ANÁLISE DE DADOS CATEGÓRICOS

POR MODELOS LINEARES

CICÍLIA YUKO WADA

DISSERTAÇÃO APRESENTADA AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA UNIVERSIDADE DE SÃO PAULO

PARA A OBTENÇÃO DO GRAU

MESTRE EM ESTATÍSTICA

ORIENTADOR: Prof. Dr. CLÓVIS DE ARAÚJO PERES

SÃO PAULO, NOVEMBRO DE 1977

AGRADECIMENTOS

O presente trabalho é o resultado da colaboração que recebemos ao longo de sua elaboração por parte de professores e colegas de magistério. A eles, queremos apresentar publicamente os nossos sinceros agradecimentos; em especial,

- ao Prof.Dr. Clóvis de Araújo Peres, nosso orientador, não apenas pela orientação segura, dedicação ímpar e incentivo constante, como também pela amizade com que houve por bem honrar-nos;
- ao Prof.Dr. Walter P.Canton, pelas preciosas sugestões e adendos que muito enriqueceram nosso trabalho;
- ao Prof.Dr. Domingos Viggiani, Titular de Matemática e Estatística do "campus" de Marília, UNESP, de quem sempre recebemos, desde o início de nossa carreira, todo o apoio e confiança de que necessitávamos;
- ao Prof.Dr. Enzo Del Carratore, Titular de Língua Portuguesa do "campus" de Marília, UNESP, pela revisão dos originais, eliminando as falhas de quem, como nós, está pouco afeito às sutilezas da linguagem;

- ao Prof.Dr. Jubert Sanches Cibantos, Diretor da Faculdade de Educação, Filosofia, Ciências Sociais e da Documentação de Marília, UNESP; muito nos facilitou a realização deste trabalho;
- aos colegas do Departamento de Ciências do "campus" de Marília, UNESP, pela amizade e carinho com que nos cercaram no convívio de longos anos.

ÍNDICE

Capítulo I - Introdução . . . . . p. 1

Capítulo II - Método de GSK para análise de dados categóricos 4

2.1. Introdução 4

2.2. Soluções usuais 11

2.3. Solução de GSK 15

2.3.1. Estatística de Wald e sua adaptação a dados categóricos 16

2.3.2. Equivalência algébrica entre as estatísticas de Neyman e de Wald. 20

2.3.3. A teoria de GSK 23

Capítulo III - Aplicação a problemas de uma população.

3.1. Introdução 30

3.2. Hipóteses do tipo H0: $F(\underset{\sim}{\Pi}) = A\underset{\sim}{\Pi} = 0$ . . 33

3.2.1. Hipótese de homogeneidade das marginais em modelos mistos de ordem 2 34

3.2.2. Hipótese de homogeneidade das marginais em modelos mistos de ordem 3 39

3.2.3. Hipótese de simetria em tabelas de contingência dxd 43

3.3. Hipóteses do tipo H0: $F(\underset{\sim}{\Pi})=K\ln A\underset{\sim}{\Pi}=0$ . . . 46

3.3.1. Hipótese de independência em tabelas de contingência cxd 47

3.3.2. Hipóteses de independência e de não-interação em tabelas cxdxf 51

3.3.3. Hipótese de associação parcial sobre a diagonal principal 63

Capítulo IV - Aplicação a problemas de diversas populações.

4.1. Introdução 66

4.2. Hipótese de homogeneidade entre populações 68

4.3. Hipótese linear geral no modelo $F(\underset{\sim}{\Pi}) = X\beta$ 71

4.3.1. Caso $F(\underset{\sim}{\Pi}) = A\underset{\sim}{\Pi}$ 76

4.3.2. Caso $F(\underset{\sim}{\Pi}) = K \ln A\underset{\sim}{\Pi}$ 78

Capítulo V - Extensão de Forthofer e Koch.

5.1. Introdução 95

5.2. Funções compostas 96

5.3. Aplicações 98

5.3.1. Hipótese linear geral no modelo $F(\underset{\sim}{\Pi}) = X\beta$ 99

Capítulo VI - Considerações finais 114

Bibliografia 118

CAPÍTULO I

INTRODUÇÃO

A análise de dados categóricos, isto é, de dados que se apresentam como frequências com que indivíduos, objetos, são classificados em um certo número de categorias de atributos, frequências essas que seguem uma distribuição multinomial, tem sido objeto de estudo por parte de muitos pesquisadores, devido ao crescente uso em diversas áreas, tais como biologia, medicina, psicologia e sociologia.

Constitui objetivo principal do nosso trabalho apresentar o método para análise de dados categóricos baseado em modelos lineares, proposto por GRIZZLE, STARMER e KOCH em 1969. Eles propõem uma solução geral para diversos problemas surgidos em tabelas de classificação cruzada utilizando mínimos quadrados generalizados para calcular as estatísticas de WALD.

A justificativa para o uso da estatística de WALD para solução de certos problemas surgidos em tabelas de classificação cruzada foi baseada no trabalho de BHAPKAR (1966), que mostra equivalência entre a estatística χ_1^2 modificado,

devida a NEYMAN, e a estatística de WALD, adaptada a dados categóricos, na solução de problemas relatados como funções lineares e não lineares de probabilidade das caselas. Desse estudo, conclui-se que os testes estatísticos utilizados na solução dessas hipóteses pertencem à classe dos testes χ^2_1 modificado e, como foi verificado por NEYMAN em 1949, possuem ótimas propriedades assintóticas.

A apresentação do método geral de GRIZZLE e sua fundamentação teórica são feitas no Capítulo II.

GRIZZLE, STARMER e KOCH desenvolvem no seu artigo a aplicação de seu método para hipóteses que podem ser formuladas como funções lineares e logarítmicas das probabilidades das caselas, necessitando da construção de matrizes A e K e também de uma matriz de planejamento X, diferente da usual.

Nos Capítulos III e IV, procuramos aplicar o método de GSK a vários problemas relatados como funções lineares ou logarítmicas, tendo a preocupação de identificar as situações em que funções lineares ou logarítmicas são mais apropriadas. Para isso, utilizamos a linguagem "fator" e "resposta", proposta por BHAPKAR e KOCH (1968), na identificação das variáveis.

FORTHOFER e KOCH (1973) desenvolvem a aplicação

do método de GSK à análise de funções que são relatadas como funções logarítmicas, exponenciais e lineares das probabilidades das caselas, construindo matrizes L e Q , além das acima citadas, ampliando consideravelmente o campo dos problemas que podem ser analisados em tabelas de classificação cruzada.

No Capítulo V, vamos apresentar as duas funções compostas, propostas pelos autores acima citados, e algumas de suas aplicações.

CAPÍTULO II

MÉTODO DE GRIZZLE, STARMER E KOCH (GSK) PARA ANÁLISE DE DADOS CATEGÓRICOS.

2.1. *INTRODUÇÃO*

Antes de expor o método de análise de dados categóricos proposto por GSK, em 1969 |11|, vamos apresentar dois tipos de tabelas de classificação cruzada, o primeiro com dados provenientes de uma população, e o segundo com dados de diversas populações.

É bem conhecido entre os pesquisadores o caso de dados que se apresentam em tabelas de contingência quadradas, quando n elementos de uma amostra são classificados segundo dois atributos, sendo n_{ij} o número de elementos da casela (i,j), π_{ij} a probabilidade associada à casela (i,j), com $\pi_{i.} = \sum_j \pi_{ij}$ e $\pi_{.j} = \sum_i \pi_{ij}$ sendo as probabilidades marginais, e de modo que $\sum_i \sum_j \pi_{ij} = 1$.

É o caso, por exemplo, da tabela abaixo, apresentada por GSK |11|:

Tabela 1 - 7477 mulheres; idade 30-39; distância de visão a olho nu.

Olho direito	Olho esquerdo				Total
	Alto grau	2º grau	3º grau	Baixo grau	
Alto grau	1520	266	124	66	1976
2º grau	234	1512	432	78	2256
3º grau	117	362	1772	205	2456
Baixo grau	36	82	179	492	789
Total	1907	2222	2507	841	7477

Para a análise dos dados acima podemos formular várias hipóteses estatísticas.

Uma primeira hipótese que nos poderia ocorrer seria a da verificação de independência entre as distâncias de visão dos olhos direito e esquerdo, que seria formulada por $H0: \pi_{ij} = \pi_{i.}\pi_{.j}$. Esta hipótese não é muito adequada para esses dados, porque, como os dois atributos são tomados dos mesmos indivíduos, espera-se que haja uma natural associação entre esses atributos.

Uma segunda hipótese de associação mais adequada para esses dados é a hipótese de associação sobre a diagonal principal, que nos permite verificar se as distâncias de visão dos olhos direito e esquerdo são as mesmas. Essa

hipótese seria formulada por $H0: \pi_{ii} = \pi_{i.}\pi_{.i}$

Uma terceira hipótese, a de simetria, para verificar se as caselas situadas simetricamente sobre a diagonal principal possuem as mesmas probabilidades, seria então formulada por $H0: \pi_{ij} = \pi_{ji} \quad (i \neq j)$.

E finalmente, uma quarta hipótese de grande interesse, principalmente na ausência de simetria, e a mais adequada para esses dados, seria para verificar a existência de diferença de distribuição entre as distâncias de visão do olho direito e esquerdo. Para tal, a hipótese formulada é a da homogeneidade das marginais, dada por

$$H0: \pi_{i.} = \pi_{.i}$$

Essa hipótese equivale a:

$$H0 = \begin{pmatrix} \pi_{12}+\pi_{13}+\pi_{14}-\pi_{21}-\pi_{31}-\pi_{41} \\ \pi_{21}+\pi_{23}+\pi_{24}-\pi_{12}-\pi_{32}-\pi_{42} \\ \pi_{31}+\pi_{32}+\pi_{34}-\pi_{13}-\pi_{23}-\pi_{43} \end{pmatrix} = \begin{pmatrix} f_1(\underset{\sim}{\pi}) \\ f_2(\underset{\sim}{\pi}) \\ f_3(\underset{\sim}{\pi}) \end{pmatrix} = 0$$

onde $\underset{\sim}{\pi} = (\pi_{11}, \pi_{12}, \ldots, \pi_{44})'$

isto é, uma hipótese formada por um vetor de u funções iguais a zero, linearmente independentes.

As quatro hipóteses acima podem ser formuladas como funções de probabilidades das caselas ou das probabilidades marginais, e, de um modo geral, escritas como:

$$HO: F_m(\underset{\sim}{\pi}) = 0 \ , \quad m = 1,2,\ldots,u \qquad (I)$$

Vamos apresentar, a seguir, um tipo de tabela de classificação cruzada, que descreve dados provenientes de diversas populações; GSK baseiam todo o seu estudo nesse tipo de tabela, que passaremos a chamar "modelo geral de classificação cruzada". Convém observar que uma tabela de contingência quadrada constitui caso particular desse modelo geral, onde dados coletados de apenas uma população são classificados em categorias de respostas, que são todas as combinações das categorias dos dois atributos.

Consideremos s populações das quais extraímos amostras com $n_{i.}$ elementos, $i = 1,2,\ldots,s$, sendo que os $n_{i.}$ são classificados em r categorias, de modo que n_{ij} representa o número de elementos da i-ésima população e da j-ésima categoria de respostas, $i = 1,2,\ldots,s$ e $j= 1,2,\ldots,r$.

Temos consequentemente o seguinte:

Tabela 2 - Modelo geral de classificação cruzada.

População	Categorias de respostas					Total
	1	2	3	...	r	
1	n_{11}	n_{12}	n_{13}	...	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	...	n_{2r}	$n_{2.}$
.	.	.	.	.	.	.
s	n_{s1}	n_{s2}	n_{s3}	...	n_{sr}	$n_{s.}$

Por hipótese, as amostras aleatórias independentes são tomadas de s populações e possuem distribuição multinomial com parâmetros $n_{i.}$, $i = 1,2,\ldots,s$ e $\underset{1\times r}{\pi_i'} = \left(\pi_{i1},\pi_{i2},\ldots,\pi_{ir}\right)$ onde π_{ij} representa a probabilidade de um elemento da i-ésima população ser classificado na j-ésima categoria, e ainda de modo que $\sum_{j=1}^{r} \pi_{ij} = 1$.

A tabela abaixo fornece as probabilidades esperadas para as caselas da Tabela 2:

Tabela 3 - Probabilidades esperadas.

População	Categorias de respostas					Total
	1	2	3	...	r	
1	π_{11}	π_{12}	π_{13}	...	π_{1r}	1
2	π_{21}	π_{22}	π_{23}	...	π_{2r}	1
.	.	.	.	.	.	.
s	π_{s1}	π_{s2}	π_{s3}	...	π_{sr}	1

Seja o vetor de parâmetros desconhecidos

$$\underset{1\times rs}{\underset{\sim}{\pi}'} = \left(\underset{\sim}{\pi}_1',\underset{\sim}{\pi}_2',\ldots,\underset{\sim}{\pi}_s'\right) \qquad \underset{\sim}{\pi}_i' = \left(\pi_{i1},\pi_{i2},\ldots,\pi_{ir}\right)$$

representando um ponto π no espaço cartesiano de dimensão (rs-s). Denotamos por Ω o espaço de todos os pontos possíveis de π e chamamos Ω de espaço paramétrico. Teremos:

$$\Omega = \{(\pi_{11},\pi_{12},\ldots,\pi_{1s};\pi_{21},\pi_{22},\ldots,\pi_{2s};\ldots;\pi_{s1},\pi_{s2},\ldots,\pi_{sr})/$$

$$\pi_{ij} > 0 \quad e \sum_{j=1}^{r} \pi_{ij} = 1\}. \qquad \text{(II)}$$

A função densidade de probabilidades para cada amostra i é dada por:

$$P(n_{i1},n_{i2},\ldots,n_{ir};\pi_{i1},\pi_{i2},\ldots,\pi_{ir}) = \frac{n_{i.}!}{\prod_{j=1}^{r} n_{ij}!} \prod_{j=1}^{r} \pi_{ij}^{n_{ij}}$$

e a função densidade de probabilidades conjunta das s amostras independentes é dada por:

$$\phi = \prod_{i=1}^{s} \left\{ \frac{n_{i.}!}{\prod_{j=1}^{r} n_{ij}!} \prod_{j=1}^{r} \pi_{ij}^{n_{ij}} \right\}$$

Vamos mostrar a seguir um exemplo do "modelo geral de classificação cruzada", apresentado por BHAPKAR |5| e baseado num estudo de SCHOTZ (1966), sobre motoristas desacompanhados envolvidos em acidentes que produziram ferimentos. Os dados da tabela abaixo representam a distribuição dos casos, de acordo com o tipo de acidente indicado.

Tabela 4 - Motoristas classificados de acordo com o grupo, tipo de acidente e grau de ferimento.

Grupo motorista	Tipo de acidente	Gravidade do ferimento				Total
		leve	moderado	severo	crítico	
Desacompanhado	Capotamento	21	567	1356	644	2588
	Não-capot.	996	5454	2773	1256	10479
	Subtotal	1017	6021	4129	1900	13067
Com passageiros	Capotamento	18	553	1734	869	3174
	Não-capot.	679	4561	2516	1092	8848
	Subtotal	697	5114	4250	1961	12022
Total		1714	11135	8379	3861	25089

Nestes dados, as populações seguem uma distribuição multinomial e são formadas pelas combinações de todos os níveis de "grupo motorista" e "tipo de acidente", e portanto, na notação do "modelo geral de classificação cruzada", o número de populações, denotado por s, é igual a quatro, e o número de categorias de respostas ("gravidade do ferimento"), denotado por r, é também igual a quatro.

Para a análise da Tabela 4, as hipóteses que podem ser formuladas são analogas às que aparecem na análise de variância.

Assim, considerando os fatores "grupo motorista" e "tipo de acidente", e a resposta "gravidade do ferimento", podemos verificar a existência dos efeitos dos fatores na resposta, e ainda, a interação entre os fatores.

Vamos a seguir relatar de modo sucinto os métodos usuais de solução para as hipóteses sugeridas nas Tabelas 1 e 4.

2.2. *SOLUÇÕES USUAIS*

Verificamos que, quando consideramos dados categóricos extraídos de s populações distribuídas multinomialmente em r categorias de respostas, ou quando dados categóricos são extraídos de apenas uma população multinomial, os problemas que surgem podem ser traduzidos por hipóteses do tipo

$$H0: \underset{\sim}{\pi} \in \Omega_0 \qquad \text{(III)}$$

onde Ω_0 é um subconjunto de Ω , o espaço paramétrico, dado por (II) da pág.9.

Para testar essas hipóteses, os testes estatísticos usualmente empregados são χ^2 de PEARSON, χ_1^2 modificado, e teste da razão de verossimilhança, dados respectivamente por:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i.}\hat{p}_{ij})^2}{n_{i.}\hat{p}_{ij}} \qquad \text{(IV)}$$

$$\chi_1^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i.}\hat{p}_{ij})^2}{n_{ij}} \qquad \text{(V)}$$

$$-2 \log \lambda = \sum_{i=1}^{s} \sum_{j=1}^{r} n_{ij}\{\log n_{ij} - \log n_{i.}\hat{p}_{ij}\} \qquad \text{(VI)}$$

onde $\hat{p}_{ij}$ são os estimadores de máxima verossimilhança de π_{ij} sob H0, ou são quaisquer outros estimadores pertencentes à classe dos estimadores BAN, por exemplo, χ_1^2 modificado.

As estatísticas (IV), (V) e (VI), sob H0, possuem distribuição assintótica χ^2 com grau de liberdade dado por $(\dim.\Omega - \dim.\Omega_0)$.

Para o emprego de tais testes é necessário obter os estimadores de π_{ij} sob a restrição de que $\underset{\sim}{\pi} \in \Omega_0$.

HOEFFLING, em 1965, estudou a existência e as propriedades de estimadores de máxima verossimilhança no caso geral, isto é, quando Ω_0 é um subconjunto de Ω.

Vários estudos têm sido realizados para estimadores de máxima verossimilhança em alguns subconjuntos particulares Ω_0. É bem frequente considerar Ω_0 dado por

$$\Omega_0 = \{\underset{\sim}{\pi} \in \Omega \,/\, \pi_{ij} = \phi(\theta_1, \theta_2, \ldots, \theta_m), \theta \in R^m, \quad m < rs-s\} \qquad \text{(VII)}$$

e $\hat{p}_{ij} = \phi_{ij}(\hat{\theta})$, onde $\hat{\theta}$ é o estimador de máxima verossimilhança de θ. Nesse caso, inclui-se a hipótese de independência, onde $\hat{p}_{ij}$ são facilmente obtidos.

Para outras hipóteses, como por exemplo, associação sobre a diagonal principal, simetria e homogeneidade das marginais, os estimadores de máxima verossimilhança não podem ser obtidos de maneira analítica simples, exigindo-se o uso de métodos iterativos complicados. BOWKER, em 1948, e STUART, em 1955, verificaram a inaplicabilidade do método de máxima verossimilhança para obtenção de estimadores sob a hipótese de homogeneidade das marginais.

Por outro lado, NEYMAN, em 1949, apresentou um estudo sobre as propriedades e métodos de obtenção de estimadores χ_1^2 modificado, nos casos particulares de Ω_0 dado em (VII)

e também quando Ω_o é dado por:

$$\Omega_o = \{\pi \in \Omega \ / \ F_m(\underset{\sim}{\pi}) = 0 \ , \ m=1,2,..,u \ , \quad u \leq rs-s\} \qquad \text{(VIII)}$$

onde as F são u funções linearmente independentes de $\underset{\sim}{\pi}$ possuindo derivadas parciais contínuas de segunda ordem em relação a $\underset{\sim}{\pi}$, e sendo u o posto da matriz

$$\left[\frac{\partial F_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right]_{u \times (rs-s)}$$

Verificou NEYMAN que os estimadores χ_1^2 modificado, sob Ω_o, dado em (VIII), quando as funções F são lineares de π_{ij}, são mais fáceis de serem obtidos do que os estimadores de máxima verossimilhança, e nesse caso obtidos por um sistema de equações lineares. Para os casos onde as F não são lineares de π_{ij}, NEYMAN fornece um método de linearização, tornando possível a obtenção de estimadores através de um sistema de equações lineares.

Considerando ainda o particular Ω_o dado em (VIII) NEYMAN demonstra que as estatísticas χ^2 e χ_1^2 usando estimadores de máxima verossimilhança ou estimadores χ_1^2 modificado, são consistentes e equivalentes ao teste da razão de verossimilhança e, além disso, possuem distribuição assintóti

ca χ^2 com u graus de liberdade.

Tratamos aqui de dois tipos de estimadores, o de máxima verossimilhança, porque é o mais utilizado, e o de χ_1^2 modificado, por ter dado origem, através de seu estudo, aos trabalhos de BHAPKAR (1966), os quais constituíram a motivação fundamental da solução apresentada por GRIZZLE *et alii* (1969).

2.3. *SOLUÇÃO DE GRIZZLE, STARMER E KOCH*

Para solução dos problemas apresentados no item 2.1., envolvendo tabelas de contingência, GRIZZLE, STARMER e KOCH apresentaram, em 1969, uma teoria utilizando modelos de regressão linear e mínimos quadrados ponderados.

Essa metodologia apresenta vantagens na proposição de hipóteses e testes adequados a uma grande variedade de situações experimentais, além de apresentar facilidades na solução de problemas oriundos dessas situações.

A teoria desenvolvida por GSK foi baseada nos estudos de BHAPKAR (1966) |3|, sobre equivalência da estatística de WALD, quando adaptada a dados categóricos, e a estatística de NEYMAN, para testar hipótese geral do tipo

$$H0: F_m(\underset{\sim}{\pi}) = 0 \ , \quad m=1,2,3,\ldots,u$$

Iremos a seguir relatar de modo sucinto o estudo de BHAPKAR, sobre a adaptação da estatistica de WALD a dados categóricos e a equivalência desta última à estatística de NEYMAN.

2.3.1. ESTATÍSTICA DE WALD E SUA ADAPTAÇÃO A DADOS CATEGÓRICOS

Considerando N variáveis aleatórias X_r, $r=1,2,3,\ldots,N$, independentes, identicamente distribuídas, com densidade de probabilidades conjuntas $f(x_1,x_2,\ldots,x_N;\underset{\sim}{\Theta})$, $\underset{\sim}{\Theta} = (\Theta_1,\Theta_2,\ldots,\Theta_k)$ sendo o vetor de parâmetros desconhecidos, WALD, 1943 |26|, estudou o problema de testar hipóteses do tipo

$$H0: F_m(\underset{\sim}{\Theta}) = 0 \;, \quad m=1,2,\ldots,u \quad u \leq k \qquad \text{(IX)}$$

isto é, hipóteses constituídas por u funções de $\underset{\sim}{\Theta}$, que possuem derivadas parciais contínuas de segunda ordem em relação a $\underset{\sim}{\Theta}$, de modo que a matriz da transformação

$$H = \left[\frac{\partial F_m(\underset{\sim}{\Theta})}{\partial \underset{\sim}{\Theta}}\right]_{u \times k} \quad \text{seja definida.}$$

Para testar a hipótese (IX), WALD propôs a estatística

$$W = \Sigma \; \Sigma \; \lambda^{mn}(\hat{\underset{\sim}{\Theta}}) \; F_n(\hat{\underset{\sim}{\Theta}}) F_m(\hat{\underset{\sim}{\Theta}}) \quad m,n=1,2,\ldots,u \qquad (X)$$

onde

$\hat{\underset{\sim}{\Theta}} = (\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k)$ é o vetor dos estimadores de máxima verossimilhança de Θ,

λ^{mn} é o recíproco de λ_{mn} que é a matriz de dispersão assintótica de $F_m(\underset{\sim}{\Theta})$, definida por

$$\lambda_{mn} = H' \; I^{-1} H \; ,$$

com

$$I = E_{\underset{\sim}{\Theta}} \left\{ \frac{- \partial^2 \log f(X;\underset{\sim}{\Theta})}{\partial \underset{\sim}{\Theta}_m \; \partial \underset{\sim}{\Theta}_n} \right\} \quad m,n=1,2,\ldots,u \; ,$$

que é a Informação de FISHER.

Desta forma, a estatística de WALD pode ser escrita por:

$$W = F'(\hat{\underset{\sim}{\Theta}}) \left[H(\hat{\underset{\sim}{\Theta}}) \; I^{-1} \; H(\hat{\underset{\sim}{\Theta}}) \right]^{-1} F(\hat{\underset{\sim}{\Theta}}) \qquad (XI)$$

WALD demonstrou também que as estatísticas W e $-2 \log \lambda$ se equivalem assintoticamente em distribuição, e possuem distribuição χ^2 com u graus de liberdade sob H0. Além disso, têm "poderes equivalentes" e possuem ótimas propriedades assintóticas.

Para adaptar a teoria de WALD a dados categóricos, sob o "modelo geral de classificação cruzada", BHAPKAR considerou as variáveis aleatórias X^j_{mi} tais que:

$$X^j_{mi} = \begin{cases} 1 \text{ se a m-ésima observação da i-ésima amostra pertence à j-ésima categoria de resposta;} \\ 0 \text{ nos outros casos.} \end{cases}$$

$i = 1,2,\ldots,s$, $m = 1,2,\ldots,n_{i.}$, $j = 1,2,\ldots,r$ e

$x'_{mi} = \left[x^1_{mi}, x^2_{mi},\ldots,x^r_{mi}\right]$

A função densidade de probabilidades é dada por:

$$f(x_{11},x_{12},\ldots,x_{n_{i.},s}; \underset{\sim}{\pi}) = \prod_{i=1}^{s} \prod_{j=1}^{r} \pi_{ij}^{n_{ij}}$$

Fazendo-se $\underset{\sim}{\theta} = \underset{\sim}{\pi}$, com $u \leq rs-s$, a hipótese a ser testada será:

$$H0: F_m(\underset{\sim}{\pi}) = 0 \ , \quad m=1,2,\ldots,u \ , \quad u \leq rs-s \qquad \text{(XII)}$$

onde as F são u funções de π_{ij}, linearmente independentes, possuindo derivadas contínuas de segunda ordem em relação a π_{ij} e $H = \left(\dfrac{\partial F_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right)_{u \times rs}$ é de posto completo u.

A estatística de WALD dada em (XI) para testar a hipótese definida em (XII) assume a forma:

$$W = F'(\underset{\sim}{p}) \left[H(\underset{\sim}{p})\ V(\underset{\sim}{p})\ H'(\underset{\sim}{p}) \right]^{-1} F(\underset{\sim}{p}) \qquad \text{(XIII)}$$

onde

$$\underset{1xrs}{\underset{\sim}{p}} = \left(\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s' \right) \qquad \text{sendo}$$

$$\underset{1xr}{\underset{\sim}{p}_i'} = \left(p_{i1}, p_{i2}, \ldots, p_{ir} \right), \quad i = 1,2,\ldots,s \quad \text{e} \quad p_{ij} = \frac{n_{ij}}{n_{i.}}$$

$V(\underset{\sim}{p})$ é o estimador consistente da matriz de covariância de $\underset{\sim}{\pi}$, a qual é dada por $Var(\underset{\sim}{p}) = V(\underset{\sim}{\pi}) = I^{-1}(\underset{\sim}{\pi})$, onde I é a Informação de FISHER, no caso:

$$I(\underset{\sim}{\pi}) = \begin{bmatrix} n_{1.}(D_{\pi_1}^{-1} + \pi_{1r}^{-1} L) & 0 & \ldots & 0 \\ 0 & n_{2.}(D_{\pi_2}^{-1} + \pi_{2r}^{-1} L) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & n_{s.}(D_{\pi_s}^{-1} + \pi_{sr}^{-1} L) \end{bmatrix}$$

onde $D_{\pi_i}^{-1} = diag(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir})$ e $L = |1|_{rxr}$.

O inverso da Informação de FISHER, que é a matriz de covariância de $\underset{\sim}{\pi}$, é dado, no caso, por:

$$I^{-1}(\underset{\sim}{\pi}) = V(\underset{\sim}{\pi}) = \begin{bmatrix} n_{1.}^{-1}(D_{\pi_1}^{-1} - \pi_1\pi_1^{-}) & 0 & \cdots & 0 \\ 0 & n_{2.}^{-1}(D_{\pi_2}^{-1} - \pi_2\pi_2^{-}) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_{s.}^{-1}(D_{\pi_s}^{-1} - \pi_s\pi_s^{-}) \end{bmatrix}$$

O estimador dessa matriz é obtido trocando-se $\underset{\sim}{\pi}$ por $\underset{\sim}{p}$, e é denominado $V(\underset{\sim}{p})$.

A matriz

$$\underset{u \times u}{S} = H(\underset{\sim}{p})\ V(\underset{\sim}{p})\ H^{-}(\underset{\sim}{p})$$

é um estimador consistente da matriz de covariância de F.

A estatística dada em (XIII), como demonstrou WALD, sob a hipótese H0, definida em (XII), é uma χ^2 central, com u graus de liberdade.

2.3.2. *EQUIVALÊNCIA ALGÉBRICA ENTRE AS ESTATÍSTICAS DE NEYMAN E DE WALD.*

BHAPKAR (1966) mostrou ainda a equivalência algébrica entre as estatísticas de NEYMAN e de WALD. A estatística de NEYMAN já havia sido descrita por BHAPKAR, em

1961 |4| , para testar hipóteses lineares do tipo

$$H0: \sum_{i=1}^{s} \sum_{j=1}^{r} a_{mij}\, \pi_{ij} + h_m = F_m(\underset{\sim}{\pi}) = 0 \qquad \text{(XIV)}$$

onde a_{mij} e h_m são constantes conhecidas. Essa estatística é obtida pela minimização da estatística χ_1^2 sujeito a restrições $F_m(\underset{\sim}{\pi}) = 0$ e $\sum_{j=1}^{r} \pi_{ij} - 1 = 0$, e é dada por

$$\underset{\text{sob } H0}{\text{Min}} \; \chi_1^2 = c' G^{-1} c \qquad \text{(XV)}$$

onde $c' = \left(c_1, c_2, \ldots, c_u\right)$,com

$$c_m' = \sum_{i=1}^{s} \sum_{j=1}^{r} a_{mij}\, p_{ij} + h_m = F_m(p) \;,$$

$$G = g_{m,m'} \;, \qquad m,m' = 1,2,\ldots,u \;, \text{ com}$$

$$g_{m,m'} = \sum_{i=1}^{s} n_{i.}^{-1} \left\{ \sum_{j=1}^{r} a_{mij}\, a'_{mij}\, p_{ij} - (\Sigma a_{mij} p_{ij})(\Sigma a'_{mij} p_{ij}) \right\}$$

A estatística (XV), através do Lema 12 de NEYMAN, possui distribuição χ^2 sob H0, em amostras grandes, com u graus de liberdade.

Por outro lado, a estatística de WALD adaptada a dados categóricos, dada em (XIII), se compõe dos seguintes termos equivalentes aos termos da estatística de NEYMAN:

$$c = \left(c_1, c_2, \ldots, c_u\right) ,$$

com: $c'_m = F_m(\underset{\sim}{p})$, $m = 1, 2, \ldots, u$,

$$S = H\, V(\underset{\sim}{p})\, H' ,$$

onde $H' = \left(a_1, a_2, \ldots, a_u\right)$, com $a_m = \left(a'_{m1}, a'_{m2}, \ldots, a'_{ms}\right)$ e

$a'_{mi} = \left(a_{mi1}, a_{mi2}, \ldots, a_{mir}\right)$, e portanto:

$$G = S = H\, V(\underset{\sim}{p})\, H' = \left(a'_m\, V(\underset{\sim}{p})\, a_{m'}\right) =$$

$$= \left(\sum_{i=1}^{s} n_{i.}^{-1}\, a'_{mi}\, (D_{\underset{\sim}{p}i} - \underset{\sim}{p}_i \underset{\sim}{p}'_i)\, a_{mi}\right)$$

Portanto, as duas estatísticas (XIII) e (XV) se equivalem algebricamente. Se as funções não são lineares de π_{ij} , usando a linearização de NEYMAN, BHAPKAR obteve a mesma equivalência.

Desta forma, para testar hipóteses do tipo (XII) a estatística utilizada é a de WALD, que através da equivalência demonstrada por BHAPKAR, pertence à classe dos testes χ_1^2 modificado, com distribuição assintótica χ^2 com u graus de liberdade, sob H0.

2.3.3. A TEORIA DE GRIZZLE, STARMER E KOCH (GSK).

Baseado nos estudos de BHAPKAR, GSK formularam um método para resolver problemas que aparecem em tabelas de classificação cruzada, sob o "modelo geral de classificação cruzada", isto é, quando s amostras são retiradas de s populações independentes, e os elementos de cada amostra classificados em r categorias de respostas.

Os problemas são formulados por funções $F_m(\underset{\sim}{\pi})$, $m = 1,2,\ldots,u$, linearmente independentes, possuindo derivadas parciais contínuas de segunda ordem em relação a π_{ij}, e sendo $H = \left[\dfrac{\partial F_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right]_{\pi_{ij}=p_{ij}}$ uma matriz de posto completo u.

GSK baseiam o desenvolvimento do método nas mesmas considerações feitas para a adaptação da estatística de WALD a dados categóricos. Assim, para dados categóricos sob o "modelo geral de classificação cruzada", consideram para

$$\underset{1\times r}{\underset{\sim}{\pi}_i'} = \left(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir}\right) \quad \text{e}$$

$$\underset{1\times rs}{\pi'} = \left(\underset{\sim}{\pi}_1', \underset{\sim}{\pi}_2', \ldots, \underset{\sim}{\pi}_s'\right), \quad \text{os estimadores}$$

$$\underset{1\times r}{\underset{\sim}{p}_i'} = (p_{i1}, p_{i2}, \ldots, p_{ir}) \quad \text{com} \quad p_{ij} = n_{ij}/n_{i.} \text{ , e}$$

$$\underset{1\times rs}{\underset{\sim}{p}'} = (\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s') \text{ ; e ainda, para}$$

$$\mathrm{var}(\underset{\sim}{p}_i) = \underset{r\times r}{V(\underset{\sim}{\pi}_i)} = 1/n_{i.} \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ r \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & & r \end{array} \\ \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \cdots & -\pi_{i1}\pi_{ir} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \cdots & -\pi_{i2}\pi_{ir} \\ \cdots & \cdots & \cdots & \cdots \\ -\pi_{ir}\pi_{i1} & -\pi_{ir}\pi_{i2} & \cdots & \pi_{ir}(1-\pi_{ir}) \end{bmatrix} \end{array}$$

o estimador amostral é obtido trocando-se $\underset{\sim}{\pi}$ por $\underset{\sim}{p}$, e é denotado por $\underset{r\times r}{V(\underset{\sim}{p}_i)}$; e finalmente, para $\underset{rs\times rs}{V(\underset{\sim}{\pi})}$, matriz de covariância de π, o estimador é denotado por $V(\underset{\sim}{p})$, matriz bloco diagonal, tendo as matrizes $V(\underset{\sim}{p}_i)$ na diagonal principal, isto é:

$$V(\underset{\sim}{p}) = \begin{bmatrix} n_{1.}^{-1}(D_{\underset{\sim}{p}_1} - \underset{\sim}{p}_1\underset{\sim}{p}_1') & 0 & \cdots & 0 \\ 0 & n_{2.}^{-1}(D_{\underset{\sim}{p}_2} - \underset{\sim}{p}_2\underset{\sim}{p}_2') & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_{s.}^{-1}(D_{\underset{\sim}{p}_s} - \underset{\sim}{p}_s\underset{\sim}{p}_s') \end{bmatrix}$$

onde $D_{\underset{\sim}{p}_i} = \mathrm{diag}(p_{i1}, p_{i2}, \ldots, p_{ir})$, para $i = 1, 2, \ldots, s$.

As funções $f_m(\underset{\sim}{\pi})$ consideradas por GSK compõem o vetor

$$\left[F(\underset{\sim}{\pi})\right]' = \left[f_1(\underset{\sim}{\pi}), f_2(\underset{\sim}{\pi}), \ldots, f_u(\underset{\sim}{\pi})\right]$$

que, quando $\underset{\sim}{\pi} = \underset{\sim}{p}$, é escrito por

$$F' = \left[F(\underset{\sim}{p})\right]' = \left[f_1(\underset{\sim}{p}), f_2(\underset{\sim}{p}), \ldots, f_u(\underset{\sim}{p})\right]$$

A matriz de covariância de F é estimada por

$$S = H\, V(\underset{\sim}{p})\, H' ,$$

onde $H = \left[\dfrac{\partial F_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right]_{\pi_{ij}=p_{ij}}$ é uma matriz uxu . Quando

$f_m(\underset{\sim}{p})$ é uma função linear de elementos de $\underset{\sim}{p}$, S é a matriz exata de covariância de F ; quando $f_m(\underset{\sim}{p})$ é uma função não linear de elementos de $\underset{\sim}{p}$, S é a matriz de covariância assintótica de F, a qual é obtida pelo método "delta".

As funções $f_i(\underset{\sim}{\pi})$ devem ser independentes umas das outras e sujeitas à restrição $\sum_{j=1}^{r} \pi_{ij} = 1$, $i = 1,2,..,s$, isto é, ambos H e S são de posto u . Entretanto, para alguns tipos de dados, se qualquer dos n_{ij} for igual a zero, S será de posto menor que u . BERKSON, em 1955 |1| , sugere a substituição desses n_{ij} iguais a zero por 1/r , que tem o efeito de fazer o estimador de π_{ij} ser $1/rn_{i.}$.

Definidas as u funções de $\underset{\sim}{\pi}$, $\left(f_m(\underset{\sim}{\pi})\right)$, e os seus estimadores $\left(f_m(\underset{\sim}{p})\right)$, e sua matriz de covariância $S = H\ V(\underset{\sim}{p})\ H'$, GSK consideraram a análise das u funções através de modelo linear do tipo

$$\underset{u\times 1}{F}(\underset{\sim}{\pi}) = \underset{u\times v}{X}\ \underset{v\times 1}{\underset{\sim}{\beta}} + \varepsilon \qquad \text{(XVI)}$$

onde

X é uma matriz de planejamento de posto $v \leq u$,

$\underset{\sim}{\beta}$ é um vetor de parâmetros desconhecidos, e

$\underset{\sim}{\varepsilon}$ é o vetor dos erros,

de tal modo que

$$E\ \{F(\underset{\sim}{\pi})\} = X\ \underset{\sim}{\beta} \quad \text{e} \quad \text{var}\ \{F(\underset{\sim}{\pi})\} = \text{var}(\underset{\sim}{\varepsilon}) = S = H.\text{Var}(\underset{\sim}{p}).H'$$

GSK utilizaram o método dos mínimos quadrados generalizados a fim de estimar $\underset{\sim}{\beta}$; o valor de $\underset{\sim}{\beta}$ que minimiza $(F-X\underset{\sim}{\beta})'S^{-1}(F-X\underset{\sim}{\beta})$ é o estimador de $\underset{\sim}{\beta}$, e é dados por:

$$\underset{\sim}{b} = (X'S^{-1}X)^{-1}XS^{-1}F \qquad \text{(XVII)}$$

que muitos trabalhos provam pertencer à classe de estimadores BAN (por exemplo, THEIL, 1973). A variância de $\underset{\sim}{b}$ é dada por

$$\text{var}(\underset{\sim}{b}) = (XS^{-1}X)^{-1}\ . \qquad \text{(XVIII)}$$

O teste estatístico do ajuste do modelo é dado por

$$SS\left(F(\underset{\sim}{\pi}) = X\underset{\sim}{\beta}\right) = F'S^{-1}F - \underset{\sim}{b}'(X'S^{-1}X)\underset{\sim}{b} \qquad \text{(XIX)}$$

que possui distribuição χ^2 assintótica com u-v graus de liberdade sob a hipótese do ajuste do modelo.

Em muitos casos de dados provenientes de apenas uma população, as hipóteses estatísticas apropriadas são formuladas por $\underset{u\times 1}{F(\underset{\sim}{\pi})} = 0$; assim, não é necessária a construção da matriz de planejamento X, e o teste do ajuste do modelo se resume a $F'S^{-1}F$, que é uma χ^2 central com u graus de liberdade.

Se o modelo descreve adequadamente os dados, podemos necessitar de testes para testar os componentes do vetor $\underset{\sim}{\beta}$. Podemos formular hipóteses da forma

$$H0: \underset{d\times v}{C} \; \underset{v\times 1}{\underset{\sim}{\beta}} = 0 \qquad \text{(XX)}$$

onde C é uma matriz de constantes adequadas de posto $d \leq v$.

Para testar tais hipóteses, utilizamos:

$$SS\left(C\underset{\sim}{\beta} = 0\right) = \underset{\sim}{b}'C'\left(C(X'S^{-1}X)^{-1}C'\right)^{-1} C\underset{\sim}{b} \qquad \text{(XXI)}$$

que possui distribuição assintótica χ^2 com d graus de liberdade, sob a hipótese definida em (XX).

A forma de S depende de H e, através de H, da função $F(\underset{\sim}{\pi})$. Além disso, para cada família de funções $F(\underset{\sim}{\pi})$, S será diferente. GSK consideraram dois tipos de funções, linear e logarítmica, definindo assim duas famílias de funções:

$$\text{A)}\ \underset{u\times 1}{F(\underset{\sim}{\pi})} = \underset{u\times rs}{A}\ \underset{rs\times 1}{\underset{\sim}{\pi}} \qquad \text{(XXII)}$$

onde A é uma matriz de constantes $a_{\gamma ij}$, de posto $u \leq rs-s$ e é dada por:

$$A = \begin{pmatrix} a_{111} & a_{112} & \cdots a_{11r}; & a_{121} & a_{122} & \cdots a_{12r}; \cdots; & a_{1s1} & a_{1s2} & \cdots a_{1sr} \\ a_{211} & a_{212} & \cdots a_{21r}; & a_{221} & a_{222} & \cdots a_{22r}; \cdots; & a_{2s1} & a_{2s2} & \cdots a_{2sr} \\ \cdot & \cdot & \cdots \cdot & \cdot & \cdot & \cdots \cdot \cdots & \cdot & \cdot & \cdots \cdot \\ a_{u11} & a_{u12} & \cdots a_{u1r}; & a_{u21} & a_{u22} & \cdots a_{u2r}; \cdots; & a_{us1} & a_{us2} & \cdots a_{usr} \end{pmatrix}$$

A matriz de derivadas parciais da transformação $F(\underset{\sim}{\pi}) = A(\underset{\sim}{\pi})$ é

$$H = \left[\frac{\partial F_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right]_{\pi_{ij}=p_{ij}} = A$$

e, portanto, a matriz de covariância estimada de F é dada por

$$S = A\ V(\underset{\sim}{p})\ A' \qquad \text{(XXIII)}$$

$$\text{B)} \quad \underset{t\times 1}{F}(\underset{\sim}{\pi}) = \underset{t\times u}{K} \log_e \underset{u\times rs}{A} \underset{rs\times 1}{\pi} \qquad \text{(XXIV)}$$

onde K é uma matriz $t \times u$ de constantes arbitrárias, cujo posto é $t \leq u \leq rs$; A é a matriz definida no item anterior, e $\log_e$ transforma um vetor no vetor correspondente de logaritmos naturais.

A matriz de derivadas parciais da transformação $F(\underset{\sim}{\pi}) = K \log_e A\underset{\sim}{\pi}$ é dada por

$$H = \left[\frac{\partial F_m(\underset{\sim}{\pi})}{\partial \pi_{ij}}\right]_{\pi_{ij}=p_{ij}} = K D_a^{-1} A$$

onde D_a é uma matriz diagonal com elementos do vetor $a=Ap$ na diagonal principal.

A matriz de covariância estimada de F é dada por

$$S = K D_a^{-1} A V(p) A' D_a^{-1} K' . \qquad \text{(XXV)}$$

O estudo de hipóteses como aquelas que propusemos para análise dos dados da Tabela 1, bem como para análise de outros tipos de tabelas, como a Tabela 4, através do método de GSK, será feito nos capítulos seguintes.

CAPÍTULO III

APLICAÇÃO A PROBLEMAS DE UMA POPULAÇÃO

3.1. *INTRODUÇÃO*

Vimos, no Capítulo II, que a análise estatística de dados categóricos provindos de uma população com distribuição multinomial, poderia ser feita, dependendo do objetivo, por hipóteses como:

a) homogeneidade das marginais, formulada por:

$$H0: \pi_{i.} = \pi_{.i}$$

b) simetria, formulada por:

$$H0: \pi_{ij} = \pi_{ji} \qquad i \neq j$$

c) independência, formulada por:

$$H0: \pi_{ij} = \pi_{i.}\, \pi_{.j}$$

d) associação sobre a diagonal principal, formulada por:

$$H0: \pi_{ii} = \pi_{i.}\, \pi_{.i}$$

as quais poderiam ser formuladas, de um modo geral, por:

$$H0: F_m(\underset{\sim}{\pi}) = 0 \ , \quad m=1,2,\ldots,u$$

Neste capítulo, pretendemos estudar estas hipóteses através do método de GSK, isto é, transformando essas funções de π_{ij} em funções do tipo $F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$ ou ainda, $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$; além disso, pretendemos identificar o modelo de classificação cruzada para o qual alguma dessas quatro hipóteses é mais adequada.

A identificação de modelos de dados categóricos pode ser feita através dos termos "fator" e "resposta", cuja distinção foi salientada por BHAPKAR e KOCH, em 1968, | 5 | e que propicia a formulação de hipóteses adequadas.

Devemos entender o termo "fator" como uma classificação caracterizada por uma descrição de alguns aspectos da população ou subpopulação às quais os sujeitos pertencem (por exemplo, blocos, tratamentos), e nesse caso as marginais totais referentes aos níveis do fator são fixas; pelo termo "resposta", entende-se uma classificação caracterizada pela descrição do que acontece com os sujeitos durante o experimento (por exemplo, morre ou vive, é baixo, médio, alto), e nesse caso as marginais totais não são fixas antes da realização do experimento.

Os dados provindos de uma população possuem somente o tamanho n da amostra fixo, e as marginais totais são

aleatórias, qualquer que seja o número de entradas considerado na tabela.

As situações em que os n sujeitos retirados de uma população são classificados em respostas, constituem os modelos "multirespostas, nenhum fator", por exemplo, situações em que n sujeitos são classificados de acordo com o seu status e o status do pai, constituindo um modelo "duas respostas, nenhum fator".

Existem ainda situações experimentais nas quais os n elementos de uma amostra retirados de uma população são expostos aos m níveis de um fator e classificados em d categorias de resposta associada a cada nível. Seria o caso, por exemplo, de n mulheres submetidas a dois tipos de diagnóstico do Ca do colo uterino: exame histopatológico e exame colposcópico. Tipo de exame seria o fator, com dois níveis, o histopatológico e o colposcópico, e onde cada exame, por sua vez, seria classificado de acordo com , por exemplo, três categorias de respostas - leve, moderado e severo. Esse modelo é denominado "modelo misto de ordem dois" e é traduzido por uma tabela de contingência 3x3 . De um modo geral, para m níveis do fator e d categorias de respostas, temos um"modelo misto de ordem m" , e os dados sumarizados em tabelas de contingências dxdx...xd.

Ambos os modelos, "multiresposta, nenhum fator" e "modelo misto de ordem m", representam experimentos envolvendo apenas uma distribuição multinomial. A identificação dos modelos é vantajosa, pois favorece a formulação da hipótese adequada.

Como vamos restringir o estudo a tabelas de classificação cruzada bidimensionais e tridimensionais, consideramos os modelos:

"duas respostas, nenhum fator"

"três respostas, nenhum fator"

"modelos mistos de ordem dois"

"modelos mistos de ordem três".

Considerando esses modelos de tabelas, vamos aplicar a metodologia de GSK para testar hipóteses mais adequadas para cada modelo.

3.2. *HIPÓTESES DO TIPO* $H0: F(\tilde{\pi}) = A\pi = 0$

Vamos tratar aqui das hipóteses de homogeneidade das marginais e de simetria, desenvolvidas como funções lineares de π_{ij}, e formuladas por $H0: F(\tilde{\pi}) = A\tilde{\pi} = 0$, sendo A uma matriz de constantes uxrs e π o vetor de probabilidades das caselas.

Para testar essas hipóteses, considerando $F = A\underset{\sim}{p}$ o vetor estimado de $F(\underset{\sim}{\pi})$ e S a sua matriz de covariância, dada em (XXIII), a soma dos quadrados é fornecida por $F'S^{-1}F$, que possui distribuição assintótica χ^2 com u graus de liberdade.

3.2.1. *HIPÓTESE DE HOMOGENEIDADE DAS MARGINAIS EM MODELOS MISTOS DE ORDEM DOIS.*

Vamos considerar a seguir situações em que n sujeitos retirados de uma amostra são expostos segundo dois níveis de um fator e classificados em d categorias de respostas, constituindo tabelas de contingências d x d.

Além do exemplo citado na pág.32, referente ao diagnóstico do Ca uterino, podemos assinalar outras situações que se enquadrem nesse modelo. É o caso, por exemplo, de trabalhos experimentais que envolvem classificações de cada sujeito em respostas às situações "antes e depois", "controle e tratamento", onde cada sujeito é o seu próprio controle. Podemos citar como exemplo dessa situação um experimento em que são tomadas as opiniões de um grupo de adolescentes sobre a aplicação de uma maior ou menor ação punitiva diante de uma certa forma de delinquência, antes e

depois da exibição de um certo filme. Os dados são colocados em uma tabela de contingência 2 x 2, que é considerada um "modelo misto de ordem dois", sendo "fator" a ação punitiva com dois níveis, antes e depois da exibição do filme, e "resposta" o nível da ação punitiva, com duas categorias, maior e menor.

No exemplo citado na pág.32, a hipótese conveniente para a análise é a de que não existe diferença entre os dois métodos, o colposcópico e o histopatológico, para o diagnóstico do Ca uterino; no exemplo referente às opiniões sobre a ação punitiva, podemos estar interessados em verificar se as opiniões sobre o nível da ação punitiva diferem ou não antes e depois da exibição do filme.

As hipóteses propostas para a análise dos exemplos citados são hipóteses de homogeneidade das marginais ou hipóteses de homogeneidade dos níveis do fator, e são formuladas, para tabelas de contingência d x d , por:

$$H0: \pi_{i.} = \pi_{.j} \quad i = j = 1,2,\ldots,d$$

A hipótese de homogeneidade das marginais é mais conveniente para a análise de modelos mistos; nesses modelos, a hipótese de independência é irrelevante, desde que se espera que as respostas de um mesmo sujeito para diferentes níveis do fator sejam naturalmente associadas.

Essa hipótese havia sido anteriormente estudada por BOWKER (1948) e STUART (1955), os quais verificaram a inaplicabilidade dos métodos usuais; IRELAND, KU e KULLBACK em 1969, utilizaram o princípio da informação para obtenção de estimadores.

Vamos descrever a utilização do método de GSK para o estudo da hipótese de homogeneidade das marginais.

Para isso, consideramos um "modelo misto de ordem dois", com d categorias de respostas. O modelo é descrito por uma tabela de contingência $d \times d$, portanto, as frequências n_{ij}, $i, j = 1,2,\ldots,d$ são variáveis aleatórias com distribuição multinomial, com parâmetros n e π_{ij}, $i, j = 1,2,\ldots,d$, sendo $\sum_i \sum_j \pi_{ij} = 1$.

Segundo as notações do "modelo geral de classificação cruzada", os valores para o número de populações e para o número de categorias de respostas, são $s = 1$ e $r = d^2$.

A hipótese de homogeneidade das marginais, formulada por

$$H0: \pi_{i.} = \pi_{.j} \quad , \quad i = j = 1,2,\ldots,d$$

onde $\pi_{i.} = \sum_{j=1}^{d} \pi_{ij}$ e $\pi_{.j} = \sum_{i=1}^{d} \pi_{ij}$, é composta de $d - 1$ restrições linearmente independentes, e pode ser es

crita por:

$$H0: F_m(\underset{\sim}{\pi}) = \pi_{m.} - \pi_{.m} = 0 \quad , \quad m = 1,2,\ldots,d-1$$

ou, de forma equivalente, por:

$$H0: F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$$

onde A é uma matriz dada por

$$\underset{(d-1)\times d^2}{A} = \begin{pmatrix} 0\ 1\ 1\ ..\ 1\ 1; -1\ 0\ 0\ ..\ 0\ 0; -1\ 0\ 0\ 0\ ..\ 0\ 0; \ldots; -1\ 0\ 0\ ..\ 0\ 0 \\ 0\ -1\ 0\ ..\ 0\ 0;\ 1\ 0\ 1\ ..\ 1\ 1;\ 0\ -1\ 0\ 0\ ..\ 0\ 0; \ldots;\ 0\ -1\ 0\ ..\ 0\ 0 \\ 0\ 0\ -1\ ..\ 0\ 0;\ 0\ 0\ -1\ ..\ 0\ 0;\ 1\ 1\ 0\ 1\ ..\ 1\ 1; \ldots;\ 0\ 0\ -1\ ..\ 0\ 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 0\ 0\ 0\ ..\ -1\ 0;\ 0\ 0\ 0\ ..\ -1\ 0;\ 0\ 0\ 0\ 0\ ..\ -1\ 0; \ldots;\ 0\ 0\ 0\ ..\ -1\ 0 \end{pmatrix}$$

e $\pi' = \left(\pi_{11}, \pi_{12}, \ldots, \pi_{1d}; \pi_{21}, \pi_{22}, \ldots, \pi_{2d}; \ldots; \pi_{d1}, \pi_{d2}, \ldots, \pi_{dd}\right)$

Para testar a hipótese de homogeneidade das marginais, numa tabela d x d, a soma dos quadrados é fornecida por $F'S^{-1}F$, que é uma χ^2 assintótica com (d-1) graus de liberdade sob H0.

Vamos ilustrar esse método de solução para hipótese de homogeneidade das marginais, com um exemplo de "modelo misto de ordem dois", fornecido e resolvido por GSK |11|, e cujos dados já mostramos na Tabela 1 da pág.5.

Para esses dados, a hipótese de homogeneidade das marginais é formulada por:

$$H0: \begin{cases} \pi_{1.} = \pi_{.1} \\ \pi_{2.} = \pi_{.2} \\ \pi_{3.} = \pi_{.3} \\ \pi_{4.} = \pi_{.4} \end{cases} \iff H0: \begin{cases} \pi_{1.} - \pi_{.1} = 0 \\ \pi_{2.} - \pi_{.2} = 0 \\ \pi_{3.} - \pi_{.3} = 0 \\ \pi_{4.} - \pi_{.4} = 0 \end{cases}$$

(somente 3 são l.i., pois $\sum_i \sum_j \pi_{ij} = 1$)

ou, em outros termos,

$$H0: F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$$

onde A é dada por:

$$A_{3x16} = \begin{array}{c} \begin{array}{cccccccccccccccc} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} & \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} & \pi_{31} & \pi_{32} & \pi_{33} & \pi_{34} & \pi_{41} & \pi_{42} & \pi_{43} & \pi_{44} \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \end{pmatrix} \end{array}$$

e $\underset{\sim}{\pi}' = \left(\underset{\sim}{\pi}_1', \underset{\sim}{\pi}_2', \underset{\sim}{\pi}_3', \underset{\sim}{\pi}_4'\right)$ onde $\underset{\sim}{\pi}_i' = \left(\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4}\right)$, $i = 1,2,3,4$.

Os valores de F e S obtidos foram:

$$F' = \left(0{,}00923 \quad 0{,}00455 \quad 0{,}00682\right)$$

$$S = A\,V(\underset{\sim}{p})\,A' = 10^{-4} \begin{pmatrix} 0{,}1507 & -0{,}0894 & -0{,}0430 \\ -0{,}0894 & 0{,}2601 & -0{,}1420 \\ -0{,}0430 & -0{,}1420 & 0{,}2538 \end{pmatrix}$$

Obteve-se o valor $F'S^{-1}F = 11{,}98$, e como $\chi^2_{3;0,05} = 7{,}81$, rejeita-se H0, isto é, não há homoge-

neidade entre as marginais, ou seja, existem diferenças significativas entre as distâncias de visão do olho direito e esquerdo, ao nível de 5%.

3.2.2. *HIPÓTESES DE HOMOGENEIDADE DAS MARGINAIS EM MODELOS MISTOS DE ORDEM TRÊS.*

Vamos considerar aqui situações em que cada um dos elementos de uma amostra é exposto a três níveis de um fator e classificado segundo d categorias de respostas, cada resposta sendo associada a cada nível do fator, constituindo "modelo misto de ordem três".

Seria o caso, por exemplo, de um experimento realizado por LESSLER (1962), para estudo de respostas quanto à natureza do simbolismo sexual em objetos. Para isso, um grupo de pessoas foi solicitado a classificar como masculinos ou femininos certos objetos, os quais haviam sido previamente caracterizados como sendo culturalmente masculinos e anatomicamente femininos, em três diferentes tempos de exposição: 1/5 seg., 1/100 seg., e 1/1000 seg. Assim, esse modelo é um "modelo misto de ordem três", sendo "fator" o tempo de exposição, com três níveis citados acima, e "resposta" o sexo, com duas categorias, masculino e feminino.

O objetivo desse estudo é a comparação das respostas nos diferentes tempos de observação, portanto, a hipótese é a da homogeneidade dos níveis do fator.

Para formulação desta hipótese e seu estudo através do método de GSK, vamos descrever um "modelo misto de ordem três", traduzido por uma tabela de três entradas dxdxd, sendo cada entrada um nível do fator com d categorias de respostas. Assim, sejam:

n_{ijk} , $i,j,k = 1,2,\ldots,d$, o número de pessoas classificadas segundo a i-ésima categoria de resposta associada ao primeiro nível do fator, j-ésima categoria de resposta associada ao segundo nível do fator, e k-ésima categoria de resposta associada ao terceiro nível do fator; e

π_{ijk} a probabilidade associada à casela (i,j,k), de modo tal que $\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\pi_{ijk} = 1$ e $\pi_{i..} = \sum_{j=1}^{d}\sum_{k=1}^{d}\pi_{ijk}$, $\pi_{.j.} = \sum_{i=1}^{d}\sum_{k=1}^{d}\pi_{ijk}$ e $\pi_{..k} = \sum_{i=1}^{d}\sum_{j=1}^{d}\pi_{ijk}$ as probabilidades marginais.

Nesse modelo, a hipótese de homogeneidade sobre os níveis do fator é formulada por:

$$H0: \pi_{i..} = \pi_{.j.} = \pi_{..k} \qquad i=j=k= 1,2,\ldots,d$$

Desde que temos u funções linearmente independentes, a hipótese acima equivale a:

$$H0: \begin{cases} \pi_{i..} - \pi_{.j.} = 0 \\ \pi_{i..} - \pi_{..k} = 0 \end{cases} \qquad i=j=k= 1,2,\ldots,d-1$$

que se enquadra na forma $F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$; portanto, a soma dos quadrados para testar essa hipótese é dada por $F'S^{-1}F$, que é uma χ^2 assintótica com (m-1)(d-1) graus de liberdade, onde m é o número de entradas da tabela e d é o número de categorias de respostas.

Vamos ilustrar a solução da hipótese de homogeneidade das marginais utilizando o modelo misto estudado por GSK (1969) e, anteriormente, por BHAPKAR (1965) e COCHRAN (1950). O exemplo consiste de dados indicando as respostas "favorável" ou "não favorável", de 46 sujeitos submetidos às drogas A, B e C. Esse é um exemplo de "modelo misto de ordem três", com duas categorias de respostas, sendo o fator a droga, com três níveis - A, B e C -, portanto, m=3, e resposta com duas categorias - "favorável" e "não favorável" -, portanto, d=2. Os dados obtidos estão na tabela abaixo:

Tabela 5 - Tabulação de respostas às drogas A, B e C.

Resposta A		favorável			não favorável			Total
Resposta B		fav.	não fav.	tot.	fav.	não fav.	tot.	
Resp.C	favor.	6	2	8	2	6	8	16
	não fav.	16	4	20	4	6	10	30
Total		22	6	28	6	12	18	46

A hipótese de interesse é a da igualdade da probabilidade da resposta favorável para cada uma das três drogas, isto é, igualdade das correspondentes distribuições marginais. Logo,

$$H0: \pi_{i..} = \pi_{.j.} = \pi_{..k} \quad i=j=k= 1,2 \quad \text{ou}$$

$$H0: \pi_{1..} = \pi_{.1.} = \pi_{..1} \quad \text{que equivale a}$$

$$H0: \begin{cases} \pi_{121} + \pi_{122} - \pi_{211} - \pi_{212} = 0 \\ \pi_{112} + \pi_{122} - \pi_{211} - \pi_{221} = 0 \end{cases}$$

Desta forma, H0 é do tipo $F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$, onde A é uma matriz 2 x 8 dada por:

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 \end{pmatrix}$$

A matriz de covariância de F é dada por:

$$S = 1/46 \begin{pmatrix} 6/23 & 3/23 \\ 3/23 & 12/23 \end{pmatrix}$$ e sua inversa será:

$$S^{-1} = \frac{46 \times 23}{21} \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}$$

e a soma dos quadrados $F'S^{-1}F$ resulta 6,58 com 2 graus de liberdade, portanto rejeita-se H0, isto é, as drogas não são igualmente eficientes.

3.2.3. *HIPÓTESE DE SIMETRIA EM TABELAS DE CONTINGÊNCIA dxd*

Em situações onde dados categóricos são classificados em uma tabela de contingência dxd , possuindo os dois atributos as mesmas categorias de respostas, como descritos para "modelos mistos de ordem dois" (por exemplo, visão do olho direito e esquerdo, com categorias alto grau, segundo grau, terceiro grau e baixo grau), ou em modelos "duas respostas, nenhum fator", num caso particular onde o número de categorias das duas respostas é igual (por exemplo, status do pai e status do filho, com as categorias alto, médio e baixo), além das hipóteses de independência e homogeneidade das marginais, podemos testar outras hipóteses como a de simetria.

A hipótese de simetria é utilizada quando queremos verificar se as caselas situadas simetricamente sobre a diagonal principal têm a mesma probabilidade de ocorrência, e é formulada por:

$$H0: \pi_{ij} = \pi_{ji} \qquad i,j = 1,2,\ldots,d \ , \quad i \neq j$$

STUART (1955) já havia estudado o problema de testar simetria em tabelas de contingência dxd , usando estimadores de máxima verossimilhança, e KULLBACK (1968) propôs uma estatística baseada no princípio da informação.

Aplicando o método de GSK à hipótese de simetria, construímos $d(d-1)/2$ funções lineares de π_{ij} , de modo que essa hipótese equivale a

$$H0: F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$$

onde A é uma matriz de constantes $\left(\frac{d(d-1)}{2} \times d^2\right)$ e $\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

Para testar a hipótese de simetria, a soma dos quadrados é fornecida por $F'S^{-1}F$, que possui distribuição χ^2 assintótica com $d(d-1)/2$ graus de liberdade sob H0.

Vamos ilustrar o uso do método de GSK para a hipótese de simetria utilizando os dados da tabela 1. No caso, a hipótese de simetria é formulada por:

$$H0: \begin{cases} \pi_{12} = \pi_{21} \\ \pi_{13} = \pi_{31} \\ \pi_{14} = \pi_{41} \\ \pi_{23} = \pi_{32} \\ \pi_{24} = \pi_{42} \\ \pi_{34} = \pi_{43} \end{cases} \quad \text{que equivale a} \quad H0: \begin{cases} \pi_{12} - \pi_{21} = 0 \\ \pi_{13} - \pi_{31} = 0 \\ \pi_{14} - \pi_{41} = 0 \\ \pi_{23} - \pi_{32} = 0 \\ \pi_{24} - \pi_{42} = 0 \\ \pi_{34} - \pi_{43} = 0 \end{cases}$$

que é do tipo $H0: F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi} = 0$, onde A é dada por:

$$\underset{6\times 16}{A} = \begin{array}{c} \begin{array}{cccccccccccccccc} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} & \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} & \pi_{31} & \pi_{32} & \pi_{33} & \pi_{34} & \pi_{41} & \pi_{42} & \pi_{43} & \pi_{44} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \end{pmatrix} \end{array}$$

Os estimadores de $F(\underset{\sim}{\pi})$ e da sua matriz de covariância resultam:

$$F(\underset{\sim}{p}) = \begin{pmatrix} 0{,}0042798 & 0{,}0009362 & 0{,}0040123 & 0{,}009362 & -0{,}00053497 \\ 0{,}003477 \end{pmatrix}$$

$$S = 10^{-3} \begin{pmatrix} 894{,}12 & -0{,}0536 & -0{,}2297 & -0{,}5359 & 0{,}0306 & -0{,}1990 \\ -0{,}0536 & 431{,}07 & -0{,}0502 & -0{,}1172 & 0{,}0067 & -0{,}0435 \\ -0{,}2297 & -0{,}0502 & 182{,}24 & -0{,}5024 & 0{,}0287 & -0{,}1866 \\ -0{,}5359 & -0{,}1172 & -0{,}5024 & 1419{,}1 & 0{,}0669 & -0{,}4354 \\ 0{,}0306 & 0{,}0067 & 0{,}0287 & 0{,}0669 & 286{,}19 & 0{,}0248 \\ -0{,}1990 & -0{,}0435 & -0{,}1866 & -0{,}4354 & 0{,}02488 & 686{,}71 \end{pmatrix}$$

O χ^2 para testar essa hipótese resulta 19,155, com 6 graus de liberdade, o que significa rejeição de H0, isto é, não existe simetria.

5.3. HIPÓTESES DO TIPO $H0: F(\pi) = K \ln A\pi = 0$.

Vamos tratar agora do estudo de hipóteses como as de independência e associação sobre a diagonal principal, as quais, para aplicação do método de GSK necessitam ser escritas sob a forma $H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$.

Sabemos que as hipóteses de independência são facilmente solucionadas por métodos usuais, devido à facilidade com que são obtidos os estimadores de máxima verossimilhança, e que, em algumas tabelas com as de três entradas (modelo "três respostas, nenhum fator"), na utilização do método de GSK, existe grande dificuldade para a construção de matrizes K adequadas. Assim, para facilitar a obtenção dessas matrizes, para as hipóteses de independência e de não interação, apresentamos uma técnica sugerida por GRIZZLE e WILLIAMS (1972) |12| , e que consiste na construção de uma tabela de contrastes.

Para testar hipóteses do tipo $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, se $F = K \ln A\underset{\sim}{p}$ é o vetor estimado de $F(\underset{\sim}{\pi})$, e S a matriz de covariância de F , dada em (XXV), a soma dos quadrados é fornecida por $F'S^{-1}F$, que possui distribuição assintótica χ^2 com t graus de liberdade sob H0.

3.3.1. HIPÓTESE DE INDEPENDÊNCIA EM TABELAS DE CONTINGÊNCIA c x d.

Uma situação que aparece comumente na pesquisa é a que se refere a dados coletados de uma população, classificados segundo categorias de respostas, A e B, cada uma com c e d níveis respectivamente, e assumindo as frequências uma simples distribuição multinomial.

Na notação do "modelo geral de classificação cruzada", o número de populações é dado por s=1 e o número de categorias de respostas, que são todas as combinações dos níveis de A e B, é dado por r=cxd.

Utilizando os termos "fator" e "resposta", temos um modelo "duas respostas, nenhum fator". Situações "duas respostas, nenhum fator", nas quais nenhuma das marginais é fixa (por isso o modelo é chamado "nenhum fator") e onde somente o tamanho da amostra é fixo, são, por exemplo, fornecidas em tabelas cxd (por exemplo, sexo x reação de Mitsuda) e em tabelas particulares dxd (por exemplo, status do pai x status do filho).

Consideremos, para desenvolver a hipótese de independência, a seguinte tabela de contingência cxd:

Tabela 6 - Tabela de contingência c x d .

Resposta A	Resposta B					Total
	B_1	B_2	B_3	...	B_d	
A_1	π_{11}	π_{12}	π_{13}	...	π_{1d}	$\pi_{1.}$
A_2	π_{21}	π_{22}	π_{23}	...	π_{2d}	$\pi_{2.}$
A_3	π_{31}	π_{32}	π_{33}	...	π_{3d}	$\pi_{3.}$
.	.	.	.	...	.	.
A_c	π_{c1}	π_{c2}	π_{c3}	...	π_{cd}	$\pi_{c.}$
Total	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	...	$\pi_{.d}$	1

Seja π_{ij} a probabilidade associada à i-ésima categoria de A, $i=1,2,\ldots,c$, e j-ésima categoria de B, $j=1,2,\ldots,d$, e $\pi_{i.} = \sum_j \pi_{ij}$, $\pi_{.j} = \sum_i \pi_{ij}$ as probabilidades marginais aleatórias, de tal modo que $\sum_i \sum_j \pi_{ij} = 1$.

Em modelos desse tipo, o experimentador pode estar interessado na análise da associação entre as respostas (por exemplo, se a intensidade da reação de Mitsuda independe de sexo, se status do filho independe do status do pai), cuja hipótese é a de independência entre as respostas, formulada por:

$$H0: \pi_{ij} = \pi_{i.}\pi_{.j} \qquad i=1,2,\ldots,c \quad j=1,2,\ldots,d$$

Numa tabela particular 2 x 4 teremos:

$$H0: \pi_{ij} = \pi_{i.}\,\pi_{.j} \qquad i=1,2 \qquad j=1,2,3,4 \qquad \text{ou}$$

$$H0: \begin{cases} \pi_{11} = \pi_{1.}\pi_{.1} \\ \pi_{12} = \pi_{1.}\pi_{.2} \\ \pi_{13} = \pi_{1.}\pi_{.3} \\ \pi_{14} = \pi_{1.}\pi_{.4} \\ \pi_{21} = \pi_{2.}\pi_{.1} \\ \pi_{22} = \pi_{2.}\pi_{.2} \\ \pi_{23} = \pi_{2.}\pi_{.3} \\ \pi_{24} = \pi_{2.}\pi_{.4} \end{cases} \Longleftrightarrow H0: \begin{cases} \dfrac{\pi_{11}\,\pi_{24}}{\pi_{21}\,\pi_{14}} = 1 \\ \dfrac{\pi_{12}\,\pi_{24}}{\pi_{22}\,\pi_{21}} = 1 \\ \dfrac{\pi_{13}\,\pi_{24}}{\pi_{23}\,\pi_{14}} = 1 \end{cases}$$

que equivale a

$$H0: \begin{cases} \ln \pi_{11} - \ln \pi_{14} - \ln \pi_{21} + \ln \pi_{24} = 0 \\ \ln \pi_{12} - \ln \pi_{21} - \ln \pi_{22} + \ln \pi_{24} = 0 \\ \ln \pi_{13} - \ln \pi_{14} - \ln \pi_{23} + \ln \pi_{24} = 0 \end{cases}$$

que é uma hipótese do tipo $H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, onde:

$$\underset{3x8}{K} = \begin{pmatrix} 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{pmatrix},$$

A uma matriz identidade 8x8 , e

$$\pi' = \left\{\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{24}\right\} .$$

De um modo geral, numa tabela de contingência c x d, a hipótese de independência entre duas respostas é descrita por (c-1)(d-1) funções logarítmicas do tipo

$$F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0 ,$$

onde K é uma matriz (c-1)(d-1) x cd , A é uma matriz identidade cd x cd , e π' é o vetor das probabilidades das caselas (1 x cd).

A fim de encontrarmos essas funções logarítmicas com maior facilidade, podemos nos utilizar do conceito de não interação, que, em tabelas de contingência c x d, equivale ao de independência, estudado por BHAPKAR e KOCH (1968); segundo esses autores, o conceito de não interação entre duas respostas está relacionado com uma medida de associação Δ_{ij} dada por

$$\Delta_{ij} = \frac{\pi_{ij}\ \pi_{cd}}{\pi_{cj}\ \pi_{id}} \qquad i=1,2,..,c-1 \quad j=1,2,..,d-1 ,$$

e que satisfaz $\Delta_{ij} = 1$ quando não existe interação.

Desse modo, a hipótese de independência

$$H0: \pi_{ij} = \pi_{i.}\ \pi_{.j} \qquad i=1,2,..,c \quad j=1,2,..,d$$

equivale a

$$H0: \Delta_{ij} = 1 \qquad i = 1,2,...,c-1 \quad j = 1,2,..,d-1 .$$

Se tomarmos os logaritmos de Δ_{ij}, obtemos:

$$H0: \ln \Delta_{ij} = \ln \pi_{ij} - \ln \pi_{cj} - \ln \pi_{id} + \ln \pi_{cd} = 0$$

que é da forma $H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$.

O teste adequado para essa hipótese é fornecido por $F'S^{-1}F$, que possui distribuição assintótica χ^2 com $(c-1)(d-1)$ graus de liberdade sob H0.

3.3.2. *HIPÓTESES DE INDEPENDÊNCIA E DE NÃO INTERAÇÃO EM TABELAS c x d x f.*

Consideremos uma amostra de tamanho n retirada de uma população e classificada segundo três respostas, com c, d e f categorias, respectivamente, assumindo os resultados uma simples distribuição multinomial.

Desse modo, na notação de GSK, s=1 e o número de categorias de respostas é r=cxdxf; na linguagem de "fator" e "resposta", é um modelo "três respostas, nenhum fator". São situações como, por exemplo, sexo, idade e aproveitamento escolar; técnicos, tipos de testes e tipos de defeitos.

Seja π_{ijk} a probabilidade associada à casela (i,j,k), $i = 1,2,\ldots,c$, $j = 1,2,\ldots,d$ e $k = 1,2,\ldots,f$ e $\pi_{i..} = \sum_j \sum_k \pi_{ijk}$, $\pi_{.j.} = \sum_i \sum_k \pi_{ijk}$ e $\pi_{..k} = \sum_i \sum_j \pi_{ijk}$

as probabilidades marginais aleatórias, tal que

$$\sum_i \sum_j \sum_k \pi_{ijk} = 1 .$$

Na análise estatística desse modelo, podemos estar interessados em testar várias hipóteses, como por exemplo, as de independência e de não interação.

Em primeiro lugar, iremos desenvolver as hipóteses de independência do tipo:

a) independência mútua completa, formulada por

$$H0: \pi_{ijk} = \pi_{i..} \pi_{.j.} \pi_{..k} \qquad \begin{array}{l} i = 1,2,\ldots,c \\ j = 1,2,\ldots,d \\ k = 1,2,\ldots,f \end{array}$$

b) independência parcial, quando duas classificações são independentes da terceira, formulada por

$$H0: \pi_{ijk} = \pi_{i..} \pi_{.jk} \qquad \begin{array}{l} i = 1,2,\ldots,c \\ j = 1,2,\ldots,d \\ k = 1,2,\ldots,f \end{array}$$

c) independência condicional, para verificar se, dada uma classificação, as outras duas são indpendentes, formulada por

$$H0: \pi_{ijk} = \pi_{i.k} \pi_{.jk}/\pi_{..k} \qquad \begin{array}{l} i = 1,2,\ldots,c \\ j = 1,2,\ldots,d \\ k = 1,2,\ldots,f \end{array}$$

Essas hipóteses de independência são formuladas em termos de hipótese de tipo $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, onde K é uma matriz de constantes (t x cdf), A é uma matriz identidade (cdf x cdf), e $\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas (cdf x 1).

A soma dos quadrados para testar essa hipótese é dada por $F'S^{-1}F$, que possui distribuição χ^2 assintótica, com t graus de liberdade, sob H0.

Verifiquemos, por exemplo, um experimento 4x2x2, onde os dados representam a produção defeituosa de quatro máquinas A_i , i = 1,2,3,4 , classificada segundo dois tipos de testes T_j , j = 1,2 , em dois tipos de defeitos D_k, K = 1,2 .

Tabela 7 - Número de produtos manufaturados segundo tipo de defeito, tipo de teste e máquina.

	A_1		A_2		A_3		A_4		Total
	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2	
T_1	24	8	7	13	7	2	5	2	68
T_2	11	13	2	8	7	7	3	5	56
Total	35	21	9	21	14	9	8	7	124

Utilizaremos esse exemplo para ilustrar a aplicação dos testes de independência usando a metodologia de GSK.

a) Para testar se máquinas, testes e tipos de defeitos são independentes, a hipótese é

$$H0: \pi_{ijk} = \pi_{i..}\ \pi_{.j.}\ \pi_{..k}$$

ou, equivalentemente, $H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, onde

$$\underset{10\times16}{K} = \begin{pmatrix}
1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\
0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\
0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{pmatrix}$$

A é uma matriz identidade 16x16 , e

$\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

Com a matriz K da forma descrita acima, a estatística $F'S^{-1}F$ resulta 20,642 com 10 graus de liberdade, que é estatisticamente significante, ao nível de 5%. Isto significa que tipo de defeito, tipo de teste e máquina não são independentes.

De um modo geral, em tabelas cxdxf, a soma dos quadrados para testar a hipótese de independência mútua completa é uma χ^2 com cdf-c-d-f+2 graus de liberdade.

b) Se o teste de independência mútua completa resultar significante, isto é, com rejeição de H0, podemos estar interessados em verificar se duas classificações são independentes da terceira. No nosso exemplo, se quisermos testar se máquinas e testes indpendem do tipo de defeito, isto é, verificar independência parcial, formulada por

$$H0: \pi_{ijk} = \pi_{ij.}\ \pi_{..k}$$

ou equivalentemente, $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, K assume a forma

$$\underset{7\times17}{K} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

A é uma matriz identidade 16x16 , e

$\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

A estatística $F'S^{-1}F$ resulta 16,047 com 7 graus de liberdade, portanto significante ao nível de 5%, isto é, máquinas e tipos de testes não são independentes do tipo de defeito.

Se quisermos testar se máquinas e defeitos independem do tipo de teste, a hipótese é formulada por

$$H0: \pi_{ijk} = \pi_{i.k}\ \pi_{.j.}$$

ou, equivalentemente, $H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, onde

$$K_{7\times16} = \begin{pmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

A é uma matriz identidade 16x16, e

$\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

Assim, a estatística $F'S^{-1}F$ resulta 16,494 com 7 graus de liberdade, portanto significante ao nível de 5%, isto é, máquinas e defeitos não independem do tipo de teste.

Se ainda quisermos testar se tipo de teste e tipo de defeito independem de máquina, a hipótese adequada é

$$H0: \pi_{ijk} = \pi_{.jk}\, \pi_{i..}$$

ou, equivalentemente, $H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, onde

$$K_{9\times16} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

A é uma matriz identidade 16x16, e

$\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

A estatística $F'S^{-1}F$ resulta 19,158 com 9 graus de liberdade, portanto significante ao nível de 5%, isto é, tipo de teste e tipo de defeito não independem de máquina.

De um modo geral, em tabelas cxdxf , a soma dos quadrados para testar independência parcial entre (i,j) e k é dada por $F'S^{-1}F$, que é uma χ^2 assintótica com (cd-1) x (f-1) = cdf-cd-f+1 graus de liberdade, sob H0.

c) Algumas vezes, estamos interessados em testar a hipótese de que, dada uma classificação, as outras duas são independentes.

No exemplo, se quisermos verificar se, dado o tipo de defeito, máquinas e testes são independentes, formulamos a hipótese:

$$H0: \pi_{ijk} = \pi_{i.k}\, \pi_{.jk}/\pi_{..k}$$

ou, equivalentemente, H0: $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$, onde

$$\underset{6\times16}{K} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

A é uma matriz identidade 16x16, e

$\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

A soma dos quadrados para testar essa hipótese, $F'S^{-1}F$ resulta 14,052, com 6 graus de liberdade, portanto significante ao nível de 5%, isto é, dado o tipo de defeito, máquina e teste não são independentes.

De um modo geral, para tabelas cxdxf , a soma dos quadrados para o teste de independência condicional, entre i e j, dado k, possui distribuição χ^2 com f(c-1)(d-1) graus de liberdade, sob H0.

Para facilitar a construção das matrizes K , para hipóteses de independência em tabelas de três entradas, vamos desenvolver um procedimento geral, proposto por GRIZZLE e WILLIAMS (1973), e que consiste na elaboração de uma tabela de contrastes. Nessas tabelas, os contrastes dentro de um mesmo efeito não são mutuamente ortogonais; no entanto, os contrastes entre os efeitos são mutuamente ortogonais.

Vamos ilustrar tal procedimento através da construção de uma tabela de contrastes para um experimento 4x2x2:

Tabela 8 - Tabela de contrastes para um experimento 4x2x2.

Efeitos.	A_1				A_2				A_3				A_4			
	T_1		T_2		T_1		T_2		T_1		T_2		T_1		T_2	
	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2
	1	1	1	1	0	0	0	0	0	0	0	0	-1	-1	-1	-1
A	0	0	0	0	1	1	1	1	0	0	0	0	-1	-1	-1	-1
	0	0	0	0	0	0	0	0	1	1	1	1	-1	-1	-1	-1
T	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
D	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
	1	-1	-1	1	0	0	0	0	0	0	0	0	-1	-1	1	1
AT	0	0	0	0	1	1	-1	-1	0	0	0	0	-1	-1	1	1
	0	0	0	0	0	0	0	0	1	1	-1	-1	1	1	-1	-1
	1	-1	1	-1	0	0	0	0	0	0	0	0	-1	1	-1	1
AD	0	0	0	0	1	-1	1	-1	0	0	0	0	-1	1	-1	1
	0	0	0	0	0	0	0	0	1	-1	1	-1	-1	1	-1	1
DT	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1
	1	-1	-1	1	0	0	0	0	0	0	0	0	-1	1	1	-1
ATD	0	0	0	0	1	-1	-1	1	0	0	0	0	-1	1	1	-1
	0	0	0	0	0	0	0	0	1	-1	-1	1	-1	1	1	-1

Nessa tabela, notamos que, para testar a independência mútua completa entre máquina (A), teste (T) e defeito (D), a matriz K corresponde aos contrastes de 1 e -1 formados pelas linhas das interações AT, AD, DT e ATD; para testar as independências parciais, isto é, se:

- máquina (A), teste (T), independe de defeito (D),
- máquina (A), defeito (D), independe de teste (T),
- teste (T), defeito (D), independe de máquina (A),

as matrizes K correspondem aos contrastes de 1 e -1 formados pelas linhas das interações:

- AD, DT e ATD,
- AT, DT e ATD, e
- AT, AD e ATD respectivamente; e finalmente, para testar a independência condicional entre máquina (A) e teste (T), dado o tipo de defeito (D), a matriz K corresponde aos contrastes de 1 e -1 formados pelas linhas das interações AT e ATD.

Vamos agora tratar da hipótese de não interação. Observa-se que, até este momento, utilizamos a palavra não interação ligada ao problema de independência; assim, por exemplo, a independência mútua completa entre três respostas A, B e C implica em não interação entre A e B, A e C, B e C, e não interação entre A, B e C; a independência parcial entre (A, B) e C implica na não interação entre B e C, entre A e C, e não interação entre A, B e C; e a independência condicional entre A e B, dado C, implica na não interação entre A e B, e não interação entre A, B e C.

Em tabelas de classificação cruzada "três respostas, nenhum fator", o conceito de não interação está ligado à natureza da associação entre as respostas; assim, a hipótese de não interação entre três respsotas estabelece que alguma medida de associação entre duas respostas é constante sobre os níveis da terceira resposta; esta hipótese é chamada de não interação de segunda ordem.

Essa medida de não interação de segunda ordem, apresentada por BHAPKAR e KOCH,

$$\lambda_{ijk} = \frac{\pi_{ijk}\ \pi_{cdk}\ \pi_{cjk}\ \pi_{idf}}{\pi_{cjk}\ \pi_{idk}\ \pi_{ijf}\ \pi_{cdf}} \qquad \begin{array}{l} i = 1,2,..,c-1 \\ j = 1,2,..,d-1 \\ k = 1,2,..,f-1 \end{array}$$

satisfaz $\lambda_{ijk} = 1$, quando não existe interação de segunda ordem entre as três respostas. Se tomarmos os logaritmos naturais dessa relação, obtemos uma hipótese do tipo

$$H0:\ F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0 ,$$

onde K é uma matriz de 1 e -1 relativos aos $\ln \pi_{ijk}$.

Portanto, a soma dos quadrados para testar a não interação de segunda ordem, é dada por $F'S^{-1}F$, que possui distribuição χ^2 assintótica com $(c-1)(d-1)(f-1)$ graus de liberdade sob H0.

No nosso exemplo, a hipótese de não interação de segunda ordem é dada por:

$$H0:\ F(\pi) = \begin{cases} \ln\pi_{111} - \ln\pi_{112} - \ln\pi_{121} + \ln\pi_{122} - \ln\pi_{411} + \ln\pi_{412} + \ln\pi_{421} - \ln\pi_{422} = 0 \\ \ln\pi_{211} - \ln\pi_{212} - \ln\pi_{221} + \ln\pi_{222} - \ln\pi_{411} + \ln\pi_{412} + \ln\pi_{421} - \ln\pi_{422} = 0 \\ \ln\pi_{311} - \ln\pi_{312} - \ln\pi_{321} + \ln\pi_{322} - \ln\pi_{411} + \ln\pi_{412} + \ln\pi_{421} - \ln\pi_{422} = 0 \end{cases}$$

ou, de modo equivalente, por:

$$H0:\ F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0 \ , \qquad \text{onde}$$

$$\underset{3\times16}{K} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

A é uma matriz identidade 16x16, e

$\underset{\sim}{\pi}$ é o vetor de probabilidades das caselas.

A estatística $F'S^{-1}F$ resulta $\chi^2 = 1{,}8266$ com 3 graus de liberdade, portanto não significante ao nível de 5%, isto é, tipo de defeito, tipo de teste e máquina não interagem.

Notamos que essa matriz K corresponde aos contrastes formados pelas linhas da interação ATD na tabela dos contrastes do experimento 4x2x2.

Dessa forma, outras hipóteses de não interação podem ser formuladas, tomando-se matrizes K adequadas, na tabela de contrastes.

3.3.3. HIPÓTESE DE ASSOCIAÇÃO PARCIAL SOBRE A DIAGONAL PRINCIPAL.

Em dados classificados em tabelas de contingência dxd, nas situações "duas respostas, nenhum fator", ou "modelos mistos de ordem dois", podemos estar interessados na associação dos elementos na diagonal principal, como, por exemplo, verificar se a visão do olho direito e esquerdo se associam segundo os mesmos graus. A hipótese adequada para esse tipo de análise é dada por:

$$H0: \pi_{ii} = \pi_{i.}\ \pi_{.i} \qquad i = 1,2,\ldots,d$$

que equivale a uma hipótese do tipo

$$H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0$$

Vamos desenvolver essa hipótese utilizando os dados da Tabela 1. No caso, a hipótese é formulada por:

$$H0: \begin{cases} \pi_{11} = \pi_{1.}\pi_{.1} \\ \pi_{22} = \pi_{2.}\pi_{.2} \\ \pi_{33} = \pi_{3.}\pi_{.3} \\ \pi_{44} = \pi_{4.}\pi_{.4} \end{cases}$$

Fazendo a transformação logarítmica, obtemos:

$$H0: \begin{cases} \ln \pi_{11} - \ln \pi_{1.} - \ln \pi_{.1} = 0 \\ \ln \pi_{22} - \ln \pi_{2.} - \ln \pi_{.2} = 0 \\ \ln \pi_{33} - \ln \pi_{3.} - \ln \pi_{.3} = 0 \\ \ln \pi_{44} - \ln \pi_{4.} - \ln \pi_{.4} = 0 \end{cases}$$

que equivale a uma hipótese do tipo

$$H0: F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi} = 0 \quad , \qquad \text{onde}$$

$$\underset{12\times16}{A} = \begin{pmatrix} 1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1 \\ 1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\ 1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0 \\ 0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0 \\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0 \\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1 \end{pmatrix}$$

$$\underset{4\times12}{K} = \begin{pmatrix} 1&0&0&0&-1&0&0&0&-1&0&0&0 \\ 0&1&0&0&0&-1&0&0&0&-1&0&0 \\ 0&0&1&0&0&0&-1&0&0&0&-1&0 \\ 0&0&0&1&0&0&0&-1&0&0&0&-1 \end{pmatrix}$$

Os estimadores de $F(\underset{\sim}{\pi})$ e de sua matriz de covariância obtidos com os dados foram:

$$F' = \begin{pmatrix} 1,1039 & 0,8133 & 0,7663 & 1,7127 \end{pmatrix}$$

$$S = 10^{-4} \begin{pmatrix} 3,0044 & -0,1877 & -0,8373 & -0,7010 \\ -0,1877 & 2,3758 & 0,0897 & 0,4586 \\ -0,8373 & 0,0897 & 2,0013 & 0,5600 \\ -0,7010 & -0,4586 & 0,5600 & 9,2522 \end{pmatrix}$$

A estatística $\chi^2 = F'S^{-1}F$ assume o valor 17909,0 com 4 graus de liberdade, portanto rejeitamos a hipótese de não associação sobre a diagonal principal, isto é, as distâncias de visão dos olhos direito e esquerdo se associam segundo os mesmos graus.

De um modo geral, em tabelas de contingência dxd, a soma dos quadrados para testar a hipótese de associação parcial sobre a diagonal principal é fornecida por $F'S^{-1}F$, que possui distribuição χ^2 assintótica com d graus de liberdade, sob H0.

CAPÍTULO IV

APLICAÇÕES A PROBLEMAS DE DIVERSAS POPULAÇÕES

4.1. *INTRODUÇÃO*

Neste capítulo, iremos tratar da análise de dados categóricos provindos de diversas populações ou subpopulações. Esses dados serão classificados em vários tipos de tabelas de classificação cruzada.

Já apresentamos uma delas, no Capítulo II, referente a motoristas classificados de acordo com "grupo, tipo de acidente e grau de ferimento" (Tabela 4). Nessa tabela, os sujeitos pertencentes a todas as combinações de níveis dos fatores "grupo" e "tipo de acidente" foram classificados de acordo com a "gravidade do ferimento".

Assim, de um modo geral, uma tabela de classificação cruzada contém as frequências com que indivíduos pertencentes às várias combinações dos níveis dos fatores são classificados em várias combinações das categorias de respostas. Deste modo, as marginais totais dependentes da combinação dos níveis dos fatores e o tamanho da amostra são números fixos conhecidos antes da realização do experimento.

Identificando pelos termos "fator" e "resposta" as tabelas de classificação cruzada, vamos tratar dos modelos seguintes:

"um fator, uma resposta"

"multifator, uma resposta"

"um fator, multiresposta"

"multifator, multiresposta".

Para cada um destes modelos existem análises estatísticas apropriadas. Por exemplo, em modelos "um fator, uma resposta", podemos estar interessados em testar a homogeneidade entre os níveis do fator; como veremos adiante, essa hipótese pode ser formulada como $H0: F(\pi) = A\pi = 0$. Para outros modelos, podemos estar interessados em testar hipóteses análogas às que aparecem na análise de variância de dados contínuos por modelos lineares; assim, GSK propõem o ajuste do modelo linear $F(\pi) = X\beta$ às funções que pertencem ou à classe $F(\pi) = A\pi$ ou à $F(\pi) = K \ln A\pi$.

A formulação de funções, do tipo linear ou log-linear, deve ser verificada com cuidado para cada situação exposta acima. Parece-nos que a dificuldade do método de GSK consiste na identificação do tipo de função apropriada para os dados e consequentemente na construção de matrizes A e K adequadas.

A seguir, vamos descrever os modelos de tabelas de classificação cruzada citadas acima, verificar para cada um as hipóteses de maior interesse e a solução através da metodologia de GSK.

4.2. *HIPÓTESE DE HOMOGENEIDADE ENTRE POPULAÇÕES.*

Vamos considerar o modelo geral de classificação cruzada, em que as amostras são retiradas de s populações ou subpopulações, cada amostra sendo de tamanho $n_{i.}$, para $i = 1,2,\ldots,s$. Os elementos de cada amostra são classificados em r categorias de respostas, sendo n_{ij} o número de elementos da i-ésima população classificados na j-ésima categoria de resposta.

Utilizando os termos "fator" e "resposta", vamos identificar o modelo geral descrito acima como um modelo "um fator, uma resposta", cada um com s e r níveis respectivamente.

Em modelos deste tipo, a hipótese de principal interesse é a da homogeneidade entre as populações ou subpopulações, que equivale à hipótese de não-interação entre fator e resposta, isto é,

$$H0: \pi_{ij} = \pi_{j.} \qquad j = 1,2,\ldots,r$$

onde π_j é um valor comum para todas as populações dentro da j-ésima categoria de resposta.

Consideremos um exemplo deste modelo.

Tabela 9 - Classificação de indivíduos pertencentes aos grupos I e II segundo graus obtidos em um teste de habilidade.

Grupos	Graus				Total
	A	B	C	D	
I	121	120	79	33	353
II	118	95	121	30	364
Total	239	215	200	63	717

Queremos examinar a igualdade das proporções dos graus de habilidade nos dois grupos, e portanto,

$$H0: \pi_{1j} = \pi_{2j} = \pi_j \qquad j = 1,2,3,4 \ , \quad \text{ou}$$

$$H0: \begin{cases} \pi_{11} = \pi_{21} \\ \pi_{12} = \pi_{22} \\ \pi_{13} = \pi_{23} \\ \pi_{14} = \pi_{24} \end{cases} \quad \text{que equivale } H0: \begin{cases} \pi_{11} - \pi_{21} = 0 \\ \pi_{12} - \pi_{22} = 0 \\ \pi_{13} - \pi_{23} = 0 \end{cases}$$

a qual pode ser escrita como $H0: F(\pi) = \underset{uxrs}{A}\,\pi = 0$, onde A é uma matriz dada por:

$$\underset{3x8}{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \quad \text{e}$$

$$\underset{1x8}{\pi'} = \left(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{24}\right)$$

A matriz de covariância de F é estimada por $S = A\,V(\underset{\sim}{p})\,A'$, onde $V(\underset{\sim}{p})$ é o estimador consistente para a matriz de covariância do estimador de π, resultando:

$$S = 10^{-3} \begin{pmatrix} 1,239 & -0,563 & -0,514 \\ -0,563 & 1,165 & -0,453 \\ -0,514 & -0,453 & 1,103 \end{pmatrix}$$

O teste de homogeneidade é dado por $F'S^{-1}F$, que possui distribuição assintótica χ^2 com u graus de liberdade sob H0.

Assim,

$$F'S^{-1}F = \left(0,019 \quad 0,079 \quad -0,110\right) \; 10^3 \begin{pmatrix} 2,57 & 2,05 & 2,05 \\ 2,02 & 2,64 & 2,02 \\ 2,05 & 2,02 & 2,67 \end{pmatrix} \begin{pmatrix} 0,019 \\ 0,079 \\ -0,110 \end{pmatrix}$$

$= 11,94$, com 3 graus de liberdade.

Para testar a mesma hipótese, RAO obteve, usando estimador de máxima verossimilhança, o valor de $\chi^2 = 11,77$ com 3 graus de liberdade. Concluímos que, ao nível de significância de 5%, existem diferenças entre as proporções de graus de habilidade entre os dois grupos.

4.3. HIPÓTESE LINEAR GERAL NO MODELO $F(\pi) = X\beta$.

Vamos descrever de início os tipos de tabelas de classificação cruzada com um número de entradas superior ou igual a três e que apresentam os dados categóricos provindos de diversas populações. A identificação das variáveis é feita em termos de "fator" e "resposta" favorecendo a formulação de hipóteses adequadas.

Um primeiro modelo denominado "multifator, uma resposta" traduz os resultados experimentais, de indivíduos agrupados nas diversas combinações possíveis dos níveis dos fatores, em várias categorias da resposta.

De um modo particular, se considerarmos três variáveis, sendo dois fatores A e B com c e d níveis respectivamente, e uma resposta C com r categorias, as populações ou subpopulações das quais serão extraídas as amostras, serão compostas por $s = c \times d$ combinações dos níveis dos fatores A e B.

Assim, temos uma tabela do tipo:

Tabela 10 - Tabela "dois fatores, uma resposta".

Subpopulações		Categorias de respostas					Total
		C_1	C_2	C_3	.....	C_r	
A_1	B_1	π_{111}	π_{112}	π_{113}	...	π_{11r}	1
	B_2	π_{121}	π_{122}	π_{123}	...	π_{12r}	1
	.	.	.	.	...	.	.
	B_d	π_{1d1}	π_{1d2}	π_{1d3}	...	π_{1dr}	1
A_2	B_1	π_{211}	π_{212}	π_{213}	...	π_{21r}	1
	B_2	π_{221}	π_{222}	π_{223}	...	π_{22r}	1
	.	.	.	.	...	.	.
	B_d	π_{2d1}	π_{2d2}	π_{2d3}	...	π_{2dr}	1
.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
A_c	B_1	π_{c11}	π_{c12}	π_{c13}	...	π_{c1r}	1
	B_2	π_{c21}	π_{c22}	π_{c23}	...	π_{c2r}	1
	.	.	.	.	...	.	.
	B_d	π_{cd1}	π_{cd2}	π_{cd3}	...	π_{cdr}	1

Aqui, π_{ijk} é a probabilidade de um indivíduo do i-ésimo nível de A e j-ésimo nível de B ser classificado na k-ésima categoria da resposta C , e assim,

$$\sum_{k=1}^{r} \pi_{ijk} = 1 \quad \text{para todo } (i,j).$$

Denotamos por $\underset{\sim}{\pi}'_{ij} = \left(\pi_{ij1}, \pi_{ij2}, \pi_{ij3}, \ldots, \pi_{ijr}\right)$ o vetor de probabilidades de cada multinomial (i,j) e por

$$\underset{\sim}{\pi}' = \left(\underset{\sim}{\pi}'_{11}, \underset{\sim}{\pi}'_{12}, \ldots, \underset{\sim}{\pi}'_{cd}\right)$$ o vetor composto de probabilidades para a tabela toda.

No Capítulo II, já mostramos um exemplo desse modelo, quando apresentamos os dados de SCHOTZ, referentes ao número de motoristas classificados segundo "grupo, tipo de acidente e gravidade do ferimento".

Em modelos deste tipo, temos interesse em verificar os efeitos dos fatores A e B na resposta C, além da interação entre os fatores A e B. A interação entre os dois fatores A e B e a resposta C significa aqui homogeneidade entre as subpopulações.

Um segundo modelo denominado "um fator, multiresposta" resulta de um experimento no qual indivíduos pertencentes a um dos s níveis do fator A são classificados em todas as combinações possíveis das categorias de respostas. De um modo particular, se considerarmos três variáveis, sendo um fator A com s níveis, e duas respostas B e C com c e d categorias, respectivamente, temos r = cxd categorias de respostas, que são combinações das categorias de B e C.

Tabela 11 - Tabela "um fator, duas respostas".

Subpopulações	Categorias de respostas								
	B_1C_1	B_1C_2	...	B_1C_d	...	B_cC_1	B_cC_2	...	B_cC_d
A_1	π_{111}	π_{112}	...	π_{11d}	...	π_{1c1}	π_{1c2}	...	π_{1cd}
A_2	π_{211}	π_{212}	...	π_{21d}	...	π_{2c1}	π_{2c2}	...	π_{2cd}
.	.	.	...	.	...	.	.	...	.
A_s	π_{s11}	π_{s12}	...	π_{s1d}	...	π_{sc1}	π_{sc2}	...	π_{scd}

Nessa tabela, π_{ijk} é a probabilidade de um indivíduo do i-ésimo nível de A apresentar resposta B_jC_k e, portanto,

$$\sum_{j=1}^{c} \sum_{k=1}^{d} \pi_{ijk} = 1 \quad \text{para todo } i = 1,2,\ldots,s.$$

O vetor de probabilidades para cada população multinomial i é denotado por $\underset{\sim}{\pi}_i' = \left(\pi_{i11}, \pi_{i12}, \ldots, \pi_{icd}\right)$, e o vetor composto de probabilidades da tabela toda, então, será $\underset{\sim}{\pi}' = \left(\pi_1', \pi_2', \ldots, \pi_s'\right)$.

Neste modelo, o interesse principal é verificar o efeito do fator A nas respostas B e C, e também verificar a associação entre as respostas B e C.

Um terceiro modelo, denominado "multiresposta, multifator", traduz resultados de um experimento no qual indivíduos classificados nas combinações de níveis dos fatores apresentam resultados que são combinações das categorias das respostas. Por exemplo, se considerarmos uma tabela "dois fatores, duas respostas", com fatores A e B possuindo a e b níveis respectivamente, com respostas C e D com c e d categorias respectivamente, o número de subpopulações será $s = axb$ e o número de categorias de respostas será $r = cxd$.

Assim, π_{ijhk} é a probabilidade de um indivíduo pertencente ao i-ésimo nível de A, j-ésimo nível de B, ser classificado na h-ésima resposta C e k-ésima resposta D, e de tal modo que

$$\sum_{h=1}^{c} \sum_{k=1}^{d} \pi_{ijhk} = 1 \quad \text{para} \quad i=1,2,\ldots,a;\ j=1,2,\ldots,b.$$

O vetor de probabilidades para cada multinomial (i,j) é denotado por:

$$\underset{\sim}{\pi}'_{ij} = \left\{\pi_{ij11}, \pi_{ij12}, \ldots, \pi_{ij1d}; \pi_{ij21}, \pi_{ij22}, \ldots, \pi_{ij2d}; \ldots; \pi_{ijc1}, \ldots, \pi_{ijcd}\right\}$$

e o vetor composto para a tabela toda é dado por:

$$\underset{\sim}{\pi}' = \left\{\pi'_{11}, \pi'_{12}, \ldots, \pi'_{ab}\right\}$$

O método de GSK permite uma formulação geral para as soluções dos problemas propostos acima nos diversos modelos descritos. Deste modo, o objetivo, neste ponto, é a formulação de u funções linearmente independentes $f_m(\pi)$, $m = 1,2,\ldots,u$ adequadas aos dados, em cada caso particular. Vimos, no Capítulo II, que essas funções compõem vetores do tipo $F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi}$ ou $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi}$. Vamos verificar, nos modelos relatados acima, quais as funções que são do tipo $A\underset{\sim}{\pi}$ ou $K \ln A\underset{\sim}{\pi}$.

4.3.1. CASO $F(\pi) = A\pi$.

Para análise dos modelos descritos, podemos formular funções que são denominadas escores médios, baseadas em escores atribuídos às categorias de respostas.

Vamos considerar o modelo de classificação cruzada "dois fatores, uma resposta", com fatores A e B, cada um com c e d níveis respectivamente, e resposta C, com r categorias ordenadas. Os escores são r valores, arbitrários ou não, correspondentes a cada categoria de resposta, e denotados por a_{ijk}, $i = 1,2,\ldots,c$, $j = 1,2,\ldots,d$ e $k = 1,2,\ldots,r$. Os escores médios são funções $\bar{a}_{ij}\bar{p}_{ij}$,

onde $\quad a'_{ij} = \{a_{ij1}, a_{ij2}, \ldots, a_{ijr}\}$ e

$$p'_{ij} = \{p_{ij1}, p_{ij2}, \ldots, p_{ijr}\} .$$

Assim, para a tabela toda temos cxd escores médios, que compõem o vetor do tipo $F(\underset{\sim}{p}) = A\,\underset{\sim}{p}$, sendo A uma matriz do tipo

$$\underset{cdxcdr}{A} = \begin{pmatrix} \{a_{111} a_{112} \cdot\cdot a_{11r}\} & 0 & \cdots & 0 \\ 0 & \{a_{121} a_{122} \cdot\cdot a_{12r}\} & & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & & \{a_{cd1} a_{cd2} \cdot\cdot a_{cdr}\} \end{pmatrix}$$

e $\underset{\sim}{p}$ é o vetor das probabilidades estimadas de $\underset{\sim}{\pi}$.

Muitas vezes empregam-se escores arbitrários para cada categoria de respostas e idênticos para todas as combinações (i,j); assim, a matriz A é composta por vetores $a' = \{a_1, a_2, \ldots, a_r\}$ na diagonal principal.

Em particular, se a resposta C possui apenas duas categorias (por exemplo, "morto, sobrevivido"), normalmente atribuem-se os escores 0 e 1 (0 para insucesso, e 1 para sucesso), e os escores médios são as probabilidades da resposta "sucesso".

O vetor dos escores médios $F(\underset{\sim}{p}) = A\,\underset{\sim}{p}$ será ob-

jeto da análise de variância. Para estudar a variação dos escores médios é ajustado o modelo linear $F(\pi) = X\xi$, sendo X uma matriz de planejamento $c d x (c+d+1)$ e ξ o vetor $(c+d+1)x1$ de parâmetros desconhecidos.

4.3.2. CASO $F(\pi) = K \ln A\pi$.

Muitos problemas surgidos nos vários modelos de classificação cruzada já descritos, podem ser resolvidos analisando-se funções logarítmicas das probabilidades das caselas, funções essas que compõem o vetor $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi}$.

Vamos tratar aqui de dois problemas que julgamos de grande interesse: o primeiro, de grande utilidade em pesquisas de estatística vital, trata da estimação e análise das taxas de sobrevivência através do método de GSK, o que foi proposto por KOCH, JOHNSON e TOLLEY (1972); o segundo trata da formulação e análise dos logits, quando as tabelas de classificação cruzada apresentam respostas ordenadas. Esse problema foi proposto por WILLIAMS e GRIZZLE (1972) e resultou dos trabalhos de THURSTONE (1959) e BOCK e JONES (1967).

a) Para a estimação e análise das taxas de sobrevivência, consideremos uma situação experimental envolvendo pacientes com uma determinada doença, classificados de acordo com a idade (fator A , com dois níveis), estágio da doença (fator B, com dois níveis) e tipo de tratamento (fator C, com dois níveis), observados ao longo de um período (fator D, com três níveis). É anotado, ao fim de cada ano, o resultado do desenlace, em termos de número de óbitos durante o ano, número de sobreviventes ao ano, e número de perdidos da observação (resposta E, com três níveis).

Os resultados desse experimento são resumidos em uma tabela do tipo "dois fatore, uma resposta", sendo o primeiro fator chamado "grupo F_i", $i = 1,2,\ldots,8$, composto de oito níveis constituídos pelas combinações dos níveis dos fatores A,B e C; o segundo fator é o tempo de observação D_j, $j = 1,2,3$; e a resposta E_k , $k = 1,2,3$, isto é, sobrevivente, morto e perdido da observação.

Temos, portanto, uma tabela de classificação cruzada constituída de 24 populações (s = 24), fornecidas pelas combinações entre níveis de grupos F_i e tempo D_j, e três respostas (r = 3), que apresentaremos a seguir, apenas no que diz respeito a 12 populações, formadas pelas combinações A_l B_j C_k .

Tabela 12 - Número de pacientes pertencentes aos grupos F_i, observados durante três anos e classificados segundo o desenlace.

Grupos (F_i)			Tempo de observ.	Resultado: Sobrev.	Mortos no ano	Perdidos da observ.	Total
A_1	B_1	C_1	D_1	n_{111}	n_{112}	n_{113}	$n_{11.}$
			D_2	n_{121}	n_{122}	n_{123}	$n_{12.}$
			D_3	n_{131}	n_{132}	n_{133}	$n_{13.}$
		C_2	D_1	n_{211}	n_{212}	n_{213}	$n_{21.}$
			D_2	n_{221}	n_{222}	n_{223}	$n_{22.}$
			D_3	n_{231}	n_{232}	n_{233}	$n_{23.}$
	B_2	C_1	D_1	n_{311}	n_{312}	n_{313}	$n_{31.}$
			D_2	n_{321}	n_{322}	n_{323}	$n_{32.}$
			D_3	n_{331}	n_{332}	n_{333}	$n_{33.}$
		C_2	D_1	n_{411}	n_{412}	n_{413}	$n_{41.}$
			D_2	n_{421}	n_{422}	n_{423}	$n_{42.}$
			D_3	n_{431}	n_{432}	n_{433}	$n_{43.}$

Para essa situação, o logaritmo do estimador da razão de sobrevivência para cada grupo, é fornecido por:

$$F(\underset{\sim}{p}_i) = K^* \ln A^* \underset{\sim}{p}_i$$

com $\underset{6x9}{A} = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & A_1 & 0 \\ 0 & 0 & A_1 \end{pmatrix}$ onde $A_1 = \begin{pmatrix} 1 & 0 & 1/2 \\ 1 & 1 & 1/2 \end{pmatrix}$

$$\underset{1x6}{K} = \left(1 \ -1 \quad 1 \ -1 \quad 1 \ -1\right)$$

$$\underset{\sim}{p}_i' = \left(p_{i11}, p_{i12}, p_{i13}\right) \qquad i = 1,2,\ldots,8$$

ou

$$\log G_i = K^* \ln A^* \underset{\sim}{p}_i = \log \left\{ \prod_{j=1}^{3} \frac{p_{ij1}+0,5\ p_{ij3}}{p_{ij1}+p_{ij2}+0,5\ p_{ij3}} \right\}$$

Para os oito grupos, o vetor dos log estimadores da razão de sobrevivência é fornecido por:

$$F(\underset{\sim}{p}) = K \ln A\ \underset{\sim}{p} = \left(\log G_1, \log G_2, \ldots, \log G_8\right)$$

com:

$$\underset{48x72}{A} = \begin{pmatrix} A^* & 0 & 0 & \ldots & 0 \\ 0 & A^* & 0 & \ldots & 0 \\ 0 & 0 & A^* & \ldots & 0 \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & A^* \end{pmatrix}, \quad \underset{8x48}{K} = \begin{pmatrix} K^* & 0 & 0 & \ldots & 0 \\ 0 & K^* & 0 & \ldots & 0 \\ 0 & 0 & K^* & \ldots & 0 \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & K^* \end{pmatrix} \quad e$$

$$\underset{\sim}{p}' = \left(\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_8'\right) .$$

Estimadas as razões de sobrevivência de cada grupo, é possível, através desse método de análise, ajustando-se um modelo linear $F(\underset{\sim}{\pi}) = X\beta$, com matrizes de planejamen

tos X adequadas, testar o efeito dos fatores idade, estágio da doença e tipo de tratamento, nas taxas de sobrevivência, as interações possíveis entre fatores, e ainda verificar se as probabilidades de sobrevivência para cada ano após o diagnóstico, seguem uma curva, linear ou exponencial.

b) Iremos estudar aqui a estimação e análise dos logits. Para melhor entendimento de como estimar os logits, vamos considerar o modelo "dois fatores, uma resposta" já descrito. Neste modelo, os fatores A e B possuem c e d níveis respectivamente, e a resposta C , r categorias.

O vetor de probabilidades de cada multinomial será denotado por $\underset{\sim}{\pi}'_{ij} = \left(\pi_{ij1}, \pi_{ij2}, \ldots, \pi_{ijr}\right)$. As probabilidades acumuladas para cada (i,j) são:

$$\phi_{ij1} = \pi_{ij1}$$
$$\phi_{ij2} = \pi_{ij1} + \pi_{ij2}$$
$$\phi_{ij3} = \pi_{ij1} + \pi_{ij2} + \pi_{ij3}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\phi_{ij(r-1)} = \pi_{ij1} + \pi_{ij2} + \pi_{ij3} + \ldots + \pi_{ij(r-1)}$$
$$\phi_{ijr} = 1$$

Definem-se os logits das probabilidades acumuladas como:

$$l_{ij1} = \ln \left\{\phi_{ij1}/(1-\phi_{ij1})\right\}$$

$$l_{ij2} = \ln \left\{\phi_{ij2}/(1-\phi_{ij2})\right\}$$

. . . . . .

$$l_{ij(r-1)} = \ln \left\{\phi_{ij(r-1)}/(1-\phi_{ij(r-1)})\right\}$$

que compõem o vetor dos (r-1) logits de cada multinomial (i,j), isto é,

$$\underset{1\times(r-1)}{\check{l}_{ij}} = \left\{l_{ij1}, l_{ij2}, \ldots, l_{ij(r-1)}\right\}$$

que pode ser escrito por

$$l_{ij} = K^* \ln A^* \pi_{ij}$$

com

$$\underset{(r-1)\times 2(r-1)}{K^*} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix}$$

$$\underset{2(r-1)\times r}{A^*} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & \ldots & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & 1 & 1 & \ldots & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

O vetor dos logits composto para a tabela toda é denotado por

$$\underset{1xcd(r-1)}{\bar{l}} = \left\{\bar{l}_{11}, \bar{l}_{12}, \ldots, \bar{l}_{1d}; \ldots; \bar{l}_{c1}, \bar{l}_{c2}, \ldots, \bar{l}_{cd}\right\}$$

e, portanto, $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi}$, representa os logits para a tabela toda, com

$$\underset{cd(r-1)x2cd(r-1)}{K} = \begin{pmatrix} K^* & 0 & 0 & \ldots & 0 \\ 0 & K^* & 0 & \ldots & 0 \\ 0 & 0 & K^* & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & K^* \end{pmatrix}$$

$$\underset{2cd(r-1)xcdr}{A} = \begin{pmatrix} A^* & 0 & 0 & \ldots & 0 \\ 0 & A^* & 0 & \ldots & 0 \\ 0 & 0 & A^* & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & A^* \end{pmatrix}$$

O vetor dos estimadores dos logits, representado por ℓ , é obtido por $F(\underset{\sim}{p}) = K \ln A \underset{\sim}{p}$, onde $\underset{\sim}{p}$ é o estimador de $\underset{\sim}{\pi}$.

Para melhor visualização, vamos colocar os logits nas caselas da tabela "dois fatores, uma resposta".

Tabela 13 - Tabela dos logits.

Subpopulações		Categorias de resposta					
		C_1	C_2	C_3	...	$C_{(r-1)}$	C_r
A_1	B_1	ℓ_{111}	ℓ_{112}	ℓ_{113}	...	$\ell_{11(r-1)}$	-
	B_2	ℓ_{121}	ℓ_{122}	ℓ_{123}	...	$\ell_{12(r-1)}$	-
	.	.	.	.	...	.	.
	B_d	ℓ_{1d1}	ℓ_{1d2}	ℓ_{1d3}	...	$\ell_{1d(r-1)}$	-
A_2	B_1	ℓ_{211}	ℓ_{212}	ℓ_{213}	...	$\ell_{21(r-1)}$	-
	B_2	ℓ_{221}	ℓ_{222}	ℓ_{223}	...	$\ell_{22(r-1)}$	-
	.	.	.	.	...	.	.
	B_d	ℓ_{2d1}	ℓ_{2d2}	ℓ_{2d3}	...	$\ell_{2d(r-1)}$	-
: : : :	: : :	: : :	: : :	: : :	: : :	: : :	: : :
A_c	B_1	ℓ_{c11}	ℓ_{c12}	ℓ_{c13}	...	$\ell_{c1(r-1)}$	-
	B_2	ℓ_{c21}	ℓ_{c22}	ℓ_{c23}	...	$\ell_{c2(r-1)}$	-
	.	.	.	.	...	.	.
	B_d	ℓ_{cd1}	ℓ_{cd2}	ℓ_{cd3}	...	$\ell_{cd(r-1)}$	-

Expressando os estimadores dos logits de cada multinomial separadamente como um modelo aditivo da forma

$$\ell_{ijh} = \mu_h + \alpha^*_{ih} + \beta^*_{jh} \qquad h = 1,2,\ldots,(r-1) \ ,$$

o mesmo pode ser escrito sob a forma

$$\begin{pmatrix} \ell_{ij1} \\ \ell_{ij2} \\ \ell_{ij3} \\ \vdots \\ \ell_{ij(r-1)} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_{(r-1)} \end{pmatrix} + \begin{pmatrix} \alpha^*_{i1} \\ \alpha^*_{i2} \\ \alpha^*_{i3} \\ \vdots \\ \alpha^*_{i(r-1)} \end{pmatrix} + \begin{pmatrix} \beta^*_{j1} \\ \beta^*_{j2} \\ \beta^*_{j3} \\ \vdots \\ \beta^*_{j(r-1)} \end{pmatrix}$$

onde

μ_h é a média do h-ésimo logit;

α^*_{ih} é o efeito do i-ésimo nível de A no h-ésimo logit;

β^*_{jh} é o efeito do j-ésimo nível de B no h-ésimo logit.

Vamos ilustrar a análise dos escores médios e dos logits utilizando os dados apresentados no Capítulo II, Tabela 4, referente a um estudo feito por SCHOTZ, em 1966, sobre a relação existente entre "grupo motorista, tipo de acidente e gravidade do ferimento produzido".

Essa tabela constitui o resultado de um experimento do tipo "dois fatores, uma resposta", sendo os fatores "grupo motorista" (só e acompanhado) e "tipo de acidente" (capotamento e não-capotamento), e a resposta, "gravidade do ferimento" (leve, moderado, severo e crítico). Em termos do modelo geral de classificação cruzada, o número de populações

é igual a quatro, e o número de categorias de resposta é também quatro.

Nessa tabela, π_{ijk} é a probabilidade de um indivíduo classificado no i-ésimo tipo de motorista, $i = 1,2$, e j-ésimo tipo de acidente, $j = 1,2$, sofrer ferimento do k-ésimo grau, $k = 1,2,3,4$. Assim,

$$\pi'_{ij} = \left(\pi_{ij1}, \pi_{ij2}, \pi_{ij3}, \pi_{ij4}\right)$$

é o vetor de probabilidades de cada multinomial, tal que

$$\sum_{k=1}^{4} \pi_{ijk} = 1 \text{ , e}$$

$$\underset{\sim}{\pi}' = \left(\pi'_{11}, \pi'_{12}, \pi'_{21}, \pi'_{22}\right)$$

é o vetor composto de probabilidades para a tabela toda.

Analisamos os dados utilizando primeiramente os escores arbitrários 1,2,3 e 4 para as categorias de resposta leve, moderado, severo e crítico respectivamente. Desse modo, os escores médios compõem um vetor formulado por:

$$F(\underset{\sim}{\pi}) = \begin{cases} 1\,\pi_{111} + 2\,\pi_{112} + 3\,\pi_{113} + 4\,\pi_{114} \\ 1\,\pi_{121} + 2\,\pi_{122} + 3\,\pi_{123} + 4\,\pi_{124} \\ 1\,\pi_{211} + 2\,\pi_{212} + 3\,\pi_{213} + 4\,\pi_{214} \\ 1\,\pi_{221} + 2\,\pi_{222} + 3\,\pi_{223} + 4\,\pi_{224} \end{cases}$$

ou por $F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi}$, onde

$$\underset{4x16}{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 \end{pmatrix}$$

Os estimadores, de $F(\underset{\sim}{\pi})$ e da matriz de covariância estimada, resultam em:

$$F'(\underset{\sim}{p}) = \begin{pmatrix} 3{,}0135 & 2{,}4093 & 3{,}0882 & 2{,}4545 \end{pmatrix}$$

$$S = 10^{-4} \begin{pmatrix} 1{,}9328 & 0 & 0 & 0 \\ 0 & 0{,}6409 & 0 & 0 \\ 0 & 0 & 1{,}4585 & 0 \\ 0 & 0 & 0 & 0{,}7326 \end{pmatrix}$$

Foi considerado o modelo linear $F(\underset{\sim}{\pi}) = X\xi$, com

$$\underset{4x3}{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \qquad \text{e}$$

$$\xi' = \begin{pmatrix} \mu, \alpha_1, \beta_1 \end{pmatrix} \quad \text{com} \quad \sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{2} \beta_j = 0 \ .$$

Os estimadores $\underset{\sim}{b}$ e a sua matriz de covariância resultaram:

$$\underset{\sim}{b} = \begin{pmatrix} 2,8399 \\ -0,1285 \\ 0,3021 \end{pmatrix}$$

$$V_{\underset{\sim}{b}} = 10^{-4} = \begin{pmatrix} 0,6051 & -0,1219 & 0,4832 \\ -0,1219 & 0,2821 & -0,1602 \\ 0,4832 & -0,1602 & 0,6434 \end{pmatrix}$$

O teste do ajuste do modelo resultou $\chi^2 = 1833,12$ com 1 grau de liberdade, portanto, significante, ao nível de 5%. Isto quer dizer que o modelo não descreve adequadamente os dados e, portanto, torna-se necessária a procura de funções convenientes.

Uma reanálise dos dados foi realizada através dos logits.

Consideramos para cada multinomial (i,j),

$$\vec{l}_{ij} = \left(l_{ij1}, l_{ij2}, l_{ij3}\right)$$

o vetor dos logits, que pode ser escrito por

$$l_{ij} = K^* \ln A^* \pi_{ij} \qquad \text{onde}$$

$$\underset{3x6}{K^*} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$\underset{6x4}{A^*} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Para a tabela toda, $\underset{\sim}{l}' = \left(l'_{11}, l'_{12}, l'_{21}, l'_{22}\right)$ é o vetor composto dos logits, e pode ser escrito por

$\underset{\sim}{l} = K \ln A\underset{\sim}{\pi}$, sendo K uma matriz diagonal principal, com as matrizes K^* na diagonal principal, e A é uma matriz (24x16) diagonal principal, com as matrizes A^* na diagonal principal.

O estimador das funções $F(\underset{\sim}{\pi})$ resulta:

$$F'(\underset{\sim}{p}) = \begin{pmatrix} -4,8060 & -1,2605 & 1,1048 & -2,2535 & 0,3028 & 1,9938 \\ -5,1667 & -1,5491 & 0,9755 & -2,4875 & 0,2344 & 1,9605 \end{pmatrix}$$

Consideramos os logits para cada multinomial expressos pelo modelo aditivo:

$$\ell_{ijh} = \mu_h + \alpha^*_{ih} + \beta^*_{jh} \qquad h = 1,2,3$$

ou pelo vetor

$$\begin{pmatrix} \ell_{ij1} \\ \ell_{ij2} \\ \ell_{ij3} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \alpha^*_{i1} \\ \alpha^*_{i2} \\ \alpha^*_{i3} \end{pmatrix} + \begin{pmatrix} \beta^*_{j1} \\ \beta^*_{j2} \\ \beta^*_{j3} \end{pmatrix}$$

onde μ_h é um valor constante para as categorias de respostas;

α^*_{ih} é o efeito do i-ésimo nível do fator "grupo motorista" no h-ésimo logit;

β^*_{jh} é o efeito do j-ésimo nível do fator "tipo de acidente" no h-ésimo logit.

Para estudar a variação dos logits, ajustamos um modelo linear $E\{F(\underset{\sim}{p})\} = X\xi$, com

$$\underset{12x9}{X} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 \end{pmatrix} \quad \text{e}$$

$$\xi = \left(\mu_1, \mu_2, \mu_3, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3\right)$$

com μ_h o valor constante para cada categoria de resposta, $h = 1,2,3$;

α_h o efeito do fator "grupo motorista" no h-ésimo logit;

β_h o efeito do fator "tipo de acidente" no h-ésimo logit.

Utilizando as fórmulas

$$\underset{\sim}{b} = (X'S^{-1}X)^{-1}X'S^{-1}F \quad e \quad V(\underset{\sim}{b}) = (X'S^{-1}X)^{-1},$$

foram encontrados os estimadores dos parâmetros e os seus desvios padrões, dados na tabela abaixo:

Tabela 14 - Estimadores e desvios padrões dos parâmetros.

logit	μ		α		β	
	estim.	desvio	estim.	desvio	estim.	desvio
1	-3,684	0,0814	0,1191	0,0256	-1,306	0,0814
2	-0,569	0,0183	0,0538	0,0137	-0,837	0,0183
3	1,506	0,0186	0,0328	0,0178	-0,470	0,0187

O teste do ajuste do modelo calculado por

$$SS\ \{F(\underset{\sim}{\pi}) = X\xi\} = F'S^{-1}F - \underset{\sim}{b}'(X'S^{-1}X)^{-1}\underset{\sim}{b}$$

resulta $\chi^2 = 9,2376$ com 3 graus de liberdade, portanto não significante ($\alpha = 0,05$), isto é, o modelo se ajusta aos dados.

Assim, podemos testar hipóteses do tipo

$$H0: \underset{dxv}{C}\ \xi = 0,$$

utilizando

$$SS\ \{C\xi = 0\} = \underset{\sim}{b}'C'\{C(X'S^{-1}X)^{-1}C'\}^{-1}\ C\ \underset{\sim}{b}$$

que possui distribuição χ^2 com d graus de liberdade.

Se construirmos a matriz

$$\underset{2x9}{C_1} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

a hipótese $H0: C_1\xi = \begin{pmatrix} \beta_1 - \beta_3 \\ \beta_2 - \beta_3 \end{pmatrix} = 0$ significa hipótese de homogeneidade nos efeitos do fator "tipo de acidente" em relação aos graus de ferimento. O resultado $\chi^2 = 365,69$ com 2 graus de liberdade é significante ($\alpha = 0,05$), isto é, não há homogeneidade nos efeitos do fator "tipo de acidente" nas respostas ordenadas.

Desse modo, a hipótese para a verificação do efeito do fator "tipo de acidente" na gravidade do ferimento, é formulada por

$$H0: C_2\xi = 0 \text{ , com } C_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix},$$

produzindo $\chi^2 = 888,6329$ com 1 grau de liberdade. Rejeitamos essa hipótese, ao nível de 5%, portanto, existe efeito do fator "tipo de acidente" na gravidade do ferimento.

Do mesmo modo, a hipótese $H0: C_3\xi = 0$, com

$$\underset{2x9}{C_3} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix}$$

significa hipótese de homogeneidade nos efeitos do fator "grupo motorista" em relação aos graus de ferimento. O resultado $\chi^2 = 8,4608$ com 2 graus de liberdade é significante

(α = 0,05), isto é, não há homogeneidade nos efeitos do fator "grupo motorista" na gravidade do ferimento.

A hipótese para verificação do efeito do fator "grupo motorista" na gravidade do ferimento é formulada por

$$H0: C_4 \xi = 0 \quad , \text{ com}$$

$$C_4 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

resultando χ^2 = 44,0965 com 1 grau de liberdade, portanto significante (α = 0,05), o que significa que existe efeito do fator "grupo motorista" na gravidade do ferimento.

Em suma, a análise através da formulação dos logits mostrou ser adequada para os dados da Tabela 4, devido ao teste do ajuste do modelo resultar não significante, o que não aconteceu quando formulamos escores médios utilizando escores arbitrários para cada resposta ordenada. Ainda, a análise indicou a existência de efeito dos fatores "grupo motorista" e "tipo de acidente" na gravidade do ferimento.

CAPÍTULO V

EXTENSÃO DE FORTHOFER E KOCH

5.1. *INTRODUÇÃO.*

Vimos nos capítulos anteriores que o método introduzido por GSK (1969) restringe a análise de dados categóricos àqueles problemas que podem ser relatados como funções lineares ou log-lineares das probabilidades das caselas, isto é, funções que pertencem ao tipo $F(\underset{\sim}{\pi}) = A\underset{\sim}{\pi}$,ou $F(\underset{\sim}{\pi}) = K \ln A\underset{\sim}{\pi}$.

FORTHOFER e KOCH, em 1973, [10] apresentaram uma extensão da metodologia de GSK para análise de dados categóricos considerando funções mais complexas, obtidas por transformações compostas lineares, logarítmicas e exponenciais. Assim, muitos problemas que não podiam ser formulados por funções propostas por GSK podem ser relatados por funções compostas e analisados por técnicas derivadas de modelos lineares.

5.2. FUNÇÕES COMPOSTAS.

As funções apresentadas por FORTHOFER e KOCH pertencem a duas classes:

A) A primeira classe de funções é constituída por funções do tipo

$$F(\underset{\sim}{\pi}) = Q\{\exp\left[K \ln (A\underset{\sim}{\pi})\right]\}$$

onde A é uma matriz de constantes $u x r s$; K uma matriz $k x u$ e Q uma matriz $q x k$; ln transforma um vetor no vetor correspondente de logaritmos naturais, e exp transforma um vetor no vetor correspondente de funções exponenciais (isto é, antilog).

A matriz de derivadas parciais da transformação $F(\underset{\sim}{p}) = Q \exp K \ln A \underset{\sim}{p}$ com

$$y = \exp K \ln A \underset{\sim}{p} \quad \text{e} \quad a = A \underset{\sim}{p}$$

é dada por

$$H = \left[\frac{\delta F_m(\pi)}{\delta \pi_{ij}} / \pi_{ij} = p_{ij}\right] = Q\, D_y D_a^{-1} A$$

onde D_y é uma matriz $k x k$ com elementos do vetor y na diagonal principal, isto é,

$$D_y = \begin{pmatrix} y_1 & 0 & \dots & 0 \\ 0 & y_2 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & y_k \end{pmatrix}$$

e, D_a é também uma matriz diagonal principal uxu , com elementos do vetor a na diagonal principla, isto é,

$$D_a = \begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & a_u \end{pmatrix}$$

Assim, a matriz de covariância estimada de F será dada por

$$S = Q\, D_y\, K\, D_a^{-1}\, A\, V(\underset{\sim}{p})\, A'D_a^{-1}\, K'D_y'\, Q'$$

B) A segunda classe de funções é constituída por funções do tipo

$$F(\underset{\sim}{\pi}) = L \ln Q \exp K \ln A\underset{\sim}{\pi}$$

onde as matrizes A, K e Q são iguais às descritas na classe anterior, e L é uma matriz de constantes lxq.

A matriz das derivadas parciais da transformação $F(\underset{\sim}{p}) = L \ln Q \exp K \ln A\, \underset{\sim}{p}$ com

$g = \ln Q \exp K \ln A \underset{\sim}{p}$, $y = \exp K \ln A \underset{\sim}{p}$ e $a = A \underset{\sim}{p}$

é dada por

$$H = \left(\frac{\delta F_m(\pi)}{\delta \pi_{ij}} \Big/ \pi_{ij} = p_{ij} \right) = L\, D_g^{-1} Q\, D_y\, K\, D_a^{-1}\, A$$

onde D_g é uma matriz diagonal principal $q x q$, com elementos de g na diagonal principal, isto é,

$$D_g = \begin{pmatrix} g_1 & 0 & \dots & 0 \\ 0 & g_2 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & g_q \end{pmatrix}$$

Portanto, a matriz de covariância estimada de F será dada por

$$S = L\, D_g^{-1}\, Q\, D_y\, K\, D_a^{-1}\, A\, V(\underset{\sim}{p}) A' D_a^{-1}\, K' D_y\, Q' D_g^{-1}\, L' .$$

5.3. *APLICAÇÕES*.

FORTHOFER e KOCH apresentam diversas aplicações das duas classes de funções acima descritas; vamos tratar aqui somente de dois tipos de problemas enquadrados nessas classes de funções:

a) análise dos escores médios baseados em escores percentis, como os propostos por WILLIAMS e GRIZZLE em 1972;

b) análise dos coeficientes de correlação de postos, como os propostos por GOODMAN e KRUSKAL em 1954 e 1963.

5.3.1. *HIPÓTESE LINEAR GERAL NO MODELO $F(\pi) = X\beta$.*

A) Caso $F(\underset{\sim}{\pi}) = Q \exp K \ln A\underset{\sim}{\pi}$

Dentre as funções pertencentes a esta classe, vamos desenvolver apenas os escores médios baseados em escores percentis, propostos por WILLIAMS e GRIZZLE em 1972. Os escores percentis são calculados usando as distribuições marginais observadas, isto é, são baseados em postos; são adequadas nas situações em que a(s) resposta(s) se apresenta(m) ordenada(s).

Para ilustrar a formulação desses escores médios, vamos considerar o modelo "dois fatores, uma resposta", com A e B sendo os fatores com c e d níveis respectivamente, e C a resposta com r categorias ordenadas.

Vamos denotar os escores percentis por a_{ijk}, ca

da um deles associado a cada p_{ijk}, e escrever para cada multinomial (i,j) o vetor dos escores percentis como

$$\bar{a}_{ij} = \left(a_{ij1}, a_{ij2}, a_{ij3}, \ldots, a_{ijr}\right)$$

Se os níveis do fator A não representam blocos, calculamos os escores percentis para os dados combinados da tabela toda; assim, os a_{ij} são idênticos para todas as combinações (i,j). Se os níveis de A representam blocos não homogêneos, calculam-se os escores dentro de cada nível de A; desta forma, os $\bar{a}_{ij}$ são idênticos dentro de cada nível do fator A, e então, o vetor dos escores percentis para cada i será denotado por $\bar{a}_i = \left(a_{i1}, a_{i2}, \ldots, a_{ir}\right)$.

Considerando os níveis de A como blocos, o cálculo dos escores percentis correspondentes às r respostas ordenadas, é feito utilizando-se:

$$a_{i1} = p_{i.1} / 2$$

$$a_{i2} = p_{i.1} + p_{i.2} / 2$$

$$\ldots\ldots\ldots\ldots$$

$$a_{ir} = p_{i.1} + p_{i.2} + \ldots + p_{i.r} / 2$$

com $p_{i.k} = n_{i.k} / n_{i..}$ sendo a probabilidade marginal observada da k-ésima resposta C para o i-ésimo nível de A.

Assim, se considerarmos um i-ésimo nível de A, de uma tabela "dois fatores, uma resposta", com as frequências nas caselas

		C_1	C_2	...	C_r	Total
A_i	B_1	n_{i11}	n_{i12}	...	n_{i1r}	$n_{i1.}$
	B_2	n_{i21}	n_{i22}	...	n_{i2r}	$n_{i2.}$
	. .	. . . .	. . . .	. . . .	. . . .	. . .
	B_d	n_{id1}	n_{id2}	...	n_{idr}	$n_{id.}$
Total		$n_{i.1}$	$n_{i.2}$	...	$n_{i.r}$	$n_{i..}$
Escores percentis		a_{i1}	a_{i2}	...	a_{ir}	

os escores percentis escritos em função das probabilidades das caselas resultam:

$$a_{i1} = \frac{n_{i1.}}{2n_{i..}} p_{i11} + \frac{n_{i2.}}{2n_{i..}} p_{i21} + \ldots + \frac{n_{id.}}{2n_{i..}} p_{id1}$$

$$a_{i2} = \frac{n_{i1.}}{n_{i..}} p_{i11} + \frac{n_{i1.}}{2n_{i..}} p_{i12} + \frac{n_{i2.}}{n_{i..}} p_{i21} + \frac{n_{i2.}}{2n_{i..}} p_{i22} +$$

$$+ \ldots + \frac{n_{id.}}{n_{i..}} p_{id1} + \frac{n_{id.}}{2n_{i..}} p_{id2}$$

: :

$$a_{ir} = \frac{n_{i1.}}{n_{i..}} p_{i11} + \frac{n_{i1.}}{n_{i..}} p_{i12} + \ldots + \frac{n_{i1.}}{2n_{i..}} p_{i1r} +$$

$$+ \frac{n_{i2.}}{n_{i..}} p_{i21} + \frac{n_{i2.}}{n_{i..}} p_{i22} + \dots + \frac{n_{i2.}}{2n_{i..}} p_{i2r} + \dots +$$

$$+ \frac{n_{id.}}{n_{i..}} p_{id1} + \frac{n_{id.}}{n_{i..}} p_{id2} + \dots + \frac{n_{id.}}{2n_{i..}} p_{idr} \quad .$$

Assim, os escores médios baseados em percentis são cxd funções do tipo

$$F(\underset{\sim}{p}) = \begin{cases} a_{11}\, p_{111} + a_{12}\, p_{112} + \dots + a_{1r}\, p_{11r} \\ a_{11}\, p_{121} + a_{12}\, p_{122} + \dots + a_{1r}\, p_{12r} \\ \dots\dots\dots\dots\dots\dots \\ a_{11}\, p_{1d1} + a_{12}\, p_{1d2} + \dots + a_{1r}\, p_{1dr} \\ a_{21}\, p_{211} + a_{22}\, p_{212} + \dots + a_{2r}\, p_{2dr} \\ \vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots \\ a_{21}\, p_{2d1} + a_{22}\, p_{2d2} + \dots + a_{2r}\, p_{2dr} \\ \dots\dots\dots\dots\dots\dots \\ a_{c1}\, p_{cd1} + a_{c2}\, p_{cd2} + \dots + a_{cr}\, p_{cdr} \end{cases}$$

e que serão objeto da análise usando um modelo linear apropriado.

Vamos ilustrar a formulação e a análise dos escores médios baseados em escores percentis utilizando os dados apresentados na Tabela 4, já analisados no Capítulo IV.

Para a obtenção dos escores médios, realizamos

uma sequência de transformações do vetor das probabilidades das caselas, com a construção de matrizes Q, K e A, e utilizando transformações linear, logarítmica e exponencial.

O vetor $a = A\,\underset{\sim}{p}$, com

$$\underset{24x16}{A} = \begin{pmatrix} I_{16} \\ A_1 & \\ & A_2 \end{pmatrix}$$

onde I_{16} é uma matriz identidade 16x16,

$$\underset{4x8}{A_1} = \begin{pmatrix} 0,10 & 0 & 0 & 0 & 0,40 & 0 & 0 & 0 \\ 0,20 & 0,10 & 0 & 0 & 0,80 & 0,40 & 0 & 0 \\ 0,20 & 0,20 & 0,10 & 0 & 0,80 & 0,80 & 0,40 & 0 \\ 0,20 & 0,20 & 0,20 & 0,10 & 0,80 & 0,80 & 0,80 & 0,40 \end{pmatrix}$$

$$\underset{4x8}{A_2} = \begin{pmatrix} 0,13 & 0 & 0 & 0 & 0,36 & 0 & 0 & 0 \\ 0,26 & 0,13 & 0 & 0 & 0,73 & 0,36 & 0 & 0 \\ 0,26 & 0,26 & 0,13 & 0 & 0,73 & 0,73 & 0,36 & 0 \\ 0,26 & 0,26 & 0,26 & 0,13 & 0,73 & 0,73 & 0,73 & 0,36 \end{pmatrix}$$

é o vetor das probabilidades observadas das caselas e dos escores percentis para os dois níveis do fator "grupo motorista".

Os escores percentis resultantes são:

$$a_1' = (0,03881 \quad 0,30729 \quad 0,69437 \quad 0,92442)$$

$$a_2' = (0,02836 \quad 0,26572 \quad 0,65249 \quad 0,90874)$$

O vetor $f = K \ln A \underset{\sim}{p}$, com

$$\underset{16x24}{K} = \begin{pmatrix} I_4 & 0 & 0 & 0 & I_4 & 0 \\ 0 & I_4 & 0 & 0 & I_4 & 0 \\ 0 & 0 & I_4 & 0 & 0 & I_4 \\ 0 & 0 & 0 & I_4 & 0 & I_4 \end{pmatrix}$$

onde I_4 é uma matriz identidade 4x4 ,

0 é uma matriz 4x4 de zeros ,

é o vetor dos log dos produtos de cada p_{ijk} e seu respectivo escore percentil.

Finalmente, o vetor $g = Q \exp K \ln A \underset{\sim}{p}$, com

$$\underset{4x16}{Q} = \begin{pmatrix} 1\,1\,1\,1 & 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 1\,1\,1\,1 & 0\,0\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,1\,1\,1 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,1\,1\,1 \end{pmatrix}$$

é o vetor dos escores médios, que resulta:

$$F^{-}(\underset{\sim}{p}) = \begin{pmatrix} 0{,}66149 & 0{,}45817 & 0{,}65172 & 0{,}43685 \end{pmatrix}$$

com matriz de covariância dada por

$$S = 10^{-4} \begin{pmatrix} 0{,}15693 & -0{,}03860 & 0 & 0 \\ -0{,}03860 & 0{,}00949 & 0 & 0 \\ 0 & 0 & 0{,}11560 & -0{,}04128 \\ 0 & 0 & -0{,}04128 & 0{,}01474 \end{pmatrix}$$

A fim de verificar a variação dos escores médios, é ajustado o modelo linear $F(\underset{\sim}{\pi}) = X\xi$, com

$$\underset{4x3}{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \qquad \text{e}$$

$$\xi = \{\mu, \alpha_1, \beta_1\}$$

com: μ a média constante;

α_i o efeito do i-ésimo nível de "grupo motorista" na resposta C;

β_j o efeito do j-ésimo nível de "tipo de acidente" na resposta C, e

$$\sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{2} \beta_j = 0 \ .$$

Os estimadores de ξ e a matriz de covariância resultam:

$$\underset{\sim}{b} = \begin{pmatrix} 0,55235 \\ 0,00935 \\ 0,10474 \end{pmatrix}$$

$$V_{\underset{\sim}{b}} = 10^{-6} \begin{pmatrix} 0,82665 & 0,10064 & 1,53230 \\ 0,10064 & 0,01230 & 0,18660 \\ 1,53230 & 0,18660 & 0,28400 \end{pmatrix}$$

O teste do ajuste do modelo resultou $\chi^2 = 2,9259$ com 1 grau de liberdade, portanto não significante , isto é, o modelo se ajusta aos dados.

As hipóteses do tipo $H0: C\xi = 0$, com

$$C_1 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \quad e \quad C_2 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$$

testam a não existência dos efeitos "grupo motorista" e "tipo de acidente" respectivamente, na gravidade do ferimento produzido. Os testes resultam $\chi^2_1 = 7100,4802$ e $\chi^2_2 = 3862,5853$, ambos com um grau de liberdade, portanto significantes. Concluímos que existe efeito "grupo motorista" e "tipo de acidente" na "gravidade do ferimento", o que concorda com os resultados obtidos através da análise dos logits.

B) Caso $F(\underset{\sim}{\pi}) = L \ln Q \exp K \ln A\underset{\sim}{\pi}$

Para estudar a formulação de funções pertencentes a essa classe, vamos considerar a análise dos coeficientes de correlação de postos, descritos por GOODMAN e KRUSKAL (1954, 1963). Essa análise é adequada em situações experimentais do tipo "um fator, duas respostas", quando estamos interessados na associação entre as respostas, que po

de ser medida através do coeficiente de correlação de postos.

Os coeficientes de correlação de postos, calculados para cada nível do fator A, são baseados nas probabilidades de concordância π_s, isto é, nas probabilidades que dois indivíduos, tomados aleatoriamente, tenham a mesma ordem em ambas as classificações de respostas, e de discordância π_d, isto é, a probabilidade que eles tenham ordens diferentes de classificação, ou seja:

$$\pi_s = 2\sum_j \sum_k \pi_{ijk} \{ \sum_{j'>j} \sum_{k'>k} \pi_{ij'k'} \}$$

$$\pi_d = 2\sum_j \sum_k \pi_{ijk} \{ \sum_{j'>j} \sum_{k'<k} \pi_{ij'k'} \}$$

Os coeficientes de correlação, para cada nível do fator A, são dados por:

$$\gamma = \frac{\pi_s - \pi_d}{\pi_s + \pi_d}$$

O objetivo é construir funções que são estimadores das razões de concordância e discordância para cada nível do fator A, funções essas que pertencem ao tipo

$$F(\underset{\sim}{p}) = L \ln Q \exp K \ln A \underset{\sim}{p} \quad .$$

A análise desses estimadores das razões das probabilidades de concordância e discordância é realizada ajustando-se modelo linear $F(\underset{\sim}{\pi}) = X\beta$. Além disso, objetivamos estimar os coeficientes de correlação para cada nível do fator A.

Iremos ilustrar esta análise com os dados da Tabela 4, rearranjados da seguinte forma:

Tabela 15 - Número de motoristas com ferimentos, agrupados segundo "grupo" e "tipo de acidente".

Tipo de acidente	Grau de ferimento	Grupo motorista		Total
		Só	acompanhado	
Capotamento	leve	21	18	39
	moderado	567	553	1120
	severo	1356	1734	3090
	crítico	644	869	1513
Não-capotamento	leve	996	679	1675
	moderado	5454	4561	10015
	severo	2773	2516	5289
	crítico	1256	1092	2348

Os dados são vistos como resultado da classificação de dois grupos motoristas, "só" e "acompanhado" (i = 1,2), em tipo de acidente, "capotamento" e "não-capotamento"

(j = 1,2), e gravidade do ferimento, "leve", "moderado", "severo" e "crítico" (k = 1,2,3,4). Assim, essa tabela pode ser vista como do tipo "um fator, duas respostas", ou, na notação do modelo geral, s = 2 e r = 8.

As probabilidades de concordância e discordância para cada grupo motorista são formuladas por:

$$\pi_{s_1} = 2\pi_{111}\ (\pi_{122}+\pi_{123}+\pi_{124}) + \pi_{112}\ (\pi_{123}+\pi_{124}) + \pi_{113}\ \pi_{124}$$

$$\pi_{d_1} = 2\pi_{112}\ \pi_{121} + \pi_{113}\ (\pi_{121}+\pi_{122}) + \pi_{114}\ (\pi_{121}+\pi_{122}+\pi_{123})$$

$$\pi_{s_2} = 2\pi_{211}\ (\pi_{222}+\pi_{223}+\pi_{224}) + \pi_{212}\ (\pi_{223}+\pi_{224}) + \pi_{213}\ \pi_{224}$$

$$\pi_{d_2} = 2\pi_{212}\ \pi_{221} + \pi_{213}\ (\pi_{221}+\pi_{222}) + \pi_{214}\ (\pi_{221}+\pi_{222}+\pi_{223})$$

Os log dos estimadores das razões das probabilidades de concordância e discordância pertencem à classe de funções

$$F(\underset{\sim}{p}) = L \ln Q \exp K \ln A\ \underset{\sim}{p}$$

isto é, obtidos através de transformações do vetor $\underset{\sim}{p}$ e utilizando as matrizes L, Q, K e A adequadas.

Construímos as matrizes:

$$\underset{28\times16}{A} = \begin{pmatrix} I_8 & 0 \\ A_1 & 0 \\ 0 & I_8 \\ 0 & A_1 \end{pmatrix}$$

com I_8 uma matriz identidade 8x8;

0 uma matriz 8x8 de zeros; e

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

de tal modo que $a = A \underset{\sim}{p}$ nos fornece os estimadores das probabilidades das caselas dos grupos "só" e "acompanhado", e ainda a soma das apropriadas combinações concordantes e discordantes para cada grupo motorista;

$$\underset{12x28}{K} = \begin{pmatrix} K_1 & I_6 & 0 \\ 0 & K_1 & I_6 \end{pmatrix}$$

com I_6 uma matriz identidade 6x6;

0 uma matriz 6x14 de zeros; e

$$K_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

de tal modo que $f = K \ln A \underset{\sim}{p}$ nos fornece o vetor dos log das probabilidades de vários tipos de pares concordantes e

discordantes;

$$\underset{4x12}{Q} = \begin{pmatrix} Q_1 & 0 \\ 0 & Q_1 \end{pmatrix} \quad , \text{ com } \quad Q_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

de modo que $g = Q \exp K \ln A \underset{\sim}{p}$ nos fornece os estimadores das probabilidades concordantes e discordantes para cada grupo motorista, resultando

$$g = Q \exp K \ln A \underset{\sim}{p} = \begin{pmatrix} 0,02452 \\ 0,08922 \\ 0,02792 \\ 0,11210 \end{pmatrix}$$

e, finalmente,

$$\underset{2x4}{L} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

de modo que $F(\underset{\sim}{p}) = L \ln Q \exp K \ln A \underset{\sim}{p}$ nos dá o log das probabilidades concordantes e discordantes para cada grupo motorista, isto é,

$$F(\underset{\sim}{p}) = \log \begin{pmatrix} \dfrac{p_{111}(p_{122}+p_{123}+p_{124})+p_{112}(p_{123}+p_{124})+p_{113}p_{124}}{p_{112}p_{121}+p_{113}(p_{121}+p_{122})+p_{114}(p_{121}+p_{122}+p_{123})} \\ \dfrac{p_{211}(p_{222}+p_{223}+p_{224})+p_{212}(p_{223}+p_{224})+p_{213}p_{224}}{p_{212}p_{221}+p_{213}(p_{221}+p_{222})+p_{214}(p_{221}+p_{222}+p_{233})} \end{pmatrix}$$

$$F(\underset{\sim}{p}) = \begin{pmatrix} -1,2927 \\ -1,3899 \end{pmatrix}$$

com matriz de covariância dada por

$$S = 10^{-2} \begin{pmatrix} 0,12325 & 0 \\ 0 & 0,11919 \end{pmatrix}$$

É ajustado o modelo linear $E\{F\} = X\beta_o$, com

$X = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ e β_o sendo uma medida de associação entre tipo de acidente e grau de ferimento.

O estimador de β_o , b , e a sua matriz de covariância resultam: $b = -1,3421$ e $V_b = 0,60594 \times 10^{-3}$.

O teste do ajuste do modelo, que nos dá o teste da não-interação de segunda ordem, entre grupo motorista e tipo de acidente em relação aos seus efeitos na gravidade do ferimento, resulta $\chi^2 = 3,8977$ com 1 grau de liberdade, portanto não significante ($\alpha = 0,025$), isto é, não existe interação entre "grupo motorista", "tipo de acidente" e "gravidade do ferimento".

A hipótese $H0: C\beta_o = 0$, com $C = (1)$ nos fornece o teste da não-interação entre tipo de acidente e grau de ferimento, e resulta $\chi^2 = 2972,6841$ com 1 grau de liberdade, portanto significante, isto é, existe associação

entre gravidade do ferimento e tipo de acidente.

Com o objetivo de estimar os coeficientes de correlação para cada grupo motorista, usando os resultados obtidos em $F(\underset{\sim}{p})$, utilizamos:

$$G = (D_g + I_2)^{-1} (g - J_2)$$

$$V_G = 2(D_g + I_2)^{-2} D_g S D_g (D_g + I_2)^{-2}$$

onde $g = \exp F$, D_g é uma matriz bloco diagonal com elementos de g na diagonal principal, J_2 é um vetor 2x1 e I_2 é uma matriz identidade 2x2, e obtemos:

$$G = \begin{pmatrix} \dfrac{e^{F_1} - 1}{e^{F_1} + 1} \\ \\ \dfrac{e^{F_2} - 1}{e^{F_2} + 1} \end{pmatrix} = \begin{pmatrix} -0,57 \\ -0,60 \end{pmatrix}$$

Assim, $\gamma_1 = -0,57$ é o coeficiente de correlação entre "tipo de acidente" e "grau de ferimento", para o grupo motorista "só"; $\gamma_2 = -0,60$ é o coeficiente de correlação entre "tipo de acidente" e "grau de ferimento" para o grupo motorista "acompanhado". Esses coeficientes possuem as mesmas magnitudes para os dois grupos e indicam, nesses dados, que capotamento provoca acidente com ferimentos mais graves do que o não-capotamento.

CAPÍTULO VI

CONSIDERAÇÕES FINAIS

Pretendemos, aqui, fazer breves comentários em relação à parte formal e à parte das aplicações do método de GSK (1969) e da extensão de FORTHOFER e KOCH (1972).

Já foi dito neste trabalho que a metodologia de GSK para análise de dados categóricos apresenta uma vantagem em relação aos métodos usuais, que é a de não necessitar de estimadores de máxima verossimilhança dos parâmetros sob H0. Lembramos, mais uma vez, que algumas hipóteses formuladas em tabelas de contingências, como, por exemplo, as de homogeneidade das marginais (GSK, 1969), de associações sobre as diagonais totais (FORTHOFER e KOCH, 1972), de simetria e de associação sobre a diagonal principal (desenvolvidas por nós), apresentam dificuldadesna obtenção de estimadores sob H0, pelos métodos usuais, mas que são facilmente resolvidas pelo método aqui estudado.

Ainda, quando os dados são provenientes de diversas populações, o método de GSK e a extensão de FORTHOFER e

KOCH permitem a solução de diversos problemas através da formulação e análise de funções adequadas. Assim, o método proposto permite a análise, por exemplo, de: escores médios, logits, taxas de sobrevivência, coeficientes de correlação de postos, ridits, associações parciais, estas últimas propostas por MANTEL e HAENZEL (1959) e MANTEL (1963). Cada um desses problemas é formulado por estimadores compostos envolvendo funções linear, logarítmica e exponencial; essas funções devem obedecer às seguintes condições: ser linearmente independentes e possuir derivadas parciais contínuas de segunda ordem em relação a π_{ij} ; além disso, se algumas das caselas possuir frequência zero, esta deve ser substituída por $1/r_i n_i$, sendo r o número de categorias de respostas. Obedecidas essas condições, essas funções são escritas equivalentemente por uma sequência de matrizes A, K, Q e L apropriadas, utilizando-se transformações do vetor de probabilidades.

O programa utilizado para resolver os exemplos aqui ilustrados foi o GENCAT, pertencente ao acervo do IME.

Apesar de muitas dificuldades na aplicação do método de GSK apontadas por pesquisadores como, por exemplo, GOODMAN, no seu trabalho "Analysis of multidimensional Con-

tingency Tables: stepwise procedures and direct estimation for building models of multiple classifications"; in: TECHNOMETRICS, 1971, p.33-61, para hipóteses por ele tratadas, e em que o referido autor assinala dificuldades computacionais em modelos log-lineares, podemos, de um modo geral, concluir pela fácil aplicação do método.

Sabemos, no entanto, que o método de GSK não é o único para solucionar os problemas surgidos em dados categóricos; lembraríamos, a respeito, a existência de dois métodos alternativos: um formulado por BISHOP (1969,1971) e GOODMAN (1970,1971); o outro, por KU et alii (1971).

Outro ponto de observação, que também encontramos nos próprios autores do método, é o do desconhecimento do comportamento estatístico do método em amostras pequenas; apesar disso, reforçam eles o fato de que os testes estatísticos nele empregados apresentam ótimas propriedades assintóticas, como demonstrou NEYMAN. Ainda considerando o problema de estimadores, RAO (1963) demonstrou que os estimadores de máxima verossimilhança apresentam menor variância que os estimadores BAN utilizados no método de GSK; em contrapartida, BERKSON (1955) mostrou, em trabalhos práticos, uma equivalência dos estimadores de máxima verossi-

milhança e χ_1^2 mínimo, forma equivalente da estatística de WALD utilizada no método de GSK.

O trabalho que desenvolvemos procurou enfatizar as aplicações do método de GSK. Restringimos nossa área de aplicação a tabelas de classificação completas, lembrando que existem estudos sobre aplicação do método a tabelas incompletas, no sentido que algumas caselas têm probabilidade de ocorrência nula (GRIZZLE e WILLIAMS, 1972), e também a vetores-resposta incompletos, no sentido que a classificação não abrange todas as categorias de resposta, devido à perda de dados, respostas incorretas ou incompletas (KOCH, IMREY, REINFURT, 1972).

Esperamos, com este trabalho, poder contribuir, ainda que de forma modesta, para a solução de vários problemas que envolvem dados categóricos nas diversas áreas de pesquisa.

BIBLIOGRAFIA

1. Berkson, J. "Maximum likelihood and minimum χ^2 estimate of the logist function", in: JASA, 50, 1955, 130-162.

2. Berkson, J. "Application of minimum logit χ^2 estimate to a problem of Grizzle, with a notation on the problem of no interaction", in: BIOMETRICS, 24, 1968, 75-95.

3. Bhapkar, V.P. "A note on the equivalence of two test criteria for hypotheses in categorical data", in: JASA, 61, 1966, 228-235.

4. Bhapkar, V.P. "Some tests for categorical data", in: ANN. MATH.STAT., 32, 1961, 72-83.

5. Bhapkar, V.P. e Kock, G.G. "Hypotheses of 'no interaction' in multidimensional contingency tables", in: TECHNOMETRICS, 10, 1968a, 107-123.

6. Bhapkar, V.P. e Koch, G.G. "On the hypotheses of 'no interaction' in contingency tables", in: BIOMETRICS, 24, 1968b, 567-594.

7. Bishop, Y.M.M. "Effects of collapsing multidimensional contingency tables", in: BIOMETRICS, 27, 1971, 545-62.

8. Bross, I.D.J. "How to use 'ridit' analysis", in: BIOMETRICS, 14, 1958, 18-38.

9. Darroch, J.N. "Interactions in multi-factor contingency tables"; in: J.ROYAL STAT.SOC., Series B, 24, 1962, 251-263.

10. Forthofer, R.N. e Koch, G.G. "An analysis for compounded functions of categorical data", in: BIOMETRICS, 29, 1973, 143-157.

11. Grizzle,J.E., Starmer,C.F. e Koch, G.G. "Analysis of categorical data by linear models", in: BIOMETRICS, 25,1969,489-504.

12. Grizzle,J.E. e Williams,O.D. "Log linear models and tests of independence for contingency tables", in: BIO_ METRICS,28,1972,137-156.

13. Goodman,L.A. e Kruskal,W.H. "Measures of association for cross classification",in: JASA,49,1954,734-764.

14. Goodman,L.A. e Kruskal,W.H. "Measures of association for cross classification II; further discussion and references", in: JASA,59,1959,123-163.

15. Goodman,L.A. e Kruskal,W.H. "Measures of association for cross classification III; approximate sampling theory", in: JASA,58,1963,310-364.

16. Ireland,C.T., Ku,H.H. e Kullback,S. "Simmetry and marginal homogeneity of an rxr contingency tables",in: JASA,64,1969,1324-1341.

17. Johnson,N.L. e Leone,I.C. - STATISTICS AND EXPERIMENTAL DESIGN IN ENGINEERING AND THE PHYSICAL SCIENCES, 2 vol., New York, John Wiley & Sons, 1964.

18. Koch,G.G. e Reinfurt, D.W. "The analysis of categorical data from mixed models", in: BIOMETRICS,27,1971, 157-174.

19. Koch,G.G., Imrey,P.B. e Reinfurt,D.W. "Linear models analysis of categorical data with incomplete response vectors", in: BIOMETRICS,1972,663-690.

20. Koch,G.G.,Johnson,W.D. e Tolley,H.D. "A linear models approach to analysis of survivaland extent of disease in multidimensional contingency tables",in:JASA, 67,1972,780-795.

21. Mantel,N. "Chi-square tests with one degree of freedom; extensions of the Mantel-Haenzel procedure", in: JASA,58,1963,690-700.

22. Mantel,N. e Haenzel,W. "Statistical aspects of the analysis of data from retrospective studies of disease",in: J.NAT.CANCER INST.,22,1959,719-748.

23. Rao,C.R.- LINEAR STATISTICAL INFERENCE AND ITS APPLICATIONS; 2a. ed.,New York, John Wiley & Sons,1973.

24. Roy,S.N. e Kastenbaum "On the hypotheses of no interaction in a multiway contingency tables", in: ANN. MATH.STAT.,27,1956,749-757.

25. Searle,S.R.- LINEAR MODELS. New York, John Wiley & Sons, 1971.

26. Wald,A. "Tests of statistical hypotheses concerning several parameters when the number of observations is large", in: TRANS.AMER.MATH.SOC.,1943,54, 426-482.

27. Williams,O.D. e Grizzle,J.E. "Analysis of contingency tables having ordered response categories", in: JASA,67,1972,55-63.

28. Neyman,J. "Contributions to the theory of the χ^2 test", in: PROC.BERKELEY SYMP.MATH.STATISTICS AND PROBABILITY, Univ.California Press, Berkeley-Los Angeles, 1949.