Brit.J.Prev.Soc.Med.*, *24*: 45-46.

[21] - Fleiss, J.L. (1973), Statistical methods for rates and proportions, *Wiley*, New York.

[22] - Gart, J.J. (1962), On the combination of relative risks, *Biometrics*, *18*: 601-610.

[23] - Gart, J.J. (1970), Point and interval estimation of the comon odds ratio in the combination of 2×2 tables with fixed marginals, *Biometrika*, *57*: 471-475.

[24] - Gart, J.J. (1971), The comparison of proportions: a review of significance tests, confidence intervals and adjustments of stratification, *Int.Statist.Rev.*, *39*: 148-149.

[25] - Gehan, E.A. (1972), Letter to the editor, *Biometrics*, *28*: 239-243.

[26] - Goodman, L.A. and Kruskal, W.H. (1954), Measures of association for cross classifications, *J.Am.Statist.Ass.*, *49*: 732-764.

[27] - Goodman, L.A. and Kruskal, W.H. (1959), Measures of association for cross classifications, II: Further discussion and references, *J.Am.Statist.Ass.*, *54*: 123-163.

[28] - Guedes,M.B.L.S. (1976), A combinação de riscos relativos na medida do grau comum de associação, *Dissertação de Mestrado, Faculdade de Medicina de Ribeirão Preto, U.S.P.*, São Paulo.

[29] - Haenzel,E., Shimkin,M.B. and Mantel, N. (1958), A retrospective study of lung cancer in women, *J.Nat.Cancer Inst.,21:* 825-842.

[30] - Halperin,M., Ware,J.H., Byar,D.P., Mantel,N., Brown,C.C., Koziol, J., Gail,M. and Green,S.B. (1977), Testing for interation in a I×J×K contingency table, *Biometrika, 64(2):* 271-275.

[31] - Hannan,J. and Harkness,W.L. (1963), Normal approximation to the distribution of two-independent binomials, conditional on fixed sum, *Ann.Math.Statist., 34:* 1593-1595.

[32] - Kneale,G.W. (1971), Problems arising in estimating from retrospective survey data the latent period of juvenile cancers initiaded by obstetric radiography, *Biometrics, 27:* 563-590.

[33] - MacMahon,B. and Pugh,T.F. (1970), Epidemiology - Principles and methods, *Little, Brown and Company*, Boston.

[34] - Mantel,N. (1963), Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel Procedure,*J.Am.Statist.Ass., 58:* 690-700.

[35] - Mantel,N. (1966), Models for complex contingency tables and polychotomous dosage responses curves, *Biometrics, 22:* 83-95.

[36] - Mantel,N. and Haenszel,W. (1959), Statistical aspects of the analysis of data from restrospective studies of disease, *J.Nat. Cancer Inst., 22:* 719-748.

[37] - Medical Research Council (1948), Streptomycin treatment of pulmonary tuberculosis, *Brit.Med.J., 2:* 769-782.

[38] - Miettinen,O. (1974), Confounding and effect-modification, *Amer.J. Epid., 100(5):* 350-353.

[39] - Naylor,A.F. (1967), Small sample considerations in combining 2×2 tables, *Biometrics, 23:* 407-409.

[40] - Norton,H.W. (1945), Calculation of qui-square for complex contingency tables, *J.Am.Statist.Ass., 40:* 251-258.

[41] - Plackett,R.L. (1974), The analysis of categorical data, *Griffin*, London.

[42] - Prentice,R.L. (1976), Use of the logistic model in retrospective studies, *Biometrics*, *32*: 599-606.

[43] - Sheehe,P.R. (1966), Combination of log relative risk in retrospective studies of disease, *Amer.J.Pub.Health*, *56*: 1745-1750.

[44] - Sheps,M.C. (1958), Shall we count the living or the dead? *N.Engl. J.Med.*, *259*: 1210-1214.

[45] - Sheps,M.C. (1959), An examination of some methods of comparing several rates or proportions, *Biometrics*, *15*: 87-97.

[46] - Sheps,M.C. (1961), Marriage and mortality, *Amer.J.Pub.Health*, *51*: 547-555.

[47] - Smith,D.C.,Prentice,R.L., Thompson,D.J. and Herrmann,W.L. (1975), Exogenous estrogen and endometrial carcinoma, *N.Engl.J.Med.*, *293*: 1164-1167.

[48] - Stewart,A. and Kneallе,G.W. (1970), Age-distribution of cancers caused by obstetric X-rays and their relevance to cancer latent periods, *Lancet*, *ii*: 4-8.

[49] - Woolf,B. (1955), On estimating the relation between blood group and disease, *Ann.Human.Genet.*, *19*: 251-253.

[50] - Worcester,J. (1964), Matched samples in epidemiologic studies, *Biometrics*, *20*: 840-848.

[51] - Wynder,E.L. and Cornfield,J. (1953), Cancer of the lung in physicians, *N.Engl.J.Med.*, *248*: 441-444.

[52] - Yates,F. (1955), The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments, *Biometrika*, *42*: 382-403.

[53] - Zelen,M. (1971), The analysis of several 2×2 contingency tables, *Biometrika*, *58*: 129-137.

APENDICE I

SHEPS (1959) considerou alguns modelos de estudos epidemiológicos onde procura descrever o risco a um fator por meio de um modelo condicionado ao conhecimento "a priori" de algum fato que possa contribuir na sua descrição. Ela apresentou o risco de fumantes e não fumantes, de ter câncer de pulmão, como foi descrito na Secção 2.d.II, através da Tabela 2.d.A. Obteve ainda os estimadores de máxima-verossimilhança (2.d.12), e neste apêndice será apresentado a matriz de variância e covariância dos estimadores de p_o e p_f.

É de conhecimento geral que a matriz de variância(Σp) é o inverso da matriz de informações (I_p), onde cada elemento é obtido por:

$$-E\left[\frac{\partial^2 L}{\partial p_i \partial p_j}\right] \qquad i,j = 1,2,\ldots,k \qquad \text{(A.I.1)}$$

sendo L, o logarítmo da função de verossimilhança-

No nosso caso particupar, ocorre que só temos dois parâmetros a serem estimados: p_o e p_f. Assim:

$$L = \ell n \binom{n_o}{x_o}\binom{n_f}{x_f} + x_o \ell n\, p_o + (n_o - x_o)\ell n(1-p_o) +$$

$$+ x \ell n\,[p_o + (1-p_o)p\,] + (n_f - x_f)\ell n[(1-p_o)(1-p\,)]$$

Utilizando-se das expressões (2.d.10) e (2.d.11) temos:

$$\frac{\partial^2 L}{(\partial p_o)^2} = \frac{x_o}{p_o^2} - \frac{n_f + n_o - x_f - x_o}{q_o^2} - \frac{x_f(1-p_f)^2}{(1-q_o q_f)^2} \qquad \text{(A.I.2)}$$

$$\frac{\partial^2 L}{\partial p_o \, \partial p_f} = \frac{xf}{(1-q_o q_f)^2} \qquad (A.I.3)$$

$$\frac{\partial^2 L}{(\partial p_f)^2} = -\frac{x_f \, (1-p_o)^2}{(1 - q_o q_f)^2} - \frac{nf-xf}{q_f^2} \qquad (A.I.4)$$

Substituindo (A.I.2), (A.I.3) e (A.I.4) em (A.I.1) e considerando

$$E\left[X_o\right] = n_o \, p_o$$

e resultam:

$$E\left[X_f\right] = n_f \; p_f(1-p_o) + p_o$$

$$- E\left[\frac{\partial^2 L}{(\partial p_o)^2}\right] = \frac{n_o+n_f \, q_f p_o}{p_o q_o} + \frac{n_f \, q_f^2}{1-q_o \, q_f}$$

$$- E\left[\frac{\partial_i^2 L}{\partial p_o \, \partial p_f}\right] = \frac{n_f}{1-q_o q_f}$$

$$- E\left[\frac{\partial^2 L}{(\partial p_f)^2}\right] = \frac{n_f \, q_o}{q_f(1-q_o q_f)}$$

Assim a matriz de informação, $\underset{\sim}{I}_p$, será:

$$\underset{\sim}{I}_p = \begin{bmatrix} \dfrac{n_o+n_f q_f p_o}{p_o q_o} + \dfrac{n_f q_f^2}{1-qoqf} & \dfrac{n_f}{1-q_o q_f} \\ \dfrac{n_f}{1-qoq} & \dfrac{n_f \, q_o}{q_f(1-q_o q_f)} \end{bmatrix}$$

Portanto: $\underset{\sim}{\Sigma}_p = \underset{\sim}{I}_p^{-1} = \dfrac{1}{\det \underset{\sim}{I}_p} \cdot \left[\text{cof } \underset{\sim}{I}_p' \right]$

onde

$$\det \underset{\sim}{I}_p = n_f \, n_o \Big/ \left[p_o \, q \, (1 - q_o q_f) \right]$$

$$\underset{\sim}{\Sigma}_p = \frac{1}{\det \underset{\sim}{I}_p} \begin{bmatrix} \dfrac{n_f \, q_o}{q_f (1 - q_o q_f)} & \dfrac{- n_f}{1 - q_o q_f} \\ \dfrac{- n}{1 - q_o q_f} & \dfrac{n_o + n_f q_f q_o}{p_o q_o} + \dfrac{n_f \, q_f^2}{1 - q_o q_f} \end{bmatrix}$$

ou

$$\underset{\sim}{\Sigma}_p = \begin{bmatrix} \dfrac{p_o q_o}{n_o} & - \dfrac{p_o q_f}{n_o} \\ - \dfrac{p_o q_f}{n_o} & \dfrac{q_f}{q_o} \left(\dfrac{1 - q_o q_f}{n_f} + \dfrac{p_o p_f}{n_o} \right) \end{bmatrix}$$

Finalmente:

$$\text{Var } (\hat{p}_o) = \frac{p_o q_o}{n_o}$$

$$\text{Cov } (\hat{p}_o \, ; \, \hat{p}_f) = \frac{- p_o q_f}{n_o}$$

$$\text{Var } (\hat{p}_f) = \frac{q_f}{q_o} \left(\frac{1 - q_o q_f}{n_f} + \frac{p_o p_f}{n_o} \right)$$

APENDICE II

Neste segundo apêndice será apresentada, segundo o desenvolvimento de SHEPS(1959), a matriz de variâncias e covariâncias ($\sum q$) dos estimadores dos componentes de sobrevida dos fumantes, em várias categorias de fumantes.

Partindo da Tabela 2.d.D, da Secção 2d. e considerando o logaritmo da função de verosimilhança, sem as constantes, teremos:

$$L = x_1 \ln q_1 + (n_1 - x_1) \ln(1-q_1) + x_2 \ln q_1 q_2 + (n_2 - x_2) \ln(1-q_1 q_2) + x_3 \ln q_1 q_2 q_3 +$$

$$+ (n_3 - x_3) \ln (1-q_1 q_2 q_3).$$

Com as derivadas de primeira ordem:

$$\frac{\partial L}{\partial q_1} = \frac{x_1}{q_1} - \frac{n_1 - x_1}{1-q_1} + \frac{x_2}{q_1} - \frac{(n_2 - x_2) q_2}{1-q_1 q_2} + \frac{x_3}{q_1} - \frac{(n_3 - x_3)\, q_2 q_3}{1-q_1 q_2 q_3}$$

$$\frac{\partial L}{\partial q_2} = \frac{x_2}{q_2} - \frac{(n_2 - x_2) q_1}{1-q_1 q_2} + \frac{x_3}{q_2} - \frac{(n_3 - x_3)\, q_1 q_3}{1 - q_1 q_2 q_3}$$

$$\frac{\partial L}{\partial q_3} = \frac{x_3}{q_3} - \frac{(n_3 - x_3) q_1 q_2}{1-q_1 q_2 q_3}$$

E as derivadas de segunda ordem:

$$\frac{\partial^2 L}{(\partial q_1)^2} = -\frac{(x_1 + x_2 + x_3)}{q_1^2} - \frac{(n_1 - x_1)}{(1-q_1)^2} - \frac{(n_2 - x_2) q_2^2}{(1-q_1 q_2)^2} - \frac{(n_3 - x_3) q_2^2 q_3^2}{(1-q_1 q_2 q_3)^2} \qquad (A.II.1)$$

$$\frac{\partial^2 L}{\partial q_1 \partial q_2} = -\frac{(n_2 - x_2)}{(1 - q_1 q_2)^2} - \frac{(n_3 - x_3)\, q_3}{(1-q_1 q_2 q_3)^2} \qquad (A.II.2)$$

$$\frac{\partial^2 L}{\partial q_1 \partial q_3} = \frac{-(n_3 - x_3) q_2}{(1 - q_1 q_2 q_3)^2} \qquad \text{(A.II.3)}$$

$$\frac{\partial^2 L}{(\partial q_2)^2} = \frac{-(x_2 + x_3)}{q_2^2} - \frac{(n_2 - x_2) q_1^2}{(1 - q_1 q_2)^2} - \frac{(n_3 - x_3)\, q_1^2 q_3^2}{(1 - q_1 q_2 q_3)^2} \qquad \text{(A.II.4)}$$

$$\frac{\partial^2 L}{\partial q_2 \partial q_3} = \frac{-(n_3 - x_3)\, q_1}{(1 - q_1 q_2 q_3)^2} \qquad \text{(A.II.5)}$$

$$\frac{\partial^2 L}{(\partial q_3)^2} = \frac{-x_3}{q_3^2} - \frac{(n_3 - x_3)\, q_1^2 q_2^2}{(1 - q_1 q_2 q_3)^2} \qquad \text{(A.II.6)}$$

Mostra-se também que:

$$E\left[X_1\right] = n_1 q_1; \quad E\left[X_2\right] = n_2 q_1 q_2 \quad \text{e} \quad E\left[X_3\right] = n_3 q_1 q_2 q_3. \qquad \text{(A.II.7)}$$

Igualando a zero as tres derivadas de primeira ordem e formando o sistema:

$$\frac{x_1 + x_2 + x_3}{q_1} - \frac{n_1 - x_1}{1 - q_1} - \frac{(n_2 - x_2) q_2}{1 - q_1 q_2} - \frac{(n_3 - x_3) q_2 q_3}{1 - q_1 q_2 q_3} = 0 \qquad \text{(A.II.8)}$$

$$\frac{x_2 + x_3}{q_2} - \frac{(n_2 - x_2) q_1}{1 - q_1 q_2} - \frac{(n_3 - x_3) q_2 q_3}{1 - q_1 q_2 q_3} = 0 \qquad \text{(A.II.9)}$$

$$\frac{x_3}{q_3} - \frac{(n_3 - x_3) q_1 q_2}{1 - q_1 q_2 q_3} = 0 \qquad \text{(A.II.10)}$$

De (A.II.10)

$$\frac{n_3 - x_3}{1 - q_1 q_2 q_3} = \frac{x_3}{q_1 q_2 q_3} \qquad \text{(A.II.11)}$$

De (A.II.9) utilizando (A.II.11), resulta:

$$\frac{n_2 - x_2}{1 - q_1 q_2} = \frac{x_2}{q_1 q_2} \qquad \text{(A.II.12)}$$

De (A.II.8), utilizando (A.II.11) e (A.II.12), resulta:

$$\hat{q}_1 = \frac{X_1}{n_1} \qquad \text{(A.II.13)}$$

Levando (A.II.13) em (A.II.12), resulta:

$$\hat{q}_2 = \frac{X_2 n_1}{X_1 n_2} \qquad \text{(A.II.14)}$$

Levando (A.II.13) e (A.II.14) em (A.II.11), resulta:

$$\hat{q}_3 = \frac{X_3 n_1 n_2}{X_1 X_2 n_3} \qquad \text{(A.II.15)}$$

De (A.II.13), conclui-se $E\left[X_1\right] = n_1 q_1$

De (A.II.14), conclui-se $E\left[X_2\right] = n_2 q_1 q_2$

De (A.II.15), conclui-se $E\left[X_3\right] = n_3 q_1 q_2 q_3$

Assim, levando em conta (A.II.7), pode-se obter a matriz de informação $\underset{\sim}{I}q$, cujos elementos são definidos por:

$$-E\left[\frac{\partial^2 L}{\partial q_i \partial q_j}\right] \quad \text{na posição } (i,j), \text{ onde } i,j=1,2,3.$$

Especificamente:

$$-E\left[\frac{\partial^2 L}{(\partial q_1)^2}\right] = \frac{n_1}{q_1(1-q_1)} + \frac{n_2 q_2}{q_1(1-q_1 q_2)} + \frac{n_3 q_2 q_3}{q_1(1-q_1 q_2 q_3)}$$

$$-E\left[\frac{\partial^2 L}{\partial q_1 \partial q_2}\right] = \frac{n_2}{1-q_1 q_2} + \frac{n_3 q_3}{1-q_1 q_2 q_3}$$

$$-E\left[\frac{\partial^2 L}{\partial q_1 \partial q_3}\right] = \frac{n_3 q_2}{1-q_1 q_2 q_3}$$

$$-E\left[\frac{\partial^2 L}{(\partial q_2)^2}\right] = \frac{n_2 q_1}{q_2(1-q_1 q_2)} + \frac{n_3 q_1 q_3}{q_2(1-q_1 q_2 q_3)}$$

$$-E\left[\frac{\partial^2 L}{\partial q_2 \partial q_3}\right] = \frac{n_3 q_1}{1-q_1 q_2 q_3}$$

$$-E\left[\frac{\partial^2 L}{(\partial q_3)^2}\right] = \frac{n_3 q_1 q_2}{q_3(1-q_1 q_2 q_3)}$$

O que resulta na matriz de informação, $\underset{\sim}{I}q$:

$$\underset{\sim}{I}_q = \begin{bmatrix} \dfrac{n_1}{q_1(1-q_1)} + \dfrac{n_2 q_2}{q_1(1-q_1q_2)} + \dfrac{n_3 q_2 q_3}{q_1(1-q_1q_2q_3)} & \dfrac{n_2}{1-q_1q_2} + \dfrac{n_3 q_3}{1-q_1q_2q_3} & \dfrac{n_3 q_2}{1-q_1q_2q_3} \\ \dfrac{n_2}{1-q_1q_2} + \dfrac{n_3 q_3}{1-q_1q_2q_3} & \dfrac{n_2 q_1}{q_2(1-q_1q_2)} + \dfrac{n_3 q_1 q_3}{q_2(1-q_1q_2q_3)} & \dfrac{n_3 q_1}{1-q_1q_2q_3} \\ \dfrac{n_3 q_2}{1-q_1q_2q_3} & \dfrac{n_3 q_1}{1-q_1q_2q_3} & \dfrac{n_3 q_1 q_2}{1-q_1q_2q_3} \end{bmatrix}$$

Sabe-se que a matriz de variância é obtida, invertendo-se $\underset{\sim}{I}_q$. Assim,

$$\underset{\sim}{\Sigma}_q = \underset{\sim}{I}_q^{-1} = \frac{1}{\det \underset{\sim}{I}_q} \left[\text{cof } \underset{\sim}{I}_q \right]$$

onde

$$\det \underset{\sim}{I}_q = n_1 n_2 n_3 q_1 \Big/ \left[q_3 (1-q_1)(1-q_1q_2)(1-q_1q_2q_3) \right]$$

Então

$$\underline{\Sigma}_Q = \frac{1}{\det \underline{I}_Q} \begin{bmatrix} \dfrac{n_2 n_3 q_1^2}{q_3(1-q_1q_2)(1-q_1q_2q_3)} & \dfrac{-n_2 n_3 q_1 q_2}{q_3(1-q_1q_2)(1-q_1q_2q_3)} & 0 \\ \dfrac{-n_2 n_3 q_1 q_2}{q_3(1-q_1q_2)(1-q_1q_2q_3)} & \dfrac{n_1 n_3 q_2(1-q_1q_2)+ n_2 n_3 q_2^2(1-q_1)}{q_3(1-q_1)(1-q_1q_2)(1-q_1q_2q_3)} & \dfrac{-n_1 n_3}{(1-q_1)(1-q_1q_2q_3)} \\ 0 & \dfrac{-n_1 n_3}{(1-q_1)(1-q_1q_2q_3)} & \dfrac{n_1(n_2+n_3q_3)}{q_2(1-q_1)(1-q_1q_2)} \end{bmatrix}$$

ou

$$\underline{\Sigma}_Q = \begin{bmatrix} \dfrac{q_1(1-q_1)}{n_1} & \dfrac{-q_2(1-q_1)}{n_1} & 0 \\ \dfrac{-q_2(1-q_1)}{n_2} & \dfrac{q_2(1-q_1q_2)}{n_2q_2} + \dfrac{q_2^2(1-q_1)}{n_1q_1} & \dfrac{-q_3(1-q_1q_2)}{n_2q_1} \\ 0 & \dfrac{-q_3(1-q_1q_2)}{n_2q_1} & \dfrac{q_3(1-q_1q_2q_3)(n_2+n_3q_3)}{n_2n_3q_1q_2} \end{bmatrix}$$

Consequentemente:

$$\text{Var}\ (\hat{q}_1) = \frac{q_1(1-q_1)}{n_1} \quad , \text{cov}\ (\hat{q}_1;\hat{q}_2) = \frac{-q_2(1-q_1)}{n_2}$$

$$\text{Var}\ (\hat{q}_2) = \frac{q_2(1-q_1q_2)}{n_2q_2} + \frac{q_2^2(1-q_1)}{n_1q_1} \quad , \text{cov}\ (\hat{q}_1;\hat{q}_3) = 0$$

$$\text{Var}\ (\hat{q}_3) = \frac{q_3(1-q_1q_2q_3)(n_2+n_3q_3)}{n_2n_3\ q_1q_2} \quad , \text{cov}(\hat{q}_2;\hat{q}_3) = \frac{-q_3(1-q_1q_2)}{n_2q_1}$$

APENDICE III

GART (1962) definiu um estimador do risco relativo, num esquema de estudo epidemiológico que resulta em uma tabela 2xk, pois é composta de informações de k pares de amostras binomiais, como foi apresentado na Secção 2.c.III. Observando-se a Tabela 1.b.A, definiu-se

$$\tilde{\Omega}_G = \frac{\sum_{j=1}^{k} A_j \sum_{j=1}^{k} D_j}{\sum_{j=1}^{k} B_j \sum_{j=1}^{k} C_j} \qquad (A.III.1)$$

que não é um estimador consistente.

Será mostrado agora que, somente num caso particular, é possível se verificar a consistência.

Utilizando as notações da Tabela 1.b.A e da Tabela 2.b.C , pode se definir:

(a)
$$\begin{aligned} A_j &= N_{1j}\ \hat{p}_{1j} \\ B_j &= N_{1j}\ (1-\hat{p}_{1j}) \\ C_j &= N_{2j}\ \hat{p}_{2j} \\ D_j &= N_{2j}\ (1-\hat{p}_{2j}) \end{aligned} \qquad \text{para } j = 1,2,\ldots,k \qquad (A.III.2)$$

e $N_{i\cdot} = \sum_{j=1}^{k} N_{ij}$, $i=1,2.$

(b) O caso assintótico: para grandes amostras (isto é, $N_{i\cdot} \to \infty, i=1,2$) e sob as seguintes condições:

$$\lim_{N_{i\cdot}\to\infty} \frac{N_{ij}}{N_{i\cdot}} = c_{ij} \quad \text{e} \quad \sum_{j=1}^{k} c_{ij} = 1 \quad \text{Para } i=1,2 \qquad (A.III.3)$$

Obtém-se, retomando (A.III.1) e utilizando (A.III.2):

$$\hat{\Omega}_G = \frac{\sum_{j=1}^{k} N_{1j}\ \hat{p}_{1j}\ \sum_{j=1}^{k} N_{2j}\ (1-\hat{p}_{2j})}{\sum_{j=1}^{k} N_{1j}\ (1-\hat{p}_{1j})\ \sum_{j=1}^{k} N_{2j}\ \hat{p}_{2j}}$$

$$\hat{\Omega}_G = \frac{\sum_{j=1}^{k} \frac{N_{1j}}{N_{1\cdot}}\ \hat{p}_{1j}\ \sum_{j=1}^{k} \frac{N_{2j}}{N_{2\cdot}}\ (1-\hat{p}_{2j})}{\sum_{j=1}^{k} \frac{N_{1j}}{N_{1\cdot}}\ (1-\hat{p}_{1j})\ \sum_{j=1}^{k} \frac{N_{2j}}{N_{2\cdot}}\ \hat{p}_{2j}} \qquad \text{(A.III.4)}$$

A expressão (A.III.4), sob as condições (A.III.3), converge em probabilidade (pela generalização de SLUTSKY, da lei fraca dos grandes números(*)) para:

$$\hat{\Omega}_G = \frac{\sum_{j=1}^{k} C_{1j}\ \hat{p}_{1j}\ \sum_{j=1}^{k} C_{2j}\ (1-\hat{p}_{2j})}{\sum_{j=1}^{k} C_{1j}\ (1-\hat{p}_{1j})\ \sum_{j=1}^{k} C_{2j}\ \hat{p}_{2j}} \qquad \text{(A.III.5)}$$

Que, de modo geral, não é igual ao estimador da "razão dos produtos cruzados" quando considerada a tabela formada com os valores totais:

$$\hat{\Omega} = \frac{\hat{p}_{1\cdot}\ (1-\hat{p}_{2\cdot})}{\hat{p}_{2\cdot}\ (1-\hat{p}_{1\cdot})} \qquad \text{(A.III.6)}$$

Porém, se em (A.III.5) for considerado o caso particular onde:

$p_{1j} = p_{1\cdot}$ e $p_{2j} = p_{2\cdot}$ para todos os $j=1,2,\ldots,k$.

o limite de $\hat{\Omega}_G$ será (A.III.6).

(*)Ver CRAMÉR (1946), p. 255.

Somente neste caso é que $\hat{\Omega}_G$ será um estimador consistente, razão pela qual no exemplo citado através dos dados da Tabela 2.c.A, onde os p_{ij} são muito diferentes, houve aquela incoerência. A heterogeneidade dos p_{ij} reflete - se fortemente na disparidade do estimador.