ANÁLISE DE DADOS CATEGORIZADOS: MODELOS LOG-LINEARES E INDEPENDÊNCIA EM TABELAS DE CONTINGÊNCIA

PAULO JOSÉ OGLIARI

Orientador: Prof. Dr. CÁSSIO ROBERTO DE MELO GODOI

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", da Universidade de São Paulo, para obtenção do título de Mestre em Agronomia. Área de concentração: "Estatística e Experimentação Agronômica".

PIRACICABA
Estado de São Paulo - Brasil
Junho - 1984

À

Osório Paulo Ogliari (in memorian)

Inês Teston Ogliari

DEDICO

À

Maria de Lourdes Ogliari

Juliana Bernardi

Tânia Inês Ogliari

OFEREÇO

A G R A D E C I M E N T O S

Ao Prof. Dr. Cássio Roberto de Melo Godoi, pela orientação, sugestões, esclarecimentos, apoio, incentivo e amizade, os quais concorreram para a realização deste trabalho.

Ao Colega José Carlos Barbosa pelo valioso auxílio relativo à parte de computação e esclarecimentos prestados.

Aos Professores do Departamento de Matemática e Estatística da ESALQ/USP, pelos ensinamentos e experiências transmitidos.

À Empresa Catarinense de Pesquisa Agropecuária S/A, pelo incentivo e valioso auxílio financeiro concedido para a realização deste curso de pós-graduação.

Aos Colegas de curso pela amizade e solidariedade, os quais propiciaram um ambiente bastante produtivo.

Aos Funcionários do Departamento de Matemática e Estatística da ESALQ/USP, Djanira Ortolan Forti, Maria Izalina Ferreira Alves, Maria Alayde Penteado de Souza, Maria Aparecida Bueno de Camargo, Nilce Maria Camargo e Octávio Frassetto pela amizade e serviços prestados.

À Rosa Maria Alves pela amizade e pela realização dos serviços de datilografia.

ÍNDICE

ANÁLISE DE DADOS CATEGORIZADOS: MODELOS LOG-LINEARES E INDEPENDÊNCIA EM TABELAS DE CONTINGÊNCIA

Autor: PAULO JOSÉ OGLIARI

Orientador: Cássio Roberto de Melo Godoi

RESUMO

O objetivo deste trabalho é o de fazer-se um estudo exploratório sobre os problemas causados pela aproximação assintótica dos testes utilizados em tabelas de contingência, objetivando determinar as condições práticas pelas quais se obtenham testes de hipóteses e estimativas confiáveis dos parâmetros do modelo multinomial. Para esse estudo foram analisadas tabelas de contingência do tipo CxDxF obtidas por meio de simulação de dados de populações com tamanhos variados.

São também objetivos deste trabalho a elaboração de um programa de análise em linguagem BASIC para o ajustamento de modelos log-lineares aos dados de tabelas de contingência do tipo CxDxF e a determinação de estruturas de independência para os dados dessas tabelas.

A metodologia utilizada é a de GRIZZLE, STARMER e KOCH (1969) que descrevem como modelos lineares e quadrados mínimos ponderados

podem ser utilizados na análise de dados dispostos em uma tabela de contingencia.

Pelos resultados obtidos observa-se que com tamanhos de amostras pequenos (N = 200, N = 250 e N = 300), os testes estatísticos mostram-se rigorosos em não rejeitar a estrutura de independência estabelecida para as tabelas de contingência; porém, os modelos ajustados aos dados nem sempre correspondem exatamente aos que se estão testando.

Para tamanhos de amostras maiores (N = 1000 e N = 1500) , os testes estatísticos comportam-se de modo excelente e os modelos que explicam as relações entre as diversas variáveis em estudo ficam bem ajustados.

Conclui-se também que o procedimento de BERKSON (1955) para tabelas com frequências nulas produz bons resultados, assim como, de um modo geral, o método GSK.

Ao contrário das considerações feitas por GOODMAN (1971b) o método GSK mostra-se de um modo geral de fácil aplicação.

CATEGORICAL DATA ANALYSIS: LOG LINEAR MODELS AND INDEPENDENCE IN CONTINGENCY TABLES

Author: PAULO JOSÉ OGLIARI

Adviser: CÁSSIO ROBERTO DE MELO GODOI

SUMMARY

The first objective of this work is to explore the problems caused by the asymptotic approximation of χ^2 tests commonly used in contingency tables and to obtain pratical rules under which good properties of the tests can be reached and confidence on the parameters estimation of the multinomial model.

For this purpose CxDxF tables are simulated on a computer for several types of probability structures with several sample sizes.

A BASIC program is presented for adjusting log-linear models and variance analysis and contingency tables are presented to determine the dependence structure.

The statistical method used is the one developed by GRIZZLE, STARMER and KOCH (1969) based on linear models and weighted least square that can be used to analyse contingency tables.

The results showed that for samples as large as 200, 250 and 300, the statistical tests based on the GSK method are good to detect alternatives more complex than H_o. For sample sizes of 1000 and 1500 the GSK method detects exactly the probability structure.

Berkson's method for incomplete contingency tables showed good behavior in all cases studied.

In opposition to GOODMAN (1971b) opinion, the GSK method generally is easy to perform.

1. INTRODUÇÃO

Frequentemente os resultados de uma pesquisa são obtidos na forma de frequências (absolutas) com que as unidades experimentais (Ex. indivíduos, objetos, etc.) são classificadas (variáveis qualitativas) em um certo número de categorias de respostas, frequências essas que seguem uma distribuição multinomial.

Essas tabelas representam uma estrutura e apresentam relações das quais o pesquisador deseja avaliar certas questões de interesse, de forma que uma formulação precisa de hipóteses correspondentes a essas questões é necessária para a realização desse objetivo.

Para verificar essas hipóteses, os testes estatísticos usualmente empregados são: χ^2 de PEARSON, χ_1^2 modificado de NEYMAN e teste da razão de verossimilhança.

Neste trabalho será utilizado o método de análise proposto por GRIZZLE, STARMER e KOCH (1969), os quais introduzem o uso de modelos lineares e quadrados mínimos ponderados na análise de dados dispostos em uma tabela de contingência. Esse método tem a vantagem de permitir ao analista maior flexibilidade na escolha de modelos e testes de hipóteses a uma grande variedade de situações experimentais. A justificativa teórica do método pode ser encontrada em WALD (1943) e NEYMAN (1949). A teoria desenvolvida por GRIZZLE et alii (1969) é baseada nos estudos de BHAPKAR (1966), que mostra a equivalência da estatística de WALD e a estatística de NEYMAN.

O teste de χ^2 de PEARSON não possibilita a partição de hipóteses, enquanto que nos testes da razão de verossimilhança e método GSK fica possível estudar-se o efeito isolado de um fator e são cada vez mais comuns.

GRIZZLE e WILLIAMS (1972), descrevem métodos de ajustamento de modelos log-lineares a dados multinomiais tabulados em tabelas multidimensionais. Nesse artigo é dada atenção às similaridades entre os métodos propostos e a análise de dados contínuos por modelos lineares sendo estabelecidas relações entre os testes de independência marginal e certos testes associados com a análise de experimentos fatoriais.

No presente trabalho são testadas hipóteses de independência estabelecidas a priori, através da metodologia GSK, em tabelas de contingência obtidas por simulação de dados de situações de pequenas populações, e posteriormente procurar-se-á explorar os problemas de inferência causados pela aproximação assintótica dos testes, objetivando determinar as condições práticas para obtenção de testes e estimativas confiáveis.

Deve-se ressaltar que o trabalho limita-se ao estudo de problemas envolvendo uma única população, sobre ajustamento de modelos log-lineares e determinação de estruturas de independência em tabelas de contingência cruzadas C x D x F.

WADA (1977), coloca uma observação, também encontrada no artigo original de GRIZZLE _et alii_ (1969), que se refere ao desconhecimento do método em pequenas amostras; apesar disso, reforçam o fato de que os testes estatísticos nele empregados apresentam ótimas propriedades assintóticas, como demonstrou NEYMAN (1949).

Esta metodologia geral, em virtude da grande quantidade de produtos e inversões de matrizes envolvidos, requer um programa de computador adequado. Um programa que pode ser utilizado é o "GENCAT", desenvolvido por LANDIS _et alii_ (1976), para a análise de dados categorizados multivariados. Outro procedimento é o "FUNCAT" incorporado ao SAS - Statistical Analysis System - versão 1979. Existe também o programa "CRISCAT"-1978, mais adequado para análise de dados de sobrevivência, na área médica.

Dentre os objetivos do presente trabalho destacam-se os seguintes:

a) Viabilização computacional da metodologia GSK, linguagem BASIC, para a análise de dados categorizados multivariados provindos de uma única distribuição multinomial, a qual envolve um procedimento geral em dois estágios.

i) A construção de funções apropriadas das proporções observadas.

ii) A construção de testes estatísticos para hipóteses envolvendo essas funções e a estimação dos correspondentes parâmetros de modelos convenientes, através do método de mínimos quadrados ponderados.

Procurar-se-á inicialmente assimilar os melhores procedimentos de cálculos, a fim de elaborar-se subrotinas adequadas para a solução desse tipo de problema, envolvendo duas classes de funções: lineares e logarítmicas.

b) Procurar-se-á explorar os problemas de inferência causados pela aproximação assintótica, simulando situações de pequenas populações, estimando em correspondência as probabilidades de decisões erradas, objetivando determinar as condições práticas para obtenção de testes de hipóteses confiáveis.

c) Ajustamento de modelos log-lineares a dados multinomiais dispostos em tabelas de contingência multidimensionais; estimar os parâmetros dos mesmos, buscando o modelo que melhor explique as associações entre as diversas variáveis em estudo.

d) Determinação de hipóteses adequadas de independência em tabelas de contingência multidimensionais.

e) Testes de não interação em tabelas de contingência multidimensionais.

2. REVISÃO DE LITERATURA

PLACKETT (1961) considera uma tabela de contingência com três classificações. A razão do número três está na medida de cálculo e, como ocorre em experimentos fatoriais, a interpretação dos resultados aumenta em dificuldade com a ordem da tabela. Há muitas hipóteses provavelmente de interesse, como por exemplo, os três sistemas são mutuamente independentes, ou uma classificação é independente das outras duas, que não levantam dificuldades de definição e podem ser respondidas usando princípios padrões. Quando, entretanto, uma definição paramétrica de interação é requerida, diferentes proposições têm sido feitas; e no caso de interações de segunda ordem, testes comparáveis são avaliados, os quais, podem conduzir a resultados muitíssimo diferentes. Depois discute as questões envolvidas, desenvolve uma técnica computacional alternativa para tabela com três entradas.

BHAPKAR (1966) mostra em suas notas que para testar hipóteses lineares em dados categorizados, a estatística de Wald (1943) é equivalente (idêntica algebricamente) à estatística de χ_1^2 de NEYMAN (1949), e para testar hipóteses não lineares ela é equivalente à estatística χ_1^2 usando a técnica de linearização de Neyman, sempre que essas estatísticas χ_1^2 são definidas.

BERKSON (1968) evidencia que o χ^2 logit mínimo dá, para todos os propósitos práticos, numericamente as mesmas estimativas e testes estatísticos como aqueles obtidos por máxima verossimilhança e χ^2 de Pearson para uma variedade de problemas. Pois (GRIZZLE et alii, 1969), o χ^2 logit mínimo é apenas membro da família que aqueles testes possuem, ou seja, em grandes amostras possuem distribuição de χ^2 se a hipótese de nulidade é verdadeira e que explorando este fato conduz a aproximações unificadas, em ambos os aspectos conceituais e computacionais.

BHAPKAR e KOCH (1968a), dão ênfase para a distinção entre "fatores" (tais como tratamentos ou blocos) os quais têm os totais marginais fixados e "respostas" (tais como categorias de respostas) as quais têm os totais marginais aleatórios. Assim, quatro casos principais aparecem: (i) as tabelas "multi-respostas, nenhum fator", (ii) as tabelas "multi-respostas, um fator", (iii) as tabelas "multi-respostas, multi-fatores", (iv) as tabelas "uma resposta, multi-fatores". Para as situações (i) e (ii) o conceito de "não interação" está relacionado à questões que não considera o modelo de associação entre as respostas. Para a situação (iv) o conceito de "não-interação" está relacionado ao modo como as combinações dos fatores de terminam a distribuição de respostas. Para a situação (iii) ambos os tipos de questões aparecem. Para os diferentes tipos de tabelas o problema de formulação de hipóteses apropriadas de "não interação" são formuladas. Os testes estatísticos são baseados no critério de WALD (1943).

GOODMAN (1968) faz distinção entre dois tipos diferentes de tabelas de contingência: (A) tabelas de contingência truncadas, nas quais as frequências em algumas das caselas da tabela são omitidas da análise; e (B) as mais usuais nas quais nenhuma das caselas da tabela são omitidas. É nisso que vários métodos, apropriados para a análise de tabelas de contingência não truncadas,conduzirão (quando apropriadamente modificados) a métodos adequados para a análise de tabelas truncadas e que, inversamente, vários métodos adequados para análise de tabelas truncadas conduzirá (quando apropriadamente ampliados) a métodos adequados para a análise de tabelas não truncadas.

Para tabela de contingência não truncada RxC (R linhas e C colunas) é comum na prática calcular-se a estatística χ^2 de Karl Pearson para testar-se a hipótese de nulidade de independência usando o número apropriado de graus de liberdade. GOODMAN (1968) adiciona alguns novos métodos de análise, um dos quais é um procedimento em dois passos.

A definição do conceito de independência em uma tabela RxC e o conceito de quase-independência são apresentados. Apresenta um método iterativo que requer menos cálculos aritméticos do que os métodos de CAUSSINUS (não publicado) e FIEMBERG (1965) e é também de mais fácil aplicação no caso geral do que o método de WATSON (1956) para análise de quase-independência em tabelas de contingência truncadas.

O autor inclui neste artigo alguma discussão sobre as estimativas das interações presentes em tabelas de contingência truncadas e nao truncadas.

A maioria dos métodos apresentados neste artigo podem ser aplicados na análise de "quase-homogeneidade" entre R populações multinomiais bem como na análise de "quase-independência" em tabelas RxC. Define "quase-homogeneidade" e diz que quando há independência, as R populações multinomiais são "homogêneas".

Os métodos apresentados no artigo podem ser usados para testar a hipótese de nulidade de que as R populações multinomiais sao quase homogêneas, e eles podem também ser usados para estimar as proporções hipotéticas b_j (para $j = 1, 2, \ldots, C$). Similarmente, na análise de tabela RxC, os métodos apresentados neste artigo podem ser usados para testar a hipótese nula de quase independência, e (no caso de quase independência) eles podem também ser usados para estimar que proporção hipotética teria sido na i-ésima linha e j-ésima coluna da tabela se nenhuma das caselas for nula, e se as linhas e colunas da tabela forem independentes.

GRIZZLE et alii (1969) descrevem um procedimento geral não-iterativo para ajustar funções a um modelo linear, para testar o ajustamento do modelo, e para testar hipóteses sobre os parâmetros do modelo linear. Os casos especiais de funções lineares e funções logarítmicas dos π_{ij} são desenvolvidos em detalhes, e alguns exemplos de como a aproximação geral pode ser usada para analisar vários tipos de dados categorizados são apresentados.

ODOROFF (1970) mostra por simulação que, em tabelas pequenas para as quais o logit foi apropriado, o método χ^2 modificado mínimo dá melhores resultados do que o método de máxima verossimilhança. Também sugere uma modificação do logit e sua variância, melhorando as propriedades do método de χ^2 modificado mínimo. Não observa importantes diferenças no poder dos testes. Observa também que, em geral, os níveis exatos dos testes de χ^2 logit mínimo e os testes de χ^2 de Pearson são mais aproximados da distribuição de qui-quadrado tabelada do que os níveis dos testes da razão de verossimilhança.

JOHNSON e KOCH (1971), mostram que a técnica de quadrados mínimos ponderados é uma ferramenta útil na formação e avaliação de certos modelos apropriados para dados na forma de tabelas de contingência multi-dimensionais. O tipo dessa aproximação é análoga para aquela usada em regressão "stepwise' para dados contínuos.

GOODMAN (1971b) considera tabelas de contingência com m entradas, modelos que descrevem as possíveis interações multiplicativas entre as m variáveis da tabela, e mostra como selecionar modelos que ajustam os dados da tabela, usando métodos que são, em parte, análogos ao procedimento 'stepwise' na análise de regressão. Também apresenta um método direto de estimação para construir modelos de classificação múltipla. O autor aponta muitas dificuldades na aplicação do método GSK (para cada hipótese de interesse, a aplicação do método requer o uso de uma diferente "matriz C e/ou K", e também que a "matriz do delineamento X" é diferente da matriz usual usada na correspondente análise de regressão múltipla, se a variável dependente na análise de logit é politômica). O autor assinala também dificuldades computacionais em modelos log-lineares.

GRIZZLE e WILLIAMS (1972), apresentam métodos estatísticos para testar independência em tabelas de contingência multidimensionais baseados na correspondência entre a análise de experimentos fatoriais e testes de independência marginal. Esta relação simplifica a interpretação das análises de tabelas de contingência e conduz a testes simplificados para tabelas nas quais algumas das probabilidades das caselas é zero. Modelos lineares são usados para calcular estimativas das probabilidades.

É necessário fazer-se uma observação quanto às fórmulas de S_{11}, S_{12} e S_{22} apresentadas na página 146 do artigo original (BIOMETRICS, março 1972), as quais devem ser tomadas como

$$S_{11} = K_1 D_1^{-1} V(\underset{\sim}{p}) D_1^{-1} K_1' \;; \quad S_{12} = K_1 D_1^{-1} V(\underset{\sim}{p}) D_2^{-1} K_2' \;;$$

$$S_{22} = K_2 D_2^{-1} V(\underset{\sim}{p}) D_2^{-1} K_2' \;, \quad \text{respectivamente.}$$

Outro ponto a observar está no final da página 140, onde o correto é a hipótese $\pi_{ijk} = a_{jk} b_i$ ao invés da hipótese $\pi_{ijk} = a_{ij} b_k$. Com isso deve-se corrigir o cabeçalho da tabela 3 (pág. 141), onde lê-se $\pi_{ijk} = \pi_{ij.}\pi_{..k}$ deve-se ler $\pi_{ijk} = \pi_{.jk}\pi_{i..}$.

WILLIAMS e GRIZZLE (1972), apresentam dois métodos para analisar tabelas de contingência com categorias de respostas ordenadas. Um método é análogo ao teste de posição quando há muitos empates e o outro é baseado no modelo escalar Thurstonian. Ambos os métodos usam estimativas de χ_1^2 mínimo para os parâmetros de modelos lineares e resultados de cada análise podem ser apresentados na forma de uma análise de variância. Exemplos de respostas univariadas e bivariadas são incluídas. Adaptam o método GSK ao modelo proposto por Thurstone.

LANDIS *et alii* (1976) fazem um programa de computador, o GENCAT, que concretiza uma metodologia geral para a análise de dados categorizados multivariados. Qualquer função composta das proporções observadas a qual pode ser formulada como uma sequência das seguintes transformações do vetor de dados - linear, logarítmica, exponencial, ou adição de um vetor de constantes - pode ser analizada dentro deste programa geral. Esse algorítmo produz estatísticas de qui-quadrado mínimo modificado as quais são obtidas partindo as somas de quadrados como na ANOVA.

WADA (1977), trata no capítulo III do estudo de hipóteses como as de independência e associação sobre a diagonal principal, que, para a aplicação do método GSK, necessitam ser escritas sob a forma

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

Em algumas tabelas como as de três entradas (modelo "três respostas, nenhum fator"), na utilização do método GSK, existe grande dificuldade para a construção de matrizes K adequadas. Nesse trabalho, para facilitar a obtenção dessas matrizes, para as hipóteses de independência e de não interação, é apresentada uma técnica sugerida por GRIZZLE e WILLIAMS (1972), e que consiste na construção 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.

Um ponto a considerar é no que se refere aos resultados encontrados pela autora para as estatísticas $F'S^{-1}F$, das páginas 54, 55, 56, 57, 58 e 62 os quais merecem uma nova verificação.

HABER (1983), considera o teste exato bilateral das hipóteses de não interação em uma tabela 2x2x2 com os totais de linhas e colunas fixados. Um método de aproximar o poder do teste, foi introduzido e encontrou-se um ajuste bem próximo. Este método pode ser utilizado para calcular os tamanhos de amostras requeridos para detectar efeitos de interação com um poder pré-estabelecido.

3. METODOLOGIA

3.1. Método GSK para a Análise dos Dados

Os métodos a serem utilizados são os apresentados por GRIZZLE, STARMER e KOCH (1969) e a sua notação é utilizada como mostra o resumo a seguir. Considere os dados hipotéticos mostrados na tabela 1 e as probabilidades esperadas das caselas mostradas na tabela 2.

TABELA 1 - Distribuição de Frequências.

POPULAÇÕES (FATORES)	CATEGORIAS DE RESPOSTAS				TOTAL
	1	2	....	r	
1	n_{11}	n_{12}	....	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	....	n_{2r}	$n_{2.}$
.	.	.	....	.	.
.	.	.	....	.	.
.	.	.	....	.	.
s	n_{s1}	n_{s2}	....	n_{sr}	$n_{s.}$

TABELA 2 - Probabilidades esperadas das caselas.

POPULAÇÕES (FATORES)	CATEGORIAS DE RESPOSTAS				TOTAL
	1	2	....	r	
1	π_{11}	π_{12}	....	π_{1r}	1
2	π_{21}	π_{22}	....	π_{2r}	1
.	.	.	....	.	.
.	.	.	....	.	.
.	.	.	....	.	.
s	π_{s1}	π_{s2}	....	π_{sr}	1

O método GSK baseia-se nas mesmas hipóteses da estatística de WALD (1943) adaptada a dados categorizados. Assim, para dados categorizados sob o "modelo geral de classificação cruzada", isto é, quando s amostras simples e independentes de tamanhos $n_1, n_2, \ldots, n_s$ são retiradas de s populações independentes, e os elementos de cada amostra classificados em r categorias (distintas) de respostas, definem-se

$$\underset{1\times r}{\underset{\sim}{\pi}_i'} = \left[\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir}\right] \quad ; \quad \underset{1\times rs}{\underset{\sim}{\pi}'} = \left[\underset{\sim}{\pi}_1', \underset{\sim}{\pi}_2', \ldots, \underset{\sim}{\pi}_s'\right] \quad ;$$

os estimadores

$$\underset{1\times r}{\underset{\sim}{p}_i'} = \left[p_{i1}, p_{i2}, \ldots, p_{ir}\right] \quad ; \quad \underset{1\times rs}{\underset{\sim}{p}'} = \left[\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s'\right] \quad ;$$

com $p_{ij} = \dfrac{n_{ij}}{n_{i.}}$, e ainda

$$\underset{\underset{\sim}{\pi}_i}{\mathrm{Var}}(\underset{\sim}{p}_i) = \underset{r\times r}{V(\underset{\sim}{\pi}_i)} = \frac{1}{n_{i.}} \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \cdots & -\pi_{i1}\pi_{ir} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \cdots & -\pi_{i2}\pi_{ir} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ -\pi_{ir}\pi_{i1} & -\pi_{ir}\pi_{i2} & \cdots & \pi_{ir}(1-\pi_{ir}) \end{bmatrix}$$

$$\underset{r\times r}{V(\underset{\sim}{p}_i)} = \text{estimativa de } V(\underset{\sim}{\pi}_i) \; .$$

Para $\underset{rs\times rs}{V(\underset{\sim}{\pi})}$, matriz de variâncias e covariâncias de $\underset{\sim}{p}$, o estimador é denotado por $\underset{rs\times rs}{V(\underset{\sim}{p})}$, matriz bloco diagonal contendo $V(\underset{\sim}{p}_i)$ na diagonal principal, isto é:

$$\underset{rs\times rs}{V(\underset{\sim}{p})} = \begin{bmatrix} V(\underset{\sim}{p}_1) & \emptyset & \ldots & \emptyset \\ \emptyset & V(\underset{\sim}{p}_2) & \ldots & \emptyset \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \emptyset & \emptyset & \ldots & V(\underset{\sim}{p}_s) \end{bmatrix}$$

Sejam:

$f_m(\underset{\sim}{\pi})$: qualquer função dos elementos de $\underset{\sim}{\pi}$ que tem derivadas parciais contínuas até 2ª ordem com respeito aos elementos de π_{ij} ; $m = 1, 2, \ldots, u \leq rs - s$;

$f_m(\underset{\sim}{p})$: $f_m(\underset{\sim}{\pi})$ calculadas para $\underset{\sim}{\pi} = \underset{\sim}{p}$; funções de $\underset{\sim}{p}$ que descrevem em algum aspecto a relação entre a distribuição das respostas (categorias de respostas) e a natureza das subpopulações ;

compõem os vetores,

$$\left[\underset{1\times u}{F(\underset{\sim}{\pi})}\right]' = \left[f_1(\underset{\sim}{\pi}),\ f_2(\underset{\sim}{\pi}),\ \ldots,\ f_u(\underset{\sim}{\pi})\right] ;$$

$$\underset{1\times u}{F'} = [F(\underset{\sim}{p})]' = [f_1(\underset{\sim}{p}),\ f_2(\underset{\sim}{p}),\ \ldots,\ f_u(\underset{\sim}{p})]$$

A matriz de variâncias-covariâncias de F e estimada por:

$$\underset{u\times u}{V_F} = HV(\underset{\sim}{p})H' \quad , \quad \text{onde}$$

$$\underset{u\times rs}{H} = \left[\left. \frac{\partial\, f_m(\underset{\sim}{\pi})}{\partial\, \pi_{ij}} \right|_{\pi_{ij} = p_{ij}} \right].$$

Se as funções $f_m(\underset{\sim}{p})$ forem funções lineares de $\underset{\sim}{p}$ para $m = 1, 2, \ldots, u$, a matriz de variâncias-covariâncias de $\underset{\sim}{F}$ é exata e, quando $F_m(\underset{\sim}{p})$, $m = 1, 2, \ldots, u$, for uma função não linear dos elementos de $\underset{\sim}{p}$, $V_{\underset{\sim}{F}}$ é a matriz de estimativas de variâncias-covariâncias consistente de $\underset{\sim}{F}$, obtida pelo "método da linearização" ou "método de Neyman" (NEYMAN, 1949).

Neste ponto, é importante observar que em todas as aplicações práticas, as funções $F_m(\underset{\sim}{\pi})$, $m = 1, 2, \ldots, u$, são escolhidas de tal forma que a matriz de variâncias-covariâncias de $\underset{\sim}{F}$ seja assintóticamente não singular. Isto é, por hipótese, as funções $F_m(\underset{\sim}{\pi})$, $m= 1, 2, \ldots, u$, são independentes e também independentes das restrições

$$\sum_{j=1}^{r} \pi_{ij} = 1 \quad , \quad i = 1, 2, \ldots, s$$

(de modo que $u \leq rs - s$), e consequentemente H e $V_{\underset{\sim}{F}}$ são de posto u.

Contudo, para alguns tipos de dados, se para algum (i, j), $n_{ij} = 0$, $\underset{\sim}{V}_F$ será de posto menor que u , e a inversão de $\underset{\sim}{V}_F$ não poderá ser realizada, e neste caso o procedimento de BERKSON (1955) poderá ser usado; este determina que o valor zero seja substituído por 1/r, implicando que a estimativa de π_{ij} seja igual a $1/(n_{i.}r)$. Outros procedimentos consistem simplesmente em suprimir algumas funções de $\underset{\sim}{F}(\underset{\sim}{\pi})$ ou combinar duas ou mais funções dentro de $\underset{\sim}{F}(\underset{\sim}{\pi})$, formando médias apropriadas de maneira que a matriz de variâncias e covariâncias estimada $V_{\underset{\sim}{F}_r}$, que corresponde ao vetor reduzido $\underset{\sim}{F}_r$ seja não singular. Uma discussão mais detalhada a respeito das dificuldades ocasionadas por caselas vazias, é apresentada em KOCH et alii (1978) e KOCH *et alii* (1975).

Admita-se que $\underset{u\times 1}{F(\tilde{\pi})} = {}_u X_v \underset{\sim}{\beta} + \underset{u\times 1}{\underset{\sim}{\xi}}$, onde X é uma matriz de delineamento conhecida (que difere da matriz usual de delineamento se mais de uma função é construída dentro de cada população) de posto $v \leq u$; $\underset{\sim}{\beta}$ é um vetor de parâmetros desconhecidos, e $\underset{\sim}{\xi}$ é o vetor dos erros, de tal modo que

$$E\left[\underset{\sim}{F}(\underset{\sim}{\pi})\right] = X\underset{\sim}{\beta} \quad \text{e} \quad \text{var}\left[\underset{\sim}{F}(\underset{\sim}{\pi})\right] = \text{var}(\underset{\sim}{\xi}) = V_{\underset{\sim}{F}} \ .$$

Se o modelo dado acima ajusta os dados, o melhor estimador assintoticamente normal (BAN) de $\underset{\sim}{\beta}$ é dado por $\hat{\underset{\sim}{\beta}}$ (obtido pelo método de mínimos quadrados generalizados e que minimiza $(\underset{\sim}{F}-X\underset{\sim}{\beta})'V_{\underset{\sim}{F}}^{-1}(\underset{\sim}{F}-X\underset{\sim}{\beta})$), onde:

$$\hat{\underset{\sim}{\beta}} = \left[X' V_{\underset{\sim}{F}}^{-1} X\right]^{-1} X' V_F^{-1} \underset{\sim}{F} \quad \text{e} \quad V(\hat{\underset{\sim}{\beta}}) = \left[X' V_{\underset{\sim}{F}}^{-1} X\right]^{-1}$$

(matriz de variâncias e covariâncias de $\hat{\underset{\sim}{\beta}}$) .

O teste estatístico para o ajuste do modelo é:

$$S.Q. \left[\underset{\sim}{F}(\underset{\sim}{\pi}) = X\underset{\sim}{\beta}\right] = \underset{\sim}{F}' V_{\underset{\sim}{F}}^{-1} \underset{\sim}{F} - \hat{\underset{\sim}{\beta}}' \, (X' V_{\underset{\sim}{F}}^{-1} X) \, \hat{\underset{\sim}{\beta}} \quad ,$$

o qual é distribuído assintoticamente como uma distribuição de χ^2 (central) com u-v g.l., dado que o modelo seja verdadeiro.

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

Sendo o modelo adequado aos dados, um teste geral da hipótese H_o: ${}_dC_v \underset{\sim}{\beta}_1 = \underset{d \times 1}{\underset{\sim}{0}}$ é dado por

$$S.Q. \left[C\underset{\sim}{\beta} = \underset{\sim}{0}\right] = \hat{\underset{\sim}{\beta}}' \, C' \left[C(X' V_{\underset{\sim}{F}}^{-1} X)^{-1} \, C'\right]^{-1} C\hat{\underset{\sim}{\beta}}$$

o qual tem distribuição assintótica de χ^2 com d graus de liberdade se H_o for verdadeira, onde d é o posto de C .

Portanto, esses métodos permitem que a variação total dentro de $\underset{\sim}{F}(\underset{\sim}{\pi})$ seja subdividida em específicas fontes de variações e, consequentemente, representam uma análise da variância, estatisticamente válida para o correspondente vetor de funções estimadas $\underset{\sim}{F}(\underset{\sim}{p})$.

3.1.1. Casos especiais de $\underset{\sim}{F}(\underset{\sim}{\pi})$

A forma de $V_{\underset{\sim}{F}}$ depende de H e, através de H, das funções $\underset{\sim}{F}(\underset{\sim}{\pi})$. Sendo assim, para cada família de funções $\underset{\sim}{F}(\underset{\sim}{\pi})$, $V_{\underset{\sim}{F}}$ será diferente. Duas classes de funções cobrem o maior número de aplicações discutidas na literatura.

Relações lineares definidas pela família de funções

$$\underset{\sim}{F}(\underset{\sim}{\pi}) = {}_{u}A_{rs}\underset{\sim}{\pi}_{1} ,$$

onde A é uma matriz de constantes a_{ijk} de posto $u \leq rs - s$ dada por

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

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

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

e, portanto, uma estimativa da matriz de variâncias e covariâncias de $\underset{\sim}{F}$ é dada por

$$V_{\underset{\sim}{F}} = AV(\underset{\sim}{p})A' \quad .$$

Para relações logarítmicas pode-se definir a família de funções

$$\underset{\tilde{t}x\tilde{1}}{F(\pi)} = \underset{txu}{K} \; \ln \underset{uxrs}{A} \; \underset{rs\tilde{x}1}{\pi} \; ;$$

onde K é uma matriz de constantes arbitrárias de posto $t \leq u \leq rs$. A matriz A é uma matriz de transformação de posto u e $\log_e$ transforma um vetor no correspondente vetor de logaritmos naturais.

A matriz $V_{\tilde{F}} = HV(\tilde{p})H'$, de posto completo ficará

$$H = \left[\frac{\partial \tilde{F}}{\partial \tilde{\pi}} \Bigg|_{\tilde{\pi} = \tilde{p}} \right] = K\, D^{-1}\, A \quad ,$$

e a matriz $V_{\tilde{F}}$ fica

$$V_{\tilde{F}} = K\, D^{-1}\, AV(\tilde{p})A'\, D^{-1}\, K'$$

com,

$$D = \begin{bmatrix} \tilde{a}_1'\tilde{p} & 0 & \ldots & 0 \\ 0 & \tilde{a}_2'\tilde{p} & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & \ldots & \tilde{a}_\alpha'\tilde{p} \end{bmatrix}$$

onde $\tilde{a}_\gamma'$ representa a γ-ésima linha de A .

Essas duas classes de funções, lineares e logarítmicas não esgotam todas as possibilidades de situações práticas, onde funções mais gerais devem ser consideradas. Essas funções são formuladas em termos de funções compostas de transformações lineares, logarítmicas e exponenciais e os detalhes sobre a formulação dessas funções podem ser encontradas em FORTHOFER e KOCH (1973), WADA (1977) e CANTON (1980).

3.1.2. Relações entre independência e modelos log-lineares

Os modelos log-lineares são úteis para se expressar as inter-relações das variáveis que definem uma tabela de contingência.

Considere uma amostra de tamanho n retirada de uma população e classificada segundo três respostas, C, D e F, com c, d e f categorias, respectivamente, constituindo uma única distribuição multinomial.

Desse modo, na notação GSK, $s = 1$, e o número de categorias de respostas é $r = c.d.f$, resultando num modelo de três respostas e nenhum fator. São situações como, por exemplo, sexo, idade e aproveitamento escolar; técnicos, tipos de testes e tipos de defeitos. O modelo é chamado "nenhum fator" porque nenhuma das marginais é fixa e onde somente o tamanho da amostra é fixo.

Deve-se entender o termo "fator" como uma classificação caracterizada por uma descrição de alguns aspectos da população ou subpopulação as quais os sujeitos pertencem (por exemplo, blocos, tratamentos) e nesse caso os totais marginais referentes aos níveis do fator são fixos; pelo termo "resposta", entende-se uma classificação caracterizada pela des crição do que acontece com os sujeitos durante o experimento (por exemplo, morre ou vive; é baixo, médio, alto), e nesse caso os totais marginais não são fixos antes da realização do experimento. Assim estamos trabalhando com s (igual ao número de fatores) distribuições multinomiais independentes, cada uma com r (igual ao número de respostas) caselas e com um número total de observações fixo.

Os dados provindos de uma população possuem somente o tamanho n da amostra fixo, e os totais marginais são aleatórias, qualquer que seja o número de entradas considerado na tabela. Então, os números observados no corpo da tabela tem distribuição multinomial.

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 o 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, tem-se um "modelo misto de ordem m", e os dados resumidos em tabelas de contingência dxdxdx ... xd.

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

Seja π_{ijk} a probabilidade associada a casela ijk , $i = 1, 2, \ldots, c$; $j = 1, 2, \ldots, d$ 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 tal que $\sum_i \sum_j \sum_k \pi_{ijk} = 1$.

Hipóteses descrevendo três tipos de independência podem ser formuladas:

(1) independência mútua completa

$$H_o : \pi_{ijk} = \pi_{i..}\ \pi_{.j.}\ \pi_{..k}\ , \quad \text{se} \quad \pi_{ijk} \neq 0 \begin{cases} i = 1, 2, \ldots, c \\ j = 1, 2, \ldots, d \\ k = 1, 2, \ldots, f \end{cases}$$

(2) independência parcial, quando duas classificações são independentes da terceira:

$$H_o : \pi_{ijk} = \pi_{ij.}\ \pi_{..k}\ , \quad \text{se} \quad \pi_{ijk} \neq 0 \begin{cases} i = 1, 2, \ldots, c \\ j = 1, 2, \ldots, d \\ k = 1, 2, \ldots, f \end{cases}$$

Há três hipóteses deste tipo.

(3) independência condicional, para verificar se, dada uma classificação, as outras duas são independentes, formulada por

$$H_o : \pi_{ijk} = \pi_{i.k}\,\pi_{.jk} \,/\, \pi_{..k} \;, \quad \text{se} \quad \pi_{ijk} \neq 0 \quad \text{e} \quad \pi_{..k} \neq 0 \qquad \begin{cases} i = 1, 2, \ldots, c \\ j = 1, 2, \ldots, d \\ k = 1, 2, \ldots, f \end{cases}$$

Há três hipóteses deste tipo.

Essas hipóteses podem estar ligadas à análise de um experimento fatorial cxdxf, com c níveis da classificação C, d níveis da classificação D e f níveis da classificação F .

Considere a hipótese $\pi_{ijk} = \pi_{.jk}\,\pi_{i..}$, no qual c, d e f são iguais a três. Se esta hipótese é verdadeira, então a relação mostrada na tabela 3 também é verdadeira.

TABELA 3 - Probabilidades das caselas para a hipótese $\pi_{ijk} = \pi_{.jk}\,\pi_{i..}$

		C_1	C_2	C_3
	F_1	$\pi_{111} = \pi_{.11}\,\pi_{1..}$	$\pi_{221} = \pi_{.11}\,\pi_{2..}$	$\pi_{311} = \pi_{.11}\,\pi_{3..}$
D_1	F_2	$\pi_{112} = \pi_{.12}\,\pi_{1..}$	$\pi_{212} = \pi_{.12}\,\pi_{2..}$	$\pi_{312} = \pi_{.12}\,\pi_{3..}$
	F_3	$\pi_{113} = \pi_{.13}\,\pi_{1..}$	$\pi_{213} = \pi_{.13}\,\pi_{2..}$	$\pi_{313} = \pi_{.13}\,\pi_{3..}$
	F_1	$\pi_{121} = \pi_{.21}\,\pi_{1..}$	$\pi_{221} = \pi_{.21}\,\pi_{2..}$	$\pi_{321} = \pi_{.21}\,\pi_{3..}$
D_2	F_2	$\pi_{122} = \pi_{.22}\,\pi_{1..}$	$\pi_{222} = \pi_{.22}\,\pi_{2..}$	$\pi_{322} = \pi_{.22}\,\pi_{3..}$
	F_3	$\pi_{123} = \pi_{.23}\,\pi_{1..}$	$\pi_{223} = \pi_{.23}\,\pi_{2..}$	$\pi_{323} = \pi_{.23}\,\pi_{3..}$
	F_1	$\pi_{131} = \pi_{.31}\,\pi_{1..}$	$\pi_{231} = \pi_{.31}\,\pi_{2..}$	$\pi_{331} = \pi_{.31}\,\pi_{3..}$
D_3	F_2	$\pi_{132} = \pi_{.32}\,\pi_{1..}$	$\pi_{232} = \pi_{.32}\,\pi_{2..}$	$\pi_{332} = \pi_{.32}\,\pi_{3..}$
	F_3	$\pi_{133} = \pi_{.33}\,\pi_{1..}$	$\pi_{233} = \pi_{.33}\,\pi_{2..}$	$\pi_{333} = \pi_{.33}\,\pi_{3..}$

Da tabela 3 segue-se que:

$$\frac{\pi_{111}\,\pi_{112}\,\pi_{113}}{\pi_{131}\,\pi_{132}\,\pi_{133}} = \frac{\pi_{\cdot 11}\,\pi_{\cdot 12}\,\pi_{\cdot 13}}{\pi_{\cdot 31}\,\pi_{\cdot 32}\,\pi_{\cdot 33}} \quad \text{e} \quad \frac{\pi_{311}\,\pi_{312}\,\pi_{313}}{\pi_{331}\,\pi_{332}\,\pi_{333}} = \frac{\pi_{\cdot 11}\,\pi_{\cdot 12}\,\pi_{\cdot 13}}{\pi_{\cdot 31}\,\pi_{\cdot 32}\,\pi_{\cdot 33}}$$

tal que

$$\frac{\pi_{111}\,\pi_{112}\,\pi_{113}\,\pi_{331}\,\pi_{332}\,\pi_{333}}{\pi_{131}\,\pi_{132}\,\pi_{133}\,\pi_{311}\,\pi_{312}\,\pi_{313}} = 1 \quad ; \qquad (1)$$

$$\frac{\pi_{111}\,\pi_{121}\,\pi_{131}}{\pi_{113}\,\pi_{123}\,\pi_{133}} = \frac{\pi_{\cdot 11}\,\pi_{\cdot 21}\,\pi_{\cdot 31}}{\pi_{\cdot 13}\,\pi_{\cdot 23}\,\pi_{\cdot 33}} \quad \text{e} \quad \frac{\pi_{311}\,\pi_{321}\,\pi_{331}}{\pi_{313}\,\pi_{323}\,\pi_{333}} = \frac{\pi_{\cdot 11}\,\pi_{\cdot 21}\,\pi_{\cdot 31}}{\pi_{\cdot 13}\,\pi_{\cdot 23}\,\pi_{\cdot 33}}$$

tal que

$$\frac{\pi_{111}\,\pi_{121}\,\pi_{131}\,\pi_{313}\,\pi_{323}\,\pi_{333}}{\pi_{113}\,\pi_{123}\,\pi_{133}\,\pi_{311}\,\pi_{321}\,\pi_{331}} = 1 \quad \text{e,} \qquad (2)$$

$$\frac{\pi_{111}\,\pi_{331}}{\pi_{113}\,\pi_{333}} = \frac{\pi_{\cdot 11}\,\pi_{\cdot 31}}{\pi_{\cdot 13}\,\pi_{\cdot 33}} \quad \text{e} \quad \frac{\pi_{131}\,\pi_{311}}{\pi_{133}\,\pi_{313}} = \frac{\pi_{\cdot 31}\,\pi_{\cdot 11}}{\pi_{\cdot 33}\,\pi_{\cdot 13}}$$

tal que

$$\frac{\pi_{111}\,\pi_{331}\,\pi_{133}\,\pi_{313}}{\pi_{113}\,\pi_{333}\,\pi_{131}\,\pi_{311}} = 1 \quad . \qquad (3)$$

Tomando-se os logarítmos naturais de (1), (2) e (3), obtém-se:

$$\ln \pi_{111} + \ln \pi_{112} + \ln \pi_{113} - \ln \pi_{131} - \ln \pi_{132} - \ln \pi_{133} -$$

$$- \ln \pi_{311} - \ln \pi_{312} - \ln \pi_{313} + \ln \pi_{331} + \ln \pi_{332} +$$

$$+ \ln \pi_{333} = 0 \qquad (4)$$

$$\ln \pi_{111} - \ln \pi_{113} + \ln \pi_{121} - \ln \pi_{123} + \ln \pi_{131} - \ln \pi_{133} -$$

$$- \ln \pi_{311} + \ln \pi_{313} - \ln \pi_{321} + \ln \pi_{323} - \ln \pi_{331} +$$

$$+ \ln \pi_{333} = 0 \qquad (5)$$

e

$$\ln \pi_{111} - \ln \pi_{113} - \ln \pi_{131} + \ln \pi_{133} - \ln \pi_{311} + \ln \pi_{313} +$$

$$+ \ln \pi_{331} - \ln \pi_{333} = 0 \qquad (6)$$

Agora, considere essa tabela como um experimento fatorial. Os efeitos fatoriais de tratamentos podem ser definidos pelos contrastes mostrados na tabela 4. Os contrastes escolhidos não são mutuamente ortogonais dentro de um efeito específico; entretanto, os efeitos fatoriais são mutuamente ortogonais. Nota-se que o contraste na forma de logarítmo de π_{ijk} definido por (4) é o primeiro contraste no conjunto de contrastes para a interação CD e o (5) é o primeiro contraste no conjunto de contrastes para a interação CF e que (6) é primeiro contraste no conjunto de contrastes para a interação CDF .

Relações similares a esta para outros testes de independência marginal são resumidas na tabela 5.

A tabela 4 facilita a obtenção das matrizes K adequadas para hipóteses de independência, não interação e ajuste de modelos em tabelas 3x3x3.

TABELA 4 - Contrastes para um experimento fatorial 3x3x3.

EFEITO	C_1									C_2									C_3								
	D_1			D_2			D_3			D_1			D_2			D_3			D_1			D_2			D_3		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
	π_{111}	π_{112}	π_{113}	π_{121}	π_{122}	π_{123}	π_{131}	π_{132}	π_{133}	π_{211}	π_{212}	π_{213}	π_{221}	π_{222}	π_{223}	π_{331}	π_{332}	π_{333}	π_{311}	π_{312}	π_{313}	π_{321}	π_{322}	π_{323}	π_{331}	π_{332}	π_{333}
C	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0									
	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+									
D	+	+	+	0	0	0				+	+	+	0	0	0				+	+	+	0	0	0			
	0	0	0	+	+	+				0	0	0	+	+					0	0	0	+	+	+			
F	+	0		+	0		+	0		+	0		+	0		+	0		+	0		+	0		+	0	
	0	+		0	+		0	+		0	+		0	+		0	+		0	+		0	+		0	+	
CD	+	+	+	0	0	0				0	0	0	0	0	0	0	0	0				0	0	0	+	+	+
	0	0	0	+	+	+				0	0	0	0	0	0	0	0	0	0	0	0				+	+	+
	0	0	0	0	0	0	0	0	0	+	+	+	0	0	0							0	0	0	+	+	+
	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+				0	0	0				+	+	+
CF	+	0		+	0		+	0		0	0	0	0	0	0	0	0	0		0	+		0	+		0	+
	0	+		0	+		0	+		0	0	0	0	0	0	0	0	0	0		+	0		+	0		+
	0	0	0	0	0	0	0	0	0	+	0		+	0		+	0			0	+		0	+		0	+
	0	0	0	0	0	0	0	0	0	0	+		0	+		0	+		0		+	0		+	0		+
DF	+	0		0	0	0		0	+	+	0		0	0	0		0	+		0		0	0	0		0	
	0	+		0	0	0	0			0	+		0	0	0	0		+	0			0	0	0	0		+
	0	0	0	+	0			0	+	0	0	0	+	0			0	+	0	0	0	+	0			0	+
	0	0	0	0	+		0		+	0	0	0	0	+		0		+	0	0	0	0	+		0		+
CDF	+	0		0	0	0		0	+	0	0	0	0	0	0	0	0	0		0	+	0	0	0	+	0	
	0	+		0	0	0	0		+	0	0	0	0	0	0	0	0	0	0		+	0	0	0	0	+	
	0	0	0	+	0			0	+	0	0	0	0	0	0	0	0	0	0	0	0		0	+	+	0	
	0	0	0	0	+		0		+	0	0	0	0	0	0	0	0	0	0	0	0	0		+	0	+	
	0	0	0	0	0	0	0	0	0	+	0		0	0	0		0	+		0	+	0	0	0	+	0	
	0	0	0	0	0	0	0	0	0	0	+		0	0	0	0		+	0		+	0	0	0	0	+	
	0	0	0	0	0	0	0	0	0	0	0	0	+	0			0	+	0	0	0		0	+	+	0	
	0	0	0	0	0	0	0	0	0	0	0	0	0	+		0		+	0	0	0	0		+	0	+	

TABELA 5 - Relação entre contrastes lineares na forma de $\ln \pi_{ijk}$ e testes para independência marginal para uma tabela de contingência 3x3x3, com apenas o total fixado.

EFEITO	Nº DE CONT.	HIPÓTESES $\pi_{ijk} =$						
		$\pi_{i..}\pi_{.j.}\pi_{..k}$	$\pi_{ij.}\pi_{i.k}$	$\pi_{i.k}\pi_{.jk}$	$\pi_{ij.}\pi_{.jk}$	$\pi_{ij.}\pi_{..k}$	$\pi_{i.k}\pi_{.j.}$	$\pi_{.jk}\pi_{i..}$
C	2	N	N	N	N	N	N	N
D	2	N	N	N	N	N	N	N
F	2	N	N	N	N	N	N	N
CD	4	Y	N	Y	N	N	Y	Y
CF	4	Y	N	N	Y	Y	N	Y
DF	4	Y	Y	N	N	Y	Y	N
CDF	8	Y	Y	Y	Y	Y	Y	Y

N indica que a hipótese especificada de independência não requer que os contrastes lineares associados na forma de $\ln \pi_{ijk}$ sejam iguais a zero.

Y indica que a hipótese especificada requer que os contrastes lineares associados na forma de $\ln \pi_{ijk}$ sejam iguais a zero.

A tabela 5 torna claro que qualquer hipótese de independência marginal é rejeitada se a interação de 2ª ordem (CDF) for significativa. Todavia, a rejeição de todas as hipóteses de independência marginal deixa a possibilidade de que uma estrutura mais simples do que a admitida multinomial pode existir para a tabela; é o caso, por exemplo, se a intera ção de 2ª ordem der não significativa. Assim os dados podem ser adequadamente representados por menos do que os cxdxf-1 parâmetros da distribuição multinomial.

Considera-se agora, a relação entre os contrastes da tabela 4 na forma de $\ln \pi_{ijk}$ e modelos lineares da forma

$$\ell_{ijk} = u + \alpha_i + \beta_j + \gamma_k \qquad (7)$$

onde

$$\ell_{ijk} = \ln \pi_{ijk} \; .$$

Usando o Lema 1 de KATRI (1966) pode-se estabelecer que a soma de quadrados residual do ajustamento (7) é idêntico à soma de quadrados para testar $K \ln \underset{\sim}{\pi} = \emptyset$, onde K é uma matriz consistindo de 1's, 0's e -1's correspondentes às interações CD, CF, DF e CDF da tabela de contrastes. Com K nessa forma, a hipótese $K \ln \underset{\sim}{\pi} = 0$ corresponde à hipótese $\pi_{ijk} = a_i b_j c_k$, ou seja, a hipótese de independência mútua completa.

Assim na terminologia usual para modelos lineares, o resíduo do ajustamento do modelo

$$\ell = X \xi$$

onde

$$\underset{\sim}{\ell}' = \left[\ell_{111}, \ell_{112}, \ldots, \ell_{cdf}\right] \qquad (8)$$

$$\ell_{ijk} = \ln \pi_{ijk}$$

X : matriz do delineamento

ξ : vetor dos parâmetros do modelo,

e a soma de quadrados para testar a hipótese $Kl = 0$, onde $K X = 0$, são idênticas.

Essas hipóteses de independência e de não interação podem ser formuladas em termos de hipóteses do tipo $F(\pi) = K \ln A\pi = 0$, onde K é uma matriz de constantes (t x cdf), obtida da tabela de contrastes, de acordo com o que se quer testar. Por exemplo, para testar-se a hipótese de que C e D independem de F , a matriz K corresponde aos contrastes de 1's, o's e -1's formados pelas linhas das interações CF, DF e CDF; para se testar a independência condicional de C e D dado F, a matriz K corresponde aos contrastes de 1's, o's e -1's formados pelas linhas das interações CD e CDF, e ainda, para se testar a hipótese de não interação tripla a matriz K corresponde aos contrastes de 1's, 0's e -1's formados pelas linhas da interação CDF; A é uma matriz identidade (cdf x cdf), e π é o vetor das probabilidades das caselas (cdf x 1).

Em tabelas de classificação cruzada de "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 respostas 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.

A soma de quadrados para se testar essas hipóteses é dada por $F'S^{-1}F$, que possui distribuição assintótica de χ^2 com t graus de liberdade, sob H_o ; onde $F = K \ln Ap$ é o vetor estimado de $F(\pi)$ e S a matriz de variâncias e covariâncias de $\underset{\sim}{F}$, dada por $S = KD^{-1}AV(\underset{\sim}{p})A'D^{-1}K'$; onde $V(\underset{\sim}{p})$, como trata-se de uma única distribuição multinomial, não é mais uma matriz bloco diagonal, mas apenas uma matriz completa.

De outro modo, essas hipóteses de independência e de não interação podem ser investigadas ajustando-se modelos log-lineares aos dados de uma tabela de contingência, estimando-se os parâmetros dos mesmos, buscando o modelo que melhor explique as associações entre diversas variáveis em estudo.

Quando K exaurir os graus de liberdade da tabela, o estimador de $K\underset{\sim}{\iota}$ é $K\hat{\underset{\sim}{\iota}}$, onde $\hat{\underset{\sim}{\iota}}$ é o vetor estimado de $\underset{\sim}{\iota}$ para $\underset{\sim}{\pi} = \underset{\sim}{p}$. Esta formulação dá estimativas não ajustadas dos parâmetros do modelo (modelo completo como chama GOODMAN (1971b)). Estimativas ajustadas (quando um modelo com menos parâmetros é ajustado) geralmente têm menor variância do que as estimativas não ajustadas dos parâmetros.

O procedimento seguinte dá prontamente as estimativas ajustadas dos parâmetros. Suponha que o posto de K seja t e que deseja-se produzir estimativas ajustadas para $t' < t$ dos efeitos definidos por K. Usa-se o modelo

$$\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \underset{\sim}{\iota} = \begin{bmatrix} I \\ \emptyset \end{bmatrix} \underset{\sim}{\xi}$$

onde K_1 é uma matriz $t' \times$ c.d.f. de posto linha completo t'; I é uma matriz identidade $t' \times t'$; $\emptyset$ é uma matriz de zeros $(t - t') \times t'$; $\underset{\sim}{\iota}$ é definido como anteriormente e $\underset{\sim}{\xi}$ é o vetor dos parâmetros do modelo.

Por métodos de modelos lineares, o estimador de $\underset{\sim}{\xi}$ é

$$\hat{\underset{\sim}{\xi}} = K_1 \hat{\underset{\sim}{l}} - S_{12} S_{22}^{-1} K_2 \hat{\underset{\sim}{l}}$$

com matriz de variâncias e covariâncias estimada,

$$V_{\hat{\underset{\sim}{\xi}}} = S_{11} - S_{12} S_{22}^{-1} S'_{12}$$

onde

$$S_{11} = K_1 D_1^{-1} V(\underset{\sim}{p}) D_1^{-1} K'_1 \quad ; \quad S_{12} = K_1 D_1^{-1} V(\underset{\sim}{p}) D_2^{-1} K'_2$$

$$S_{22} = K_2 D_2^{-1} V(\underset{\sim}{p}) D_2^{-1} K'_2 \quad ,$$

onde D_1^{-1} e D_2^{-1} são partições correspondentes da matriz D^{-1} .

Neste trabalho, ajusta-se o modelo

$$\begin{bmatrix} \underset{18X27}{K_1} \\ \\ \underset{6X27}{K_2} \end{bmatrix} \underset{27X1}{\underset{\sim}{l}} = \begin{bmatrix} \underset{18X18}{I} \\ \\ \underset{8X18}{\emptyset} \end{bmatrix} \underset{18X1}{\xi}$$

onde K_1 são as primeiras 18 linhas da tabela 4, K_2 são as 8 últimas linhas, $\underset{\sim}{l}$ é como definido em (8), e $\underset{\sim}{\xi}$ são os efeitos dos parâmetros. A soma de quadrados residual desse modelo dá o teste de não interação tripla; através do teste

$$S.Q. = \underset{\sim}{F}' V_{\underset{\sim}{F}}^{-1} \underset{\sim}{F} - \hat{\underset{\sim}{\xi}}' V_{\hat{\underset{\sim}{\xi}}}^{-1} \hat{\underset{\sim}{\xi}}$$

o qual tem distribuição assintótica de χ^2 (central) com t-t'= 26-18 = 8 graus de liberdade, dado que o modelo seja verdadeiro.

Se a interação de 2ª ordem não for significativa, escolhendo-se

$$C_1 = \begin{bmatrix} \underset{4X6}{\emptyset} & : & \underset{4X4}{I} & : & \underset{4X8}{\emptyset} \end{bmatrix} ; \quad C_2 = \begin{bmatrix} \underset{4X10}{\emptyset} & : & \underset{4X4}{I} & : & \underset{4X4}{\emptyset} \end{bmatrix} ;$$

$$C_3 = \begin{bmatrix} \underset{4X14}{\emptyset} & : & \underset{4X4}{I} \end{bmatrix}$$

na hipótese geral da forma $H_o: C\underset{\sim}{\xi} = \underset{\sim}{0}$, obtém-se os testes de hipóteses para não interação de CxD, CxF e DxF respectivamente, através da seguinte soma de quadrados

$$S.Q.\left[C\underset{\sim}{\xi} = \underset{\sim}{0}\right] = \hat{\underset{\sim}{\xi}}'C' \left[C\, V_{\hat{\underset{\sim}{\xi}}}\, C'\right]^{-1} C\, \hat{\underset{\sim}{\xi}}$$

a qual tem distribuição assintótica de χ^2 com d graus de liberdade, onde d é o posto de C, se H_o for verdadeira.

Desta maneira pode-se dizer qual é o modelo adequado, na forma de $\ln \pi_{ijk}$, para os dados contidos em uma tabela de contingência. Esses testes podem também ser interpretados em termos de independência verificando a tabela 5.

3.1.3. Aplicações em tabelas incompletas

Tabelas que têm algumas das caselas obrigatoriamente nulas, devido à perda de dados, respostas incorretas ou incompletas, ou pela própria natureza dos dados, podem ser investigadas com as técnicas discutidas anteriormente com a restrição de que certos contrastes de interações não podem ser estimados. Como um exemplo considera a tabela incompleta relatada por GRIZZLE e WILLIAMS (1972).

TABELA 6 - Números observados de adolescentes classificados por sexo, idade e no que concerne a saúde.

CONDIÇÃO DE SAÚDE	IDADE	MASCULINO	FEMININO
Problemas de	12-15	4	9
Reprodução	16-17	2	7
Problemas	12-15	-	4
Menstruais	16-17	-	8
Saúde	12-15	42	19
Ótima	16-17	7	10
Não tem	12-15	57	71
Nada	16-17	20	31

Se essa tabela estivesse completa existiriam 15 graus de liberdade associados a ela, mas como existem duas caselas com probabilidades iguais a zero, devido a machos não terem problemas com menstruação, o número de graus de liberdade passa a ser 13. Desse modo, as duas caselas com probabilidades zero causam a perda de 1 g.l. para a interação SxC e 1 g.l. para a interação SxIxC.

Para efetuar-se a análise consideram-se os dados mostrados na tabela 6 como uma amostra de distribuição multinomial com 14 categorias de respostas.

3.1.4. Exemplo numérico

A tabela 7 foi obtida por simulação respeitando a hipótese $\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$, $i = 1, 2$; $j = 1, 2$ e $K = 1, 2, 3$ e 4, com o tamanho da amostra fixado em $n = 1.000$. Assim as classificações A, B e C são mutuamente independentes, por construção.

TABELA 7 - Frequências observadas das caselas para a hipótese $\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$.

		C_1	C_2	C_3	C_4	TOTAL
A_1	B_1	25	144	52	180	401
	B_2	3	35	10	50	98
A_2	B_1	17	153	50	184	404
	B_2	4	32	11	50	97
TOTAL		49	364	123	464	1.000

A matriz K dos contrastes é dada por:

$$
\underset{15\times 16}{K} =
\begin{array}{c}
\\
A \\ B \\ \\ C \\ \\ AB \\ \\ AC \\ \\ \\ BC \\ \\ \\ ABC \\ \\
\end{array}
\begin{array}{c}
\begin{array}{cccccccccccccccc}
\pi_{111} & \pi_{112} & \pi_{113} & \pi_{114} & \pi_{121} & \pi_{122} & \pi_{123} & \pi_{124} & \pi_{211} & \pi_{212} & \pi_{213} & \pi_{214} & \pi_{221} & \pi_{222} & \pi_{223} & \pi_{224}
\end{array} \\
\left[\begin{array}{cccccccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \\
0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \\
0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 \\
0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 \\
0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 \\
1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \\
0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \\
0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\
1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & -1 \\
0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 \\
0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1
\end{array}\right]
\end{array}
$$

As matrizes C's para os testes das interações AB, AC e BC são dadas por

$$C_1 = \begin{bmatrix} \underset{1\tilde{X}5}{\emptyset} & : & \underset{1X1}{I} & : & \underset{1\tilde{X}6}{\emptyset} \end{bmatrix} \quad ; \quad C_2 = \begin{bmatrix} \underset{3X6}{\emptyset} & : & \underset{3X3}{I} & : & \underset{3X3}{\emptyset} \end{bmatrix} \quad ;$$

$$C_3 = \begin{bmatrix} \underset{3X9}{\emptyset} & : & \underset{3X3}{I} \end{bmatrix}$$

respectivamente.

Ajusta-se o modelo,

$$\begin{bmatrix} \underset{12X16}{K_1} \\ \\ \underset{3X16}{K_2} \end{bmatrix} \underset{16\tilde{X}1}{\underset{\sim}{l}} = \begin{bmatrix} \underset{12X12}{I} \\ \\ \underset{3X12}{\emptyset} \end{bmatrix} \underset{12\tilde{X}1}{\underset{\sim}{\xi}}$$

onde K_1 são as primeiras 12 linhas da matriz K ; K_2 são as últimas 3 linhas da matriz K ; $\underset{\sim}{l}$ é o vetor dos logarítmos naturais de π_{ijk} ; $\underset{\sim}{\xi}$ é o vetor dos parâmetros do modelo.

O resíduo do ajustamento desse modelo incompleto corresponde ao teste para a interação tripla.

Aplicando-se a metodologia GSK, obtém-se,

$$\underset{16\times1}{\underset{\sim}{p}} = \begin{bmatrix} 0,025 \\ 0,144 \\ 0,052 \\ 0,180 \\ 0,003 \\ 0,035 \\ 0,010 \\ 0,050 \\ 0,017 \\ 0,153 \\ 0,050 \\ 0,184 \\ 0,004 \\ 0,032 \\ 0,011 \\ 0,050 \end{bmatrix} \qquad \underset{15\times1}{\underset{\sim}{F}} = K \ln \underset{\sim}{p} = \begin{bmatrix} 0,0250 \\ 0,1444 \\ 0,0520 \\ 0,1800 \\ 0,0030 \\ 0,0350 \\ 0,0100 \\ 0,0500 \\ 0,0170 \\ 0,1530 \\ 0,0500 \\ 0,1840 \\ 0,0440 \\ 0,3200 \\ 0,0110 \end{bmatrix}$$

$$\underset{16\times16}{V(p)} = \begin{bmatrix} 0{,}0000243 & -3{,}6.10^{-6} & -1{,}3.10^{-6} & -4{,}5.10^{-6} & -7{,}5.10^{-6} & -8{,}75.10^{-7} & -2{,}5.10^{-7} & -1{,}25.10^{-6} & -4{,}25.10^{-7} & -3{,}82.10^{-6} & -1{,}25.10^{-6} & -4{,}6.10^{-6} & -1{,}0.10^{-7} & -8{,}0.10^{-7} & -2{,}75.10^{-7} & -1{,}25.10^{-6} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1{,}25.10^{-6} & -7{,}2.10^{-6} & -2{,}6.10^{-6} & -9{,}0.10^{-6} & -1{,}5.10^{-7} & -1{,}75.10^{-6} & -5{,}0.10^{-7} & -2{,}5.10^{-6} & -8{,}5.10^{-7} & -7{,}65.10^{-6} & -2{,}5.10^{-6} & -9{,}2.10^{-7} & -2{,}1.10^{-7} & -1{,}6.10^{-6} & -5{,}5.10^{-7} & -4{,}75.10^{-5} \end{bmatrix}$$

$$\underset{15\times15}{S} = KD^{-1}V(\underline{p})D^{-1}K' = \begin{bmatrix} 1{,}0365 & -0{,}1088 & 0{,}0643 & -0{,}0023 & 0{,}0082 & -0{,}7115 & 0{,}6311 & 0{,}0223 & 0{,}1791 & -0{,}1022 & 0{,}0029 & -0{,}0099 & -0{,}4555 & -0{,}0173 & -0{,}1226 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -0{,}1226 & 0{,}0082 & 0{,}0001 & 0{,}0001 & -0{,}0097 & 0{,}1791 & -0{,}0290 & -0{,}0290 & -0{,}1806 & 0{,}0001 & 0{,}0001 & 0{,}0084 & 0{,}0509 & 0{,}0509 & 0{,}2811 \end{bmatrix}$$

$$\underset{3\times3}{S_{22}} = K_2 D_2^{-1} V(\underset{\sim}{p}) D_2^{-1} K_2' = \begin{bmatrix} 0,73314 & 0,05099 & 0,05099 \\ 0,05099 & 0,12429 & 0,05099 \\ 0,05099 & 0,05099 & 0,28113 \end{bmatrix}$$

$$\underset{12\times1}{K_1 \hat{\underset{\sim}{l}}} = K_1 \mathrm{Ln}\ \underset{\sim}{p} = \begin{bmatrix} 0,0250 \\ 0,1444 \\ 0,0520 \\ 0,1800 \\ 0,0030 \\ 0,0350 \\ 0,0100 \\ 0,0500 \\ 0,0170 \\ 0,1530 \\ 0,0500 \\ 0,1840 \end{bmatrix} \qquad \underset{3\times1}{K_2 \hat{\underset{\sim}{l}}} = K_2\ \ln\ \underset{\sim}{p} = \begin{bmatrix} 0,044 \\ 0,320 \\ 0,011 \end{bmatrix}$$

$$\underset{12\times 12}{V_{\tilde{\xi}}} = S_{11} - S_{12}S_{22}^{-1}S'_{12} = \begin{bmatrix} 0{,}7158 & -0{,}0655 & 0{,}0017 & -0{,}0031 & 0{,}0046 & -0{,}2631 & 0{,}3128 & 0{,}0136 & 0{,}1016 & -0{,}0625 & 0{,}0036 & -0{,}0067 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -0{,}0067 & 0{,}1789 & -0{,}0291 & -0{,}0289 & -0{,}1803 & 0{,}0034 & 0{,}0001 & 0{,}0001 & -0{,}0043 & 0{,}0511 & 0{,}0509 & 0{,}2808 \end{bmatrix}$$

$$\underset{12\times 1}{\hat{\tilde{\xi}}} = K_1\hat{\tilde{l}}_1 - S_{12}S_{22}^{-1}K_2\hat{\tilde{l}}_2 = \begin{bmatrix} 0{,}5681 \\ 12{,}2174 \\ -9{,}5906 \\ -1{,}2052 \\ -5{,}6614 \\ -0{,}0930 \\ 0{,}6140 \\ -0{,}0300 \\ 0{,}0691 \\ 0{,}9174 \\ 0{,}3914 \\ 0{,}5732 \end{bmatrix}$$

A soma de quadrados residual deste modelo é igual a

$$S.Q. = \underset{\sim}{F}'S^{-1}\underset{\sim}{F} - \hat{\underset{\sim}{\xi}}\, V_{\hat{\underset{\sim}{\xi}}}^{-1}\hat{\underset{\sim}{\xi}} = 1{,}039 \quad \text{n.s.}$$

portanto, o modelo sem a interação tripla ajusta-se bem aos dados.

Como a interação tripla não foi significativa, testam-se as interações duplas através da hipótese geral da forma $C\underset{\sim}{\xi} = \underset{\sim}{0}$.

Para as interações AB, AC e BC obtém-se,

$$S.Q.\left[C_1\underset{\sim}{\xi} = \underset{\sim}{0}\right] = \hat{\underset{\sim}{\xi}}'C_1'\left[C_1 V_{\hat{\underset{\sim}{\xi}}} C_1'\right]^{-1}C\,\hat{\underset{\sim}{\xi}} = 0{,}021149 \quad \text{n.s.}$$

$$S.Q.\left[C_2\underset{\sim}{\xi} = \underset{\sim}{0}\right] = 1{,}12711 \quad \text{n.s.}$$

$$S.Q.\left[C_3\underset{\sim}{\xi} = \underset{\sim}{0}\right] = 2{,}6035 \quad \text{n.s.}$$

respectivamente.

Assim, pode-se dizer que um modelo adequado para os dados da tabela 6, não precisa conter nenhuma das interações. Verificando-se a tabela 5, nota-se que a hipótese de independência mútua completa é satisfeita.

3.1.5. Modelos log-lineares em tabelas CxDxF e algumas considerações sobre eles

3.1.5.1. O modelo saturado ou completo

$$\log_e \pi_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} +$$

$$+ \mu_{23(jk)} + \mu_{123(ijk)} ,$$

com as restrições usuais que,

$$\mu_{1(\cdot)} = \mu_{2(.)} = \mu_{3(.)} = \ldots = \mu_{123(.jk)} = \mu_{123(i.k)} = \mu_{123(ij.)} = 0$$

Este modelo não tem interesse maior. Um motivo é que os erros das estimativas dos parâmetros são maiores.

3.1.5.2. O modelo de independência parcial

$$\log_e \pi_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} \quad .$$

Neste caso $\mu_{123(ijk)} = 0$, $\mu_{23(jk)} = 0$ e $\mu_{13(ik)} = 0$, ou seja, C e D juntamente são independentes de F.

Existem outros dois modelos de independência parcial.

3.1.5.3. O modelo de independência condicional

$$\log_e \pi_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} \quad .$$

Neste caso $\mu_{123(ijk)} = 0 \Rightarrow$ não existe interação de 2^a ordem, ou seja, a associação entre as variáveis 2 e 3 é a mesma para todos os níveis da variável 1, e $\mu_{23(jk)} = 0 \Rightarrow$ a associação entre as variáveis 2 e 3 é zero.

Se o modelo dado acima se ajusta aos dados de uma tabela, então existe uma estrutura de independência condicional do tipo $P(C|D) \times P(C|F)$.

Existem outros dois modelos de independência condicional.

3.1.5.4. O modelo de inexistência de interação de segunda ordem

$$\log_e \pi_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \; ,$$

portanto, $\mu_{123(ijk)} = 0$.

3.1.5.5. O modelo de independência mútua completa

$$\log_e \pi_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} \quad ;$$

logo $\pi_{12(ij)} = 0$, $\mu_{13(ik)} = 0$, $\mu_{23(jk)} = 0$ e $\mu_{123(ijk)} = 0$.

3.1.6. O programa (linguagem BASIC) para a análise dos dados

No exemplo numérico dado anteriormente, no que concerne aos aspéctos de cálculos, os mesmos foram realizados através do programa apresentado no Apêndice 2.

Este programa computacional foi desenvolvido para a análise de dados categorizados provindos de uma única distribuição multinomial, ou de uma única população, envolvendo funções lineares e log-lineares.

Quanto à utilização deste programa, primeiramente deve-se entrar com os valores das frequências observadas das caselas, deve-se tomar um certo cuidado para entrar com as frequências observadas de maneira correta.

Em seguida, aparece no vídeo qual o tipo de transformação desejada, deve-se digitar o número 1 ou 2 conforme deseja-se a transformação linear ou log-linear, respectivamente. No caso da transformação ser linear deve-se fornecer a matriz A_{uxrs}. No caso da transformação ser a log-linear o programa gera essa matriz de transformação, que é uma matriz identidade.

O programa também pede que se forneça a matriz dos contrastes dos efeitos fatoriais, K_{txu}, a qual deve ser fornecida no sentido das linhas. Mais adiante no programa deve-se fornecer a matriz $K_{1\,t'xu}$ e o número de linhas dessa matriz.

E, finalizando, deve-se entrar com as matrizes $C's_{dxv}$ adequadas para testar-se a hipótese geral do tipo $\underset{dxv}{C}\,\underset{vx1}{\tilde{\beta}} = \underset{dx1}{\tilde{0}}$.

3.2. Simulação

Os dados para estudo neste trabalho são obtidos com base em simulação. Esses dados possuem uma única distribuição multinomial com parâmetros

$$\underset{\sim}{\pi} = \left[\pi_{111}, \pi_{112}, \pi_{113}, \ldots, \pi_{331}, \pi_{332}, \pi_{333}\right]$$

e $$n_{...} = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} n_{ijk} ,$$

onde $\underset{\sim}{\pi}$ é o vetor das probabilidades esperadas das caselas e $n_{...}$ é o o número total de observações, fixado.

A seguir é feita uma descrição sucinta sobre o programa de simulação.

Esse programa presta-se para qualquer número de classificações (N no programa); o número de níveis pode ser qualquer (C(I) no programa).

São construídas tabelas (Apêndice 4), com valores de p_{ijk}, estimativas de máxima verossimilhança de π_{ijk}, as quais possuem por construção independência mútua completa, ou independência parcial, ou independência condicional das classificações C, D e F. Os valores das frequências relativas das caselas, p_{ijk}, correspondem ao vetor $\underset{\sim}{p}$, no programa é P(I), com c.d.f. componentes.

Pode-se gerar quantas tabelas se desejar (NT no programa), e com quaisquer tamanhos de amostras (NN no programa).

No programa, para este trabalho, é definido um vetor com 27 componentes, cada qual correspondendo à probabilidade acumulada de uma casela, de forma que o primeiro limite é dado pelo valor de p_{111}, o segundo limite é dado pela soma do primeiro limite com p_{112} e assim por diante.

A seguir, através de uma subrotina, geram-se valores uniformemente distribuídos, independentes, no intervalo zero a um, colocados de acordo com seus valores, nos componentes adequados do vetor e após, através de uma soma, são dadas as frequências da tabela de contingência cruzada.

É também usado o procedimento de BERKSON (1955), que consiste em substituir um valor zero por 1/r, onde r é o número de categorias de respostas.

O programa (linguagem BASIC) para simulação de dados multinomiais encontra-se no Apêndice 1.

Com os valores simulados das frequências observadas das caselas é feita, através da metodologia GSK, a análise dos dados, que consiste em estimar os parâmetros do modelo, ajustar um modelo não contendo a interação de segunda ordem e, se esta não for significativa, testar as interações de 1ª ordem, pois em tabelas CxDxF, todos os modelos que descrevem hipóteses de interesse são hierárquicos. Posteriormente verificando-se a tabela 5 determinar uma estrutura de independência adequada para os dados, quando couber.

Neste trabalho são geradas tabelas com diferentes hipóteses de independência e tamanhos de amostra, como descritas abaixo, estiman em correspondência o número de decisões erradas, ou seja, o número de tabelas as quais os testes estatísticos não concluem em favor da estrutura de independência estabelecida para elas, de acordo com a tabela 5.

As hipóteses de independência estabelecidas para as tabelas são as seguintes:

$P(C \cap D \cap F) = P(C)P(D)P(F)$, isto é, independência mútua completa das três classificações C, D e F.

$P[(C \cap F) \cap D)] = P(C \cap F)P(D)$, isto é, as classificações C e F juntamente são independentes de D.

$P[(C \cap D) \cap F] = P(C \cap D)P(F)$, isto é, as classificações C e D juntamente são independentes de F.

$P(C \cap D|F) = P(C|F)P(D|F)$, isto é, as classificações C e D são independentes dada a classificação F.

$P(C \cap F|D) = P(C|D)P(F|D)$, isto é, as classificações C e F são independentes dada a classificação D.

Para cada hipótese de independência citada acima são simuladas tabelas com tamanhos de amostras iguais a 200, 250, 300, 1000 e 1500 observações.

Neste trabalho são analisadas 100 tabelas para cada combinação: estrutura de independência e tamanho de amostra. Na verdade, haveria a necessidade de se gerar um número bem maior de tabelas mas que devido a limitações, no que diz respeito ao computador utilizado, isto torna-se inviável.

Um comentário sobre o modo como são tomadas as decisões (corretas ou incorretas) quanto às estruturas de independência estabelecidas para as tabelas é no sentido de que são tomadas como decisões como decisões incorretas somente quando os Y's (tabela 5) não forem verificados, ou seja, quando os contrastes na forma de ln π_{ijk} não forem iguais a zero a um nível de significância igual a 5%.

Quanto aos N's (tabela 5) são interpretados do seguinte modo: os contrastes na forma de ln π_{ijk} podem ser iguais a zero ou diferentes de zero, a um nível α de significância igual a 5% . Desse modo, os mesmos não concorrem para a aceitação ou rejeição da estrutura de independência estabelecida para a tabela.

Para exemplificar, observa-se na tabela 5 que, para ocorrer independência parcial $\pi_{ijk} = \pi_{ij.}\pi_{..k}$ os contrastes das interações CF, DF e CDF devem ser todos iguais a zero (Y), e os contrastes da interação CD podem ou não serem iguais a zero (N), a um nível de significância α ; se ocorrer de os contrastes da interação CD serem iguais a zero verifica-se (tabela 5) a condição de independência mútua completa, o que não implica que deve-se rejeitar a estrutura de independência parcial estabelecida para a tabela. Do mesmo modo, para ocorrer independência condicional do tipo $\pi_{ijk} = \pi_{i.k}\pi_{.jk}$ os contrastes das interações CD e CDF devem ser iguais a zero (Y), se ocorrer de os contrastes da interação CF serem iguais a zero verifica-se (tabela 5) que a hipótese de independência parcial

$\pi_{ijk} = \pi_{.jk}\pi_{i..}$ fica satisfeita, o que não concorre para a rejeição da hipótese $\pi_{ijk} = \pi_{i.k}\pi_{.jk}$, do mesmo modo, se os contrastes da interação DF forem iguais a zero verifica-se (tabela 5) que a hipótese de independência parcial $\pi_{ijk} = \pi_{i.k}\pi_{.j.}$ fica satisfeita, o que não concorre para a rejeição da hipótese de independência condicional $\pi_{ijk} = \pi_{i.k}\pi_{.jk}$, e ainda, se ocorrer de os contrastes das interações CF e DF serem iguais a zero, verifica-se (tabela 5) que a hipótese de independência mútua completa é satisfeita, mas que também não concorre para a rejeição da hipótese de independência condicional $\pi_{ijk} = \pi_{i.k}\pi_{.jk}$, de acordo com a tabela 5.

Relações análogas às acima mencionadas, ocorrem para as demais hipóteses de independência.

4. RESULTADOS E DISCUSSÃO

Considerando-se o que foi exposto em (3.2.) sobre o modo de como são tomadas as decisões (corretas ou incorretas), apresentam-se na tabela 8 os resultados obtidos nos trabalhos de simulação de tabelas com determinadas estruturas de independência estabelecidas a priori e na análise dos dados obtidos das mesmas, com diferentes tamanhos de amostras.

TABELA 8 - Resultados obtidos dos testes estatísticos com dados simulados de diferentes estruturas de independência e diferentes tamanhos de amostras ($\alpha = 0{,}05$).

Independência estabelecida p/ a tabela	Tamanho da amostra	Nº de decisões incorretas	Nº de decisões corretas	Total de tabelas simuladas
$P(C \cap D \cap F) = P(C)P(D)P(F)$	200	4	96	100
"	250	4	96	100
"	300	15	85	100
"	1000	14	86	100
"	1500	5	95	100
$P[(C \cap F) \cap D] = P(C \cap F)P(D)$	200	5	95	100
"	200	1	99	100
"	250	3	97	100
"	300	6	94	100
"	1000	12	88	100
"	1500	4	96	100
$P[(C \cap D) \cap F] = P(C \cap D)P(F)$	200	1	99	100
"	250	0	100	100
"	250	5	95	100
"	300	9	91	100
"	1000	4	96	100
"	1500	0	100	100
$P(C \cap D \mid F) = P(C \mid F)P(D \mid F)$	200	2	98	100
"	250	0	100	100
"	300	4	96	100
"	300	11	89	100
"	300	2	98	100
"	1000	1	99	100
"	1000	8	92	100
"	1500	3	97	100
$P(C \cap F \mid D) = P(C \mid D)P(F \mid D)$	200	5	95	100
"	250	3	97	100
"	300	4	96	100
"	350	1	99	100
"	1000	4	96	100
"	1500	4	96	100

Tomando-se agora somente as decisões incorretas, foi verificado, para todos os casos, para quais modelos alternativos os testes indicaram. Os resultados são mostrados nas tabelas 9 e 10 a seguir.

TABELA 9 - Partição do número das decisões incorretas nos respectivos modelos alternativos admitidos.

ESTRUTURA ESTABELCIDA	TAMANHO DA AMOSTRA	MODELO ALTERNATIVO ADMITIDO							
		$P(C\cap F)P(D)$	$P(C\cap D)P(F)$	$P(D\cap F)P(C)$	$P(C\|F)P(D\|F)$	$P(C\|D)P(F\|D)$	$P(D\|C)P(F\|C)$	M.S.	N.I.T.
$P(A)P(B)P(C)$	200	2						2	
	250			4					
	300	6	2	4	1			2	
$P(C\cap F)P(D)$	200				4		1		
	250				2		1		
	300			1	2		1		
$P(C\cap D)P(F)$	200			1					
	250						1	4	
	300					3	3		
$P(C\|F)P(D\|F)$	200		1				1		
	250		ZERO DECISÕES INCORRETAS						
	300					4	1		8
$P(C\|D)P(F\|D)$	200						2	3	
	250				1			1	1
	300	1					1		2

onde, M.S. = Modelo Saturado ; N.I.T. = Não Interação Tripla.

TABELA 10 - Partição do número das decisões incorretas nos respectivos modelos alternativos.

ESTRUTURA ESTABELECIDA	TAMANHO DA AMOSTRA	MODELO ALTERNATIVO ADMITIDO							
		$P(C\cap F)P(D)$	$P(C\cap D)P(F)$	$P(D\cap F)P(C)$	$P(C\|F)P(D\|F)$	$P(C\|D)P(F\|D)$	$P(D\|C)P(F\|C)$	M.S.	N.I.T.
$P(A)P(B)P(C)$	1000	3	4	4					
	1500		5						
$P(C\cap F)P(D)$	1000				8		3	1	
	1500				2		2		
$P(C\cap D)P(F)$	1000						3	1	
	1500	ZERO DECISÕES INCORRETAS							
$P(C\|F)P(D\|F)$	1000							7	2
	1500							1	2
$P(C\|D)P(F\|D)$	1000							2	2
	1500								4

onde, M.S. = Modelo Saturado; N.I.T. = Não Interação Tripla.

Pelos resultados da tabela 8 observa-se que com tamanhos de amostras razoavelmente pequenos (N = 200, N = 250 e N = 300) os testes estatísticos mostraram-se conservadores em não rejeitar a estrutura de independência estabelecida para as tabelas, a um nível α de significância, igual a 5%, porém, deve-se aqui ressaltar, que dentro das decisões corretas existem muitos casos, em que os testes indicam um modelo diferente daquele pressuposto pela estrutura de independência estabelecida para os dados da tabela.

Para a condição de independência mútua completa não existe este tipo de problema, pois os contrastes das interações CD, CF, DF e CDF devem ser todos iguais a zero (Y na tabela 5), e o questionamento aqui levantado surge quando aparece algum N (tabela 5) para satisfazer a uma certa hipótese de independência.

A tabela 11 mostra os modelos alternativos e o número deles admitidos pelos dados das tabelas, mas que não são tomados como decisões incorretas de acordo com a tabela 5.

TABELA 11 - Número de modelos alternativos admitidos pelos dados das tabelas e que não rejeitam a estrutura de independência estabalecida.

ESTRUTURA ESTABELECIDA	TAMANHO DA AMOSTRA	MODELOS ALTERNATIVOS ADMITIDOS			
		$P(C)P(D)P(F)$	$P(D \cap F)P(D)$	$P(C \cap F)P(D)$	$P(C \cap D)P(F)$
$P(C \cap F)P(D)$	200	14			
	200	54			
	250	2			
	300	19			
$P(C \cap D)P(F)$	200	15			
	250	9			
	250	19			
	300	3			
$P(C \mid F)P(D \mid F)$	200	5	9	22	
	250	15	32	2	
	300	1	19	5	
	300	1	1	3	
	300		3	5	
$P(C \mid D)P(F \mid D)$	200	7	31		24
	250	4	21		16
	300	1	9		9

Considerando a estrutura de independência parcial $P\left[(C \cap F) \cap D\right] = P(C \cap F)P(D)$ e para um tamanho de amostra N = 200 observa-se na tabela 11 que em 100 tabelas, 54 tabelas possuem independência mútua completa dos três fatores, ou seja, C, D e F são independentes entre si, mas que de acordo com a tabela 5, estas tabelas satisfazem a estrutura de independência parcial dada acima.

Tomando-se agora a estrutura de independência condicional $P(C \cap D|F) = P(C|F)P(D|F)$ e para um tamanho de amostra N = 250 observa-se na tabela 11 que em 100 tabelas, 15 possuem independência mútua completa, 32 possuem independência parcial do tipo $P(D \cap F)P(C)$ e duas possuem independência parcial $P(C \cap F)P(D)$ e todas estas tabelas não são tomadas como decisões incorretas de acordo com a tabela 5; considerando ainda esta mesma hipótese de independência condicional para um tamanho de amostra N = 300 observa-se que em 100 tabelas, uma possui independência mútua completa, 19 possuem independência parcial do tipo $P(D \cap F)P(C)$ e 5 possuem independência parcial $P(C \cap F)P(D)$, porém não são tomadas como decisões incorretas de acordo com a tabela 5.

Deve-se fazer um raciocínio análogo para os demais casos.

Um aspécto também de grande importância verificado nos resultados obtidos, é que com amostras suficientemente grandes (N = 1000 e N = 1500) o exposto acima não foi verificado, pois onde aparecem os N's na tabela 5, os contrastes na forma de $\ln \pi_{ijk}$ dão todos significativos, ou seja, diferentes de zero, considerando este aspecto, para amostras suficientemente grandes, as decisões corretas mostradas na tabela 8 não contêm um modelo diferente daquele estabelecido pela estrutura de independência.

É neste aspecto que reside a importância de se necessitar trabalhar com amostras de tamanho suficientemente grandes.

Verificando-se as tabelas 9 e 10, pode-se observar uma tendência dos resultados dada da seguinte maneira:

- Quando a estrutura estabelecida para a tabela de contingência é a de independência mútua completa das três classificações C, D e F, e esta for rejeitada, existe um direcionamento dos resultados dos testes em indicar um modelo de independência parcial para os dados.

- Quando a estrutura estabelecida para a tabela de contingência é a de independência parcial das três classificações, e esta for rejeitada, existe um direcionamento dos resultados dos testes em indicar um modelo de independência condicional para os dados da tabela.

- Quando a estrutura estabelecida para a tabela de contingência é a de independência condicional das três classificações, e esta for rejeitada, existe um direcionamento dos resultados dos testes em indicar um modelo completo ou um modelo não contendo a interação de 2ª ordem.

Para a utilização do programa de simulação há a necessidade de entrar-se com uma tabela (Apêndice 4) a qual possui uma certa estrutura de independência estabelecida, e a partir desta foi feita a simulação das demais tabelas. Verificando-se a tabela 3 (Apêndice 4) pode-se verificar a ocorrência de 10 caselas com probabilidades iguais a 0,035 e 5 caselas com probabilidades iguais a 0,030. Isto ocorreu somente por coincidência de cálculos, esta situação concorre para dificultar a decisão dos tes

tes para um modelo especificado pela estrutura dos dados. Situação similar ocorre com as tabelas 5 e 6 (Apêndice 4).

Outra situação a comentar é com respeito às tabelas 8 à 14 (Apêndice 4). Estas tabelas possuem algumas caselas com probabilidades baixíssimas e em consequência causarão a ocorrência de caselas com frequências observadas iguais a zero, neste caso utiliza-se o procedimento de BERKSON (1955) que produz bons resultados dos testes estatísticos.

O programa para computador de análise de dados categorizados, através do método proposto por GRIZZLE, STARMER e KOCH (1969), apresentado neste trabalho, foi desenvolvido para a análise de dados provindos de uma única população. Uma complementação necessária é a de fazer a viabilização computacional aplicada a problemas envolvendo várias populações.

Outros métodos alternativos, para investigar questões de interesse em tabelas de contingência, incluem o método baseado em máxima verossimilhança como formulado por BISHOP (1969) e GOODMAN (1970, 1971), e o método baseado na teoria da informação, como descrito por KULLBACK, KU e VARNER (1971).

Todos estes métodos são baseados em estimadores BAN (best asymptotic normal) e, consequentemente são assintoticamente equivalentes. Entretanto, sob o aspecto computacional, o procedimento GSK é mais vantajoso, uma vez que o "software" que o método de mínimos quadrados ponderados requer é usualmente acessível, e em consequência, as estimativas e os testes estatísticos podem ser obtidos por métodos computacionais, comparáveis aos desenvolvidos para a análise de modelos lineares em geral.

5. CONCLUSÕES

A conclusão mais relevante deste trabalho consistiu na verificação prática do comportamento assintótico dos testes.

Conclui-se que com tamanhos pequenos de amostras (N = 200, N = 250 e N = 300), os resultados dos testes para indicar um modelo que represente a estrutura de independência dos dados não são satisfatórios.

Quando se adotam tamanhos de amostras maiores (N = 1000 e N = 1500), os resultados dos testes estatísticos obtidos mostram-se bastante eficientes, no sentido de indicar um modelo que represente a estrutura de independência dos dados das tabelas.

De acordo com os comentários dos próprios autores, sobre o desconhecimento do comportamento dos testes para tamanhos de amostras pequenos, espera-se com este trabalho, que se tenha obtido maiores esclarecimentos práticos sobre o problema apontado, reforçando-se aqui as ótimas propriedades assintóticas dos testes.

A recomendação sobre um tamanho ótimo de amostra não foi totalmente possível, porém obtém-se uma boa orientação sobre o mesmo. Pode-se observar que com $N = 1500$ os resultados obtidos, a um nível de significância α igual a 5%, são excelentes.

Considerando o comentário de GOODMAN (1971) e o desenvolvimento deste trabalho, conclui-se pela facilidade de aplicação do método, de um modo geral.

O método de simulação foi eficiente para poder-se verificar as propriedades assintóticas dos testes.

Pelos resultados obtidos, conclui-se que o procedimento de BERKSON (1955), utilizado neste trabalho, o qual é uma extensão do caso binomial para o multinomial, produziu bons resultados.

6. BIBLIOGRAFIA

BERKSON, J., 1955. Maximum likelihood and minimum χ^2 estimate of the logistic function. JASA. Boston, 50: 130-162.

BERKSON, J., 1968. Application of minimum logit χ^2 estimate to a problem of Grizzle, with a natation on the problem of no interaction. Biometrics. Raleigh, 24: 75-95.

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

BHAPKAR, V.P. e G.G. KOCH, 1968a. Hypotheses of 'no interaction' in multidimensional contingency tables. Technometrics. Washington, 10: 107-123.

BHAPKAR, V.P. e G.G. KOCH, 1968b. On the hypotheses of 'no interaction' in contingency tables. Biometrics. Raleigh, 24: 567-594.

BISHOP, Y.M.M., 1969. Full contingency tables, logits and split contingency tables. Biometrics. Raleigh, 16: 618-635.

BISHOP, Y.M.M. e S.E. FIENBERG, 1969. Incomplete two-dimensional contingency tables. Biometrics. Raleigh, 25: 119-128.

CANTON, A.W.P., 1980. Análise de dados categorizados. Atas do 4º SINAPE. Rio de Janeiro, 79 p.

CANTON, A.W.P. e M.N. FESTA, 1979. Análise de dados categorizados usando modelos lineares e suas aplicações - uma revisão bibliográfica. Atas do 3º SINAPE. São Paulo, 182-189.

CAUSSINUS, H., 1965. Contribution à l'analyse statistique des tableaux de corrélation. Annales de la Faculté des Sciences de l'Université de Toulouse. Toulouse, 29: 77-182.

CHI, G.Y.H.; G.G. KOCH; S. LUI CHI; W.M. STANISH; W.D. JOHNSON e J.R. LANDIS, 1978. A computer program for the generalized chi-square analysis of computing risks agrouped survival categorical data (CRISCAT). Computer programs in Biomedicinal, impress.

FESTA, M.N., 1979. Análise de dados de sobrevivência. Instituto de Matemática e Estatística - USP. (Dissertação de Mestrado)

FIEMBERG, S.A., s.d., Preliminary grafical analysis and local independence for two-way contingency tables. Unpublished manuscript.

FORTHOFER, R.N. e G.G. KOCH, 1973. An analysis for compounded functions of categorical data. Biometrics. Raleigh, 29: 143-157.

GOODMAN, L.A., 1968. The analysis of cross-classified data; independence, quasi-independence, and interaction in contingency tables with or without missing entries. JASA. Boston, 63: 1091-131.

GOODMAN, L.A., 1970. The multivariate analysis of qualitative data: interaction among multiple classifications. JASA. Boston, 65: 226-256.

GOODMAN, L.A., 1971a. Partitioning of chi-square analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. JASA. Boston, 66: 339-344.

GOODMAN, L.A., 1971b. The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics. Washington, 13: 33-63.

GRIZZLE, J.E., C.F. STARMER e G.G. KOCH, 1969. Analysis of categorical data by linear models. Biometrics. Raleigh, 25: 489-504.

GRIZZLE, J.E. e O.D. WILLIAMS, 1972. Log linear models and tests of independence for contingency tables. Biometrics. Raleigh, 28: 137-156.

HABER, M., 1983. Sample sizes for the exact test of 'no interaction in 2x2x2 tables. Biometrics. Raleigh, 39: 493-498.

JOHNSON, W.D. e G.G. KOCH, 1971. A note on the weighted least squares analysis of the Ries-Smith contingency table data. Technometrics. Washington, 13: 438-447.

KATRI, C.G., 1966. A note on a manova model applied to problems in growth curves. Ann. Inst. Statist. Math. Tokyo, 18: 75-86.

KOCH, G.G.; FREEMAN Jr., D.H. e H.D. TOLLEY, 1975. The asymptotic covariance structure of estimated parameters from contingency table log-linear models. Institute of Statistics Nimeo Series Nº 1046. University of North Carolina.

KOCH, G.G.; J.R. LANDIS; D.H. FREEMAN Jr., J.L. FREEMAN e R.G. LEHNEN. 1978. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics. Raleigh, 33: 133--134.

KULLBACK, S.; H.H. KU e R. VARNER, 1971. Analysis of multidimensional contingency tables. JASA. Boston, 66: 55-64.

LANDIS, J.R.; W.M. STANISH, J.L. FREEMAN e G.G. KOCH, 1976. A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). Computer Programs in Biomedicine, 6: 196-231.

NEYMAN, J., 1949. Contributions to the theory of the χ^2 test. Proc. Berkley Symp. Math. Statist. Prob. University of California Press. 239-272.

ODOROF, C.L., 1970. Minimum logit chi-square estimation and maximum likelihood estimation in contingency tables. JASA. Boston, 65: 1617-32.

PLACKETT, R.L., 1961. A note on interactions in contingency tables. J.R. Statist. Soc. Londres, B, 24: 162-166.

ROY, S.N. e M.A. KASTENBRAUM, 1956. On the hypotheses of no interaction in a multi-way contingency tables. Ann. Math. Statist. Baltimore, 27: 749-757.

WADA, C.Y., 1977. Análise de dados categóricos. Instituto de Matemática e Estatística - USP. (Dissertação de Mestrado)

WALD, A., 1943. Test of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. Lancaster, 54: 426-482.

WATSON, G.S., 1956. Missing and 'mixed-up' frequencies in contingency tables. Biometrics. Raleigh, 12: 47-50.

WILLIAMS, O.D. e J.E. GRIZZLE, 1972. Analysis of contingency tables having ordered response categories. JASA. Boston, 67: 55-63.

A P Ê N D I C E 1 : Programa para simulação de dados multinomiais.

```
10   REM "MULTINOMIAL"
20   SLOW
30   PRINT AT 8, 3; "SIMULAÇÃO DE VARIÁVEIS MULTINOMIAIS"
40   PAUSE 3ØØ
50   CLS
60   PRINT "NUMERO DE TABELAS = ";
70   INPUT NT
80   PRINT NT
90   PRINT "NUMERO DE VARIAVEIS =";
100  INPUT N
110  PRINT N
120  PRINT
130  DIM C(N)
140  FOR I = 1 TO N
150  PRINT "NUM. DE CATEG. VAR."; I; " = ;
160  INPUT C(I)
170  PRINT C(I)
180  NEXT I
190  PRINT
200  PRINT "NUM. DE DADOS =";
210  INPUT NN
220  PRINT NN
230  PRINT
240  PRINT "PROBABILIDADES"
250  PRINT "________________________________________________"
260  LET CC = 1
270  FOR I = 1  TO N
280  LET CC = C(I) * CC
290  NEXT I
300  DIM P(CC)
310  FOR I = 1 TO CC
```

```
320   PRINT "P("; I; ") = ";
330   INPUT P(I)
340   PRINT P(I)
350   NEXT I
360   PRINT "------------------------------------------------"
370   PAUSE 3ØØ
380   FAST
390   DIM L(CC)
400   LET L(1) = P(1)
410   FOR I = 2 TO CC
420   LET L(I) = L(I - 1) + P(I)
430   NEXT (I)
440   RAND Ø
450   FOR K = 1 TO NT
460   DIM F(CC)
470   FOR I = 1 TO NN
480   LET A = RND
490   FOR J = 1 TO CC
500   IF A > L(J) THEN GO TO 530
510   LET F(J) = F(J) + 1
520   GO TO 540
530   NEXT J
540   NEXT I
560   CLS
570   PRINT "------------------------------------------------"
580   FOR I = 1 TO CC
581   IF F(I) = 0 THEN LET F(I) = 1/CC
590   PRINT "F("; I; ") = "; F(I)
600   NEXT I
610   PRINT "------------------------------------------------"
620   STOP
630   NEXT K
640   STOP.
```

APÊNDICE 2: Programa para análise de dados categorizados pelo método GSK.

```
10   REM "ANCAT"
20   SLOW
30   PRINT "NUMERO DE CATEGORIAS =";
40   INPUT NC
50   PRINT NC
60   PRINT
70   DIM  F(NC)
80   DIM P(NC)
90   LET TF = Ø
100  FOR I = 1  TO NC
110  PRINT "F(";I;")=";
120  INPUT F(I)
130  PRINT F(I)
140  LET TF = TF + F(I)
150  NEXT I
160  PRINT "TIPO DE TRANSFORMAÇÃO. : 1-LINEAR   2-LOG-LINEAR"
170  INPUT A$
180  IF A$ = "1" THEN GOTO 460
190  IF A$ < > "2" THEN GOTO 170
200  PRINT "MATRIZ ""K"""
210  PRINT "--------------------------------------------------"
220  PRINT "NUM DE LINHAS = " ;
230  INPUT T
240  PRINT T
250  PRINT "NUM DE COLUNAS = " ;
260  INPUT U
270  PRINT U
280  PRINT "--------------------------------------------------"
290  DIM K(T, U)
300  FOR I = 1 TO T
310  FOR J = 1 TO U
320  PRINT "K(";I; " , ";J;")= " ;
330  INPUT K(I, J)
```

```
340   PRINT K(I, J)
350   NEXT J
360   PRINT
370   NEXT I
380   PRINT "________________________________________________"
390   DIM A(NC, NC)
400   FOR I = 1 TO NC
410   FOR J = 1 TO NC
420   LET A(I, J) = (1 AND I = J) + (Ø AND I < > J)
430   NEXT J
440   NEXT I
450   GOTO 640
460   PRINT "________________________________________________"
470   PRINT "MATRIZ ""A"""
480   PRINT "________________________________________________"
490   PRINT "NUM DE LINHAS = " ;
500   INPUT U
510   PRINT U
520   PRINT "NUM DE COLUNAS = " ; NC
530   PRINT "________________________________________________"
540   FAST
550   DIM A(U, NC)
560   FOR I = 1 TO U
570   FOR J = 1 TO NC
580   PRINT "A("; I; ", "; J; ") = " ;
590   INPUT A(I, J)
600   PRINT A(I, J)
610   NEXT J
620   NEXT I
630   PRINT "________________________________________________"
640   FAST
650   FOR I = 1 TO NC
660   LET P(I) = F(I)/TF
670   NEXT I
680   DIM V(NC, NC)
```

```
690  FOR I = 1 TO NC
700  FOR J = 1 TO NC
710  LET V(I, J) = ((P(I) * (1-P(I))/TF) AND I = J) + ((-P(I)*
     P(J)/TF) AND I < > J)
720  NEXT J
730  NEXT I
740  DIM F(U)
750  FOR I = 1 TO U
760  FOR J = 1 TO NC
770  LET F(I) = F(I) + A(I, J) * P(J)
780  NEXT J
790  NEXT I
800  IF A$ = "1" THEN GOTO 980
810  DIM D(U, U)
820  FOR I = 1 TO U
830  FOR J = 1 TO U
840  LET D(I, J) = (1/F(I) AND I = J) + (Ø AND I < > J)
850  NEXT J
860  NEXT I
870  DIM W(U)
880  FOR I = 1 TO U
890  LET W(I) = LN F(I)
900  NEXT I
910  DIM F(T)
920  FOR I = 1 TO T
930  FOR J = 1 TO U
940  LET F(I) = F(I) + K(I, J) * W(J)
950  NEXT J
960  NEXT I
970  DIM T(U, NC)
980  FOR I = 1 TO U
990  FOR J = 1 TO NC
```

```
1000  FOR K = 1 TO NC
1010  LET T(I, J) = T(I, J) + A(I, K) * V(K, J)
1020  NEXT K
1030  NEXT J
1040  NEXT I
1050  DIM S(U, U)
1060  FOR I = 1 TO U
1070  FOR J = 1 TO U
1080  FOR K = 1 TO NC
1090  LET S(I, J) = S(I, J) + T(I, K) * A(J, K)
1100  NEXT K
1110  NEXT J
1120  NEXT I
1130  IF A$ = "1" THEN GOTO 1380
1140  DIM T(T, U)
1150  FOR I = 1 TO T
1160  FOR J = 1 TO U
1170  FOR K = 1 TO U
1180  LET T(I, J) = T(I, J) + K(I, K) * D(K, J)
1190  NEXT K
1200  NEXT J
1210  NEXT I
1220  DIM H(T, U)
1230  FOR I = 1 TO T
1240  FOR J = 1 TO U
1250  FOR K = 1 TO U
1260  LET H(I, J) = H(I, J) + T(I, K) * S(K, J)
1270  NEXT K
1280  NEXT J
1290  NEXT I
1300  DIM S(T, T)
1310  FOR I = 1 TO T
```

```
1320   FOR J = 1 TO T
1330   FOR K = 1 TO U
1340   LET S(I, J) = S(I, J) + H(I, K) * T(J, K)
1350   NEXT K
1360   NEXT J
1370   NEXT I
1380   IF A$ = "1" : THEN LET T = Ø
1390   LET M = (U AND A$ = "1") + (T AND A$ = "2")
1400   DIM G(M, M)
1410   FOR I = 1 TO M
1420   FOR J = 1 TO M
1430   LET T(I, J) = S(I, J)
1440   NEXT J
1450   NEXT I
1460   GOSUB 1600
1470   DIM O(M)
1480   FOR I = 1 TO M
1490   FOR J = 1 TO M
1500   LET O(I) = O(I) + F(J) * S(J, I)
1510   NEXT J
1520   NEXT I
1530   LET QQ = Ø
1540   FOR I = 1 TO M
1550   LET QQ = QQ + O(I) * F(I)
1560   NEXT I
1570   CLS
1580   PRINT AT 1Ø,Ø ; "QUI-QUADRADO = " ; QQ
1590   GO TO 1770
1600   REM "INVERSA"
1610   FOR K = 1 TO M
1620   LET D = S(K, K)
```

```
1630  FOR J = 1 TO M
1640  LET S(K, J) = S(K, J)/D
1650  NEXT J
1660  FOR I = 1 TO M
1670  IF I = K THEN GO TO 1730
1680  LET B = S(I, K)
1690  FOR J = 1 TO M
1700  LET S(I, J) = S(I, J) - B * S(K, J)
1710  NEXT J
1720  LET S(I, K) = -B/D
1730  NEXT I
1740  LET S(K, K) = 1/D
1750  NEXT K
1760  RETURN
1770  STOP
1780  IF A$ = "1" THEN GOTO 3150
1790  CLS
1800  PRINT "NUMERO DE LINHAS DE K1 = " ;
1810  INPUT NK1
1820  PRINT NK1
1830  PRINT "NUMERO DE LINHAS DE K2 = " ;
1840  LET NK2 = T - NK1
1850  PRINT NK2
1860  LET M = NK2
1870  DIM S(M, M)
1880  LET L1 = Ø
1890  FOR I = NK1 + 1 TO T
1900  LET L1 = L1 + 1
1910  LET L2 = Ø
1920  FOR J = NK1 + 1 TO T
```

```
1930 LET L2 = L2 + 1
1940 LET S(L1, L2) = T(I, J)
1950 NEXT J
1960 NEXT I
1970 GOSUB 1600
1980 DIM H(NK1, NK2)
1990 FOR I = 1 TO NK1
2000 FOR J = 1 TO NK2
2010 LET L1 = Ø
2020 FOR K = NK1 + 1 TO T
2030 LET L1 = L1 + 1
2040 LET H(I, J) = H(I, J) + T(I, K) * S(L1, J)
2050  NEXT K
2060  NEXT J
2070  NEXT I
2080  DIM G(NK1, NK1)
2090  FOR I = 1 TO NK1
2100  FOR J = 1 TO NK1
2110  LET L1 = Ø
2120  FOR K = NK1 + 1 TO T
2130  LET L1 = L1 + 1
2140  LET G(I, J) = G(I, J) + H(I, L1) * T(K, J)
2150  NEXT K
2160  NEXT J
2170  NEXT I
2180  DIM S(NK1, NK1)
2190  FOR I = 1 TO NK1
2200  FOR J = 1 TO NK1
2210  LET S(I, J) = T(I, J) - G(I, J)
2220  NEXT J
2230  NEXT I
2240  DIM G(NK1)
2250  FOR I = 1 TO NK1
```

```
2260  LET L1 = Ø
2270  FOR J = NK1 + 1 TO T
2280  LET L1 = L1 + 1
2290  LET G(I) = G(I) + H(I, L1) * F(J)
2300  NEXT J
2310  NEXT I
2320  DIM E(NK1)
2330  FOR I = 1 TO NK1
2340  LET E(I) = F(I) - G(I)
2350  NEXT I
2360  DIM G(NK1, NK1)
2370  FOR I = 1 TO NK1
2380  FOR J = 1 TO NK1
2390  LET G(I, J) = S(I, J)
2400  NEXT J
2410  NEXT I
2420  CLS
2430  PRINT "ESTIMATIVAS DOS PARAMETROS"
2440  PRINT "--------------------------------------------------"
2450  FOR I = 1 TO NK1
2460  PRINT "E("; I; ") = " ; E(I)
2470  NEXT I
2480  PRINT " ================================================ "
2490  LET M = NK1
2500  GOSUB 1600
2510  LET SQP = Ø
2520  DIM Q(NK1)
2530  FOR I = 1 TO NK1
2540  FOR J = 1 TO NK1
2550  LET Q(I) = Q(I) + E(J) * S(J, I)
2560  NEXT J
2570  LET SQP =  SQP + Q(I) * E(I)
```

```
2580  NEXT I
2590  PRINT "S.Q.P = " ; SQP
2600  PRINT "SQR = " ; QQ - SQP
2610  STOP
2620  PRINT "DESEJA TESTAR HIPOTESES:
      CE = 0 ? "
2640  INPUT C$
2650  IF C$ = "N" THEN GOTO 3150
2660  PRINT "MATRIZ C"
2670  PRINT "----------------------------------------------------"
2680  PRINT "NUM DE LINHAS = " ;
2690  INPUT MC
2700  PRINT MC
2710  PRINT "NUM DE COL = " ; NK1
2720  PRINT "----------------------------------------------------"
2730  DIM C(MC, NK1)
2740  FOR I = 1 TO MC
2750  FOR J = 1 TO NK1
2760  PRINT "C("; I; ", "; J; ") = " ;
2770  INPUT C(I, J)
2780  PRINT C(I, J)
2790  NEXT J
2800  PRINT
2810  NEXT I
2820  DIM  H(MC)
2830  FOR I = 1 TO MC
2840  FOR J = 1 TO NK1
2850  LET H(I) = H(I) + C(I, J) * E(J)
2860  NEXT J
2870  NEXT I
2880  DIM Y(MC, NK1)
2890  FOR I = 1 TO MC
```

```
2900   FOR J = 1 TO NK1
2910   FOR K = 1 TO NK1
2920   LET Y(I, J) = Y(I, J) + C(I, K) * G(K, J)
2930   NEXT K
2940   NEXT J
2950   NEXT I
2960   DIM S(MC, MC)
2970   FOR I = 1 TO MC
2980   FOR J = 1 TO MC
2990   FOR K = 1 TO NK1
3000   LET S(I, J) = S(I, J) + Y(I, K) * C(J, K)
3010   NEXT K
3020   NEXT J
3030   NEXT I
3040   LET M = MC
3050   GOSUB 1600
3060   LET SQH = Ø
3070   FOR I = 1 TO MC
3080   FOR J = 1 TO MC
3090   LET SQH = SQH + H(I) * S(I, J) * H(J)
3100   NEXT J
3110   NEXT I
3120   CLS
3130   PRINT AT 10,3; "SQH = " ; SQH
3140   GOTO 2620
3150   CLS
3160   PRINT AT 10,14; "FIM"
```

APÊNDICE 3 : Sub-rotinas para simulação de dados multinomiais, inversão de matrizes e análise de dados categorizados pelo método GSK, linguagem FORTRAN.

```
PAGE   1

// JOB

LOG DRIVE   CART SPEC   CART AVAIL   PHY DRIVE
  0000        0002        0002         0000

V2 M11   ACTUAL 16K  CONFIG 16K

// FOR
*EXTENDED PRECISION
*ONE WORD INTEGERS
*LIST SOURCE PROGRAM
C PROGRAMA PARA SIMULACAO DE VARIAVEIS MULTINOMIAIS
  SUBROUTINE MULTI(P,A,NN,NA,NB,NIC,IX,F,N)
  DIMENSION P(27),SL(27),F(27),MF(3,3,3),NTA(3),NTB(3)
  DIMENSION NTAB(3,3),A(10),NTC(3)
  SL(1)=P(1)
  DO 30 I=2,NN
  J=I-1
  SL(I)=SL(J)+P(I)
   30 CONTINUE
      DO 660 I=1,NN
  660 F(I)=0.
      DO 40 JJ=1,N
      CALL RANDU(IX,IY,YFL)
      DO 50 I=1,NN
      IF(YFL-SL(I))60,60,50
   50 CONTINUE
   60 F(I)=F(I)+1
      IX=IY
   40 CONTINUE
      L=0
      DO 100 I=1,NA
      DO 100 J=1,NB
      DO 100 K=1,NIC
      L=L+1
      IF(F(L))20,25,20
   25 F(L)=1./NN
   20 NTA(I)=0.
      NTB(J)=0
      NTC(K)=0
      NTAB(I,J)=0
      MF(I,J,K)=F(L)
  100 CONTINUE
      DO 110 I=1,NA
      DO 120 J=1,NB
      DO 130 K=1,NIC
      NTA(I)=NTA(I)+MF(I,J,K)
      NTB(J)=NTB(J)+MF(I,J,K)
      NTC(K)=NTC(K)+MF(I,J,K)
      NTAB(I,J)=NTAB(I,J)+MF(I,J,K)
  130 CONTINUE
  120 CONTINUE
  110 CONTINUE
      WRITE(3,200)(A(I),I=1,10)
  200 FORMAT('1',//,1X,20X,10A6,/,100('-'))
      NCI=NIC-1
      WRITE(3,210)(I,I=1,NCI),I
  210 FORMAT(9X,10(7X,'C',I2))
      WRITE(3,230)
```

```
PAGE   2

  230 FORMAT(1X,100('-'))
      DO 300 I=1,NA
      DO 310 J=1,NB
      WRITE(3,400)I,J,(MF(I,J,K),K=1,NIC),NTAB(I,J)
  400 FORMAT(1X,'A',I2,2X,'B',I2,11(6X,I4))
  310 CONTINUE
      WRITE(3,230)
  300 CONTINUE
      WRITE(3,500)(NTC(I),I=1,NIC),N
  500 FORMAT(1X,'TOTAL',3X,11(6X,I4))
      WRITE(3,230)
      RETURN
      END

FEATURES SUPPORTED
 ONE WORD INTEGERS
 EXTENDED PRECISION

CORE REQUIREMENTS FOR MULTI
 COMMON       0  VARIABLES     142  PROGRAM     624

RELATIVE ENTRY POINT ADDRESS IS 0OCD (HEX)

END OF COMPILATION

// DUP

*DELETE             MULTI
CART ID 0002   DB ADDR  5A18   DB CNT   0028

*STORE       WS  UA  MULTI
CART ID 0002   DB ADDR  6000   DB CNT   0028
```

```
PAGE   1

// JOB

LOG DRIVE   CART SPEC   CART AVAIL  PHY DRIVE
  0000        0002        0002        0000

V2 M11  ACTUAL 16K  CONFIG 16K

// FOR
*EXTENDED PRECISION
*ONE WORD INTEGERS
*LIST SOURCE PROGRAM
      SUBROUTINE SWEP(S,M)
      DIMENSION S(27,27)
      DO 10 K=1,M
      D=S(K,K)
      DO 20 J=1,M
      S(K,J)=S(K,J)/D
   20 CONTINUE
      DO 50 I=1,M
      IF(I-K)40,50,40
   40 B=S(I,K)
      DO 60 J=1,M
      S(I,J)=S(I,J)-B*S(K,J)
   60 CONTINUE
      S(I,K)=-B/D
   50 CONTINUE
      S(K,K)=1./D
   10 CONTINUE
      RETURN
      END

FEATURES SUPPORTED
 ONE WORD INTEGERS
 EXTENDED PRECISION

CORE REQUIREMENTS FOR SWEP
 COMMON       0  VARIABLES      12  PROGRAM     168

RELATIVE ENTRY POINT ADDRESS IS 0010 (HEX)

END OF COMPILATION

// DUP

*DELETE              SWEP
CART ID 0002   DB ADDR  5A18   DB CNT   000C

*STORE      WS  UA  SWEP
CART ID 0002   DB ADDR  60E8   DB CNT   000C
```

```
PAGE    1

// JOB

LOG DRIVE   CART SPEC   CART AVAIL  PHY DRIVE
  0000        0002        0002        0000

V2 M11   ACTUAL 16K  CONFIG 16K

// FOR
*IOCS(CARD,1132PRINTER,DISK)
*EXTENDED PRECISION
*ONE WORD INTEGERS
*LIST SOURCE PROGRAM
      DIMENSION F(27),P(27),L(27,27),V(27,27),T(27,27),S(27,27),E(27)
      DIMENSION Q(27),C(3,4,27),H(4),Y(4,27),MC(3),A(10),PP(27)
C     'ANALISE DE DADOS CATEGORIZADOS'
      READ(2,10)NC,NH,NNT,N,NA,NB,NIC,IX
   10 FORMAT(8I4)
      READ(2,15)(A(I),I=1,10)
   15 FORMAT(10A6)
      READ(2,20)(P(I),I=1,NC)
   20 FORMAT(8F10.8)
      DO 16 I=1,NC
   16 PP(I)=P(I)
      READ(2,11)(MC(I),I=1,NH)
   11 FORMAT(3I3)
      READ(2,55)NT,NU,NK1,NK2
   55 FORMAT(4I2)
      DO 22 I=1,NH
      MCC=MC(I)
      DO 22 J=1,MCC
   22 READ(2,21)(C(I,J,K),K=1,NK1)
   21 FORMAT(18F2.0)
      DO 62 I=1,NT
   62 READ(2,65)(L(I,J),J=1,NU)
   65 FORMAT(27I2)
      DO 2000 INT=1,NNT
      CALL MULTI(PP,A,NC,NA,NB,NIC,IX,F,N)
      TF=0.
      DO 28 I=1,NC
      TF=TF+F(I)
   28 CONTINUE
      DO 105 I=1,NC
      P(I)=F(I)/TF
  105 CONTINUE
      DO 110 I=1,NC
      DO 115 J=1,NC
      IF(I-J)106,107,106
  107 V(I,J)=P(I)*(1-P(I))/TF
      GO TO 115
  106 V(I,J)=-P(I)*P(J)/TF
  115 CONTINUE
  110 CONTINUE
      DO 111 I=1,NC
      DO 112 J=1,NC
      T(I,J)=0.
  112 CONTINUE
  111 CONTINUE
      DO 165 I=1,NC
      DO 170 J=1,NC
```

```
PAGE    2

      T(I,J)=T(I,J)+V(I,J)/(P(I)*P(J))
  170 CONTINUE
  165 CONTINUE
      DO 150 I=1,NU
      P(I)=ALOG(P(I))
  150 CONTINUE
      DO 155 I=1,NT
      F(I)=0.
      DO 160 J=1,NU
      F(I)=F(I)+L(I,J)*P(J)
  160 CONTINUE
  155 CONTINUE
      DO 175 I=1,NT
      DO 180 J=1,NC
      S(I,J)=0.
      DO 185 K=1,NC
      S(I,J)=S(I,J)+L(I,K)*T(K,J)
  185 CONTINUE
  180 CONTINUE
  175 CONTINUE
      DO 190 I=1,NT
      DO 195 J=1,NT
      T(I,J)=0.
      DO 200 K=1,NC
      T(I,J)=T(I,J)+S(I,K)*L(J,K)
  200 CONTINUE
  195 CONTINUE
  190 CONTINUE
      DO 205 I=1,NT
      DO 208 J=1,NT
      S(I,J)=T(I,J)
  208 CONTINUE
  205 CONTINUE
      CALL SWEP(S,NT)
      QQ=0.
      DO 210 I=1,NT
      DO 215 J=1,NT
      QQ=QQ+S(I,J)*F(I)*F(J)
  215 CONTINUE
  210 CONTINUE
      DO 220 I=1,NT
      DO 230 J=1,NT
      S(I,J)=0.
  230 CONTINUE
  220 CONTINUE
      N1=NK1+1
      L1=0.
      DO 240 I=N1,NT
      L1=L1+1
      L2=0.
      DO 250 J=N1,NT
      L2=L2+1
      S(L1,L2)=T(I,J)
  250 CONTINUE
  240 CONTINUE
      CALL SWEP(S,NK2)
      DO 260 I=1,NK1
      DO 270 J=1,NK2
      V(I,J)=0.
```

```
PAGE   3

      L1=0.
      DO 280 K=N1,NT
      L1=L1+1
      V(I,J)=V(I,J)+T(I,K)*S(L1,J)
  280 CONTINUE
  270 CONTINUE
  260 CONTINUE
      DO 290 I=1,NK1
      DO 300 J=1,NK1
      S(I,J)=0.
      L1=0.
      DO 310 K=N1,NT
      L1=L1+1
      S(I,J)=S(I,J)+V(I,L1)*T(K,J)
  310 CONTINUE
  300 CONTINUE
  290 CONTINUE
      DO 319 I=1,NK1
      DO 320 J=1,NK1
      S(I,J)=T(I,J)-S(I,J)
  320 CONTINUE
  319 CONTINUE
      DO 330 I=1,NK1
      L1=0.
      F(I)=0.
      DO 340 J=N1,NT
      L1=L1+1
      F(I)=F(I)+V(I,L1)*F(J)
  340 CONTINUE
  330 CONTINUE
      DO 350 I=1,NK1
      F(I)=F(I)-F(I)
  350 CONTINUE
      DO 360 I=1,NK1
      DO 370 J=1,NK1
      V(I,J)=0.
      V(I,J)=S(I,J)
  370 CONTINUE
  360 CONTINUE
      WRITE(3,380)
  380 FORMAT(10X,'ESTIMATIVAS DOS PARAMETROS')
      DO 390 I=1,NK1
      WRITE(3,400)(E(I),I=1,NK1)
  400 FORMAT(10F12.4)
  390 CONTINUE
      CALL SWEP(S,NK1)
      SOP=0.
      DO 410 I=1,NK1
      O(I)=0.
      DO 420 J=1,NK1
      O(I)=O(I)+E(J)*S(J,I)
  420 CONTINUE
      SOP=SOP+O(I)*E(I)
  410 CONTINUE
      SOR=OO-SOP
      IF(NH)450,460,450
  450 DO 1000 IH=1,NH
      YN=YC(IH)
      DO 500 I=1,MM

PAGE   4

      H(I)=0.
      DO 510 J=1,NK1
      H(I)=H(I)+C(IH,I,J)*E(J)
  510 CONTINUE
  500 CONTINUE
      DO 520 I=1,MM
      DO 530 J=1,NK1
      Y(I,J)=0.
      DO 540 K=1,NK1
      Y(I,J)=Y(I,J)+C(IH,I,K)*V(K,J)
  540 CONTINUE
  530 CONTINUE
  520 CONTINUE
      DO 550 I=1,MM
      DO 560 J=1,MM
      S(I,J)=0.
      DO 570 K=1,NK1
      S(I,J)=S(I,J)+Y(I,K)*C(IH,J,K)
  570 CONTINUE
  560 CONTINUE
  550 CONTINUE
      CALL SWEP(S,MM)
      SQH=0.
      DO 580 I=1,MM
      DO 590 J=1,MM
      SQH=SQH+H(I)*S(I,J)*H(J)
  590 CONTINUE
  580 CONTINUE
      WRITE(3,600)IH,SQH
  600 FORMAT(10X,'SOMA DE QUADRADOS DA HIPOTESE H',I2,'=',F12.3)
 1000 CONTINUE
  460 WRITE(3,700)SOP,SOR,OO
  700 FORMAT(10X,'SOMA DE QUAD. DE PARAMETROS=',F10.3,/,10X,'SOMA DE QUA
     *DRADOS DO RESIDUO=',F10.3,/,10X,'QUI-QUADRADO TOTAL=',F12.3)
 2000 CONTINUE
      STOP
      END

FEATURES SUPPORTED
 ONE WORD INTEGERS
 EXTENDED PRECISION
 IOCS

CORE REQUIREMENTS FOR
 COMMON        0  VARIABLES    9080  PROGRAM    1976

END OF COMPILATION

// XEQ
```

APÊNDICE 4: Tabelas com as probabilidades esperadas das caselas para diferentes estruturas de independência das classificações C, D e F.

TABELA 1 - Independência mútua completa: $\pi_{ijk} = \pi_{i..}\ \pi_{.j.}\ \pi_{..k}$

		F_1	F_2	F_3	TOTAL
	D_1	0,0393750	0,0413437	0,0505262	0,1312449
C_1	D_2	0,0393750	0,0413437	0,0505262	0,1312449
	D_3	0,0337500	0,0354375	0,0433125	0,1125000
	D_1	0,0315000	0,0330750	0,0404250	0,1050000
C_2	D_2	0,0315000	0,0330750	0,0404250	0,1050000
	D_3	0,0270000	0,0283500	0,0346500	0,0900000
	D_1	0,0341250	0,0358312	0,0437927	0,1137489
C_3	D_2	0,0341250	0,0358312	0,0437926	0,1137488
	D_3	0,0292500	0,0307250	0,0375375	0,0975125
	TOTAL	0,3000000	0,3150123	0,3849877	1,0000000

TABELA 2 - Independência parcial: $\pi_{ijk} = \pi_{ij.}\ \pi_{..k}$

		F_1	F_2	F_3	TOTAL
	D_1	0,030000	0,031500	0,038500	0,100000
C_1	D_2	0,045000	0,047250	0,057750	0,150000
	D_3	0,037500	0,039375	0,048125	0,125000
	D_1	0,015000	0,015750	0,019250	0,050000
C_2	D_2	0,030000	0,031500	0,038500	0,100000
	D_3	0,045000	0,047250	0,057750	0,150000
	D_1	0,045000	0,047250	0,057750	0,150000
C_3	D_2	0,019500	0,020475	0,025025	0,065000
	D_3	0,033000	0,034650	0,042350	0,110000
	TOTAL	0,300000	0,315000	0,385000	1,000000

TABELA 3 - Independência parcial: $\pi_{ijk} = \pi_{i.k}\pi_{.j.}$

		F_1	F_2	F_3	TOTAL
C_1	D_1	0,03500	0,05250	0,04375	0,13125
	D_2	0,03500	0,05250	0,04375	0,12125
	D_3	0,03000	0,04500	0,03750	0,11250
C_2	D_1	0,03500	0,03500	0,03500	0,10500
	D_2	0,03500	0,03500	0,03500	0,10500
	D_3	0,03000	0,03000	0,03000	0,09000
C_3	D_1	0,03500	0,02275	0,05600	0,11375
	D_2	0,03500	0,02275	0,05600	0,11375
	D_3	0,03000	0,01950	0,04800	0,09750
TOTAL		0,30000	0,31500	0,38500	1,00000

TABELA 4 - Independência parcial: $\pi_{ijk} = \pi_{.jk}\pi_{i..}$

		F_1	F_2	F_3	TOTAL
C_1	D_1	0,037500	0,037500	0,056250	0,131250
	D_2	0,056250	0,037500	0,037500	0,131250
	D_3	0,018750	0,043125	0,050625	0,112500
C_2	D_1	0,030000	0,030000	0,045000	0,105000
	D_2	0,045000	0,030000	0,030000	0,105000
	D	0,015000	0,034500	0,040500	0,090000
C_3	D	0,032500	0,032500	0,048750	0,113750
	D_2	0,048750	0,032500	0,032500	0,113750
	D_3	0,016250	0,037375	0,043875	0,097500
TOTAL		0,30000	0,315000	0,385000	1,000000

TABELA 5 - Independência condicional: $\pi_{ijk} = \pi_{i.k}\pi_{.jk}/\pi_{..k}$

		F_1	F_2	F_3	TOTAL
	D_1	0,0333335	0,0476190	0,0487012	0,1296537
C_1	D_2	0,0500000	0,0476190	0,0324675	0,1300865
	D_3	0,0166668	0,0547619	0,0438311	0,1152598
	D_1	0,0333335	0,0317460	0,0389610	0,1040405
C_2	D_2	0,0500000	0,0317460	0,0259740	0,1077200
	D_3	0,0166667	0,0365079	0,0350649	0,0882395
	D_1	0,0333335	0,0206349	0,0623376	0,1163060
C_3	D_2	0,0500000	0,0206349	0,0415584	0,1121933
	D_3	0,0166668	0,0237301	0,0561038	0,0965007
TOTAL		0,3000008	0,3149997	0,3849995	1,0000000

TABELA 6 - Independência condicional: $\pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$

		F_1	F_2	F_3	TOTAL
	D_1	0,0428571	0,0428571	0,0642858	0,1499999
C_1	D_2	0,0428571	0,0285714	0,0285715	0,0999999
	D_3	0,0208333	0,0479167	0,0562500	0,1250000
	D_1	0,0285714	0,0285714	0,0428572	0,0999999
C_2	D_2	0,0642850	0,0428571	0,0428571	0,1499999
	D_3	0,0083333	0,0191667	0,0225000	0,0500000
	D_1	0,0285714	0,0285714	0,0428572	0,0999999
C_3	D_2	0,0428571	0,0285714	0,0285715	0,0999999
	D_3	0,0208333	0,0479167	0,0562500	0,1250000
TOTAL		0,2999998	0,3149999	0,3850003	1,0000000

TABELA 7 - Independência condicional: $\pi_{ijk} = \pi_{ij.}\pi_{i.k}/\pi_{i..}$

		F_1	F_2	F_3	TOTAL
C_1	D_1	0,0400000	0,0600000	0,0500000	0,1500000
	D_2	0,0266667	0,0400000	0,0333333	0,1000000
	D_3	0,0333333	0,0500000	0,0416667	0,1250000
C_2	D_1	0,0333333	0,0333334	0,0333333	0,1000000
	D_2	0,0500000	0,0500000	0,0500000	0,1500000
	D_3	0,0166667	0,0166667	0,0166666	0,0500000
C_3	D_1	0,0307692	0,0200000	0,0492308	0,1000000
	D_2	0,0307692	0,0200000	0,0492308	0,1000000
	D_3	0,0384615	0,0250000	0,0615385	0,1250000
TOTAL		0,2999999	0,3150001	0,3850000	1,0000000

TABELA 8 - Independência mútua completa: $\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$

		F_1	F_2	F_3	TOTAL
C_1	D_1	0,02520	0,05040	0,12600	0,12600
	D_2	0,03105	0,06210	0,06210	0,15525
	D_3	0,033750	0,06750	0,06750	0,16875
C_2	D_1	0,02520	0,05040	0,05040	0,12600
	D_2	0,03105	0,06210	0,06210	0,15525
	D_3	0,03375	0,06750	0,06750	0,16875
C_3	D_1	0,00560	0,01120	0,01120	0,02800
	D_2	0,00690	0,01380	0,01380	0,03450
	D_3	0,00750	0,01500	0,01500	0,03750
TOTAL		0,20000	0,40000	0,40000	1,00000

TABELA 9 - Independência parcial: $\pi_{ijk} = \pi_{ij.}\pi_{..k}$

		F_1	F_2	F_3	TOTAL
C_1	D_1	0,016	0,032	0,032	0,080
	D_2	0,024	0,048	0,048	0,120
	D_3	0,050	0,100	0,100	0,250
C_2	D_1	0,020	0,060	0,060	0,140
	D_2	0,050	0,080	0,080	0,210
	D_3	0,020	0,040	0,040	0,100
C_3	D_1	0,010	0,020	0,020	0,050
	D_2	0,005	0,010	0,010	0,025
	D_3	0,005	0,010	0,010	0,025
TOTAL		0,200	0,400	0,400	1,000

TABELA 10 - Independência parcial: $\pi_{ijk} = \pi_{i.k}\pi_{.j.}$

		F_1	F_2	F_3	TOTAL
C_1	D_1	0,04200	0,04200	0,04200	0,12600
	D_2	0,05175	0,05175	0,05175	0,15525
	D_3	0,05625	0,05625	0,05625	0,16875
C_2	D_1	0,00840	0,05600	0,06160	0,12600
	D_2	0,01035	0,06900	0,07590	0,15525
	D_3	0,01125	0,07500	0,08250	0,16875
C_3	D_1	0,00560	0,01400	0,00840	0,02800
	D_2	0,00690	0,01725	0,01035	0,03450
	D_3	0,00750	0,01875	0,01125	0,03750
TOTAL		0,20000	0,40000	0,40000	1,00000

TABELA 11 - Independência parcial: $\pi_{ijk} = \pi_{.jk}\pi_{i..}$

		F_1	F_2	F_3	TOTAL
	D_1	0,04500	0,03600	0,04500	0,12600
C_1	D_2	0,02250	0,05400	0,07875	0,15525
	D_3	0,02250	0,09000	0,05625	0,16875
	D_1	0,04500	0,03600	0,04500	0,12600
C_2	D_2	0,02250	0,05400	0,07875	0,15525
	D_3	0,02250	0,09000	0,05625	0,16875
	D_1	0,01000	0,00800	0,01000	0,02800
C_3	D_2	0,00500	0,01200	0,1750	0,03450
	D_3	0,00500	0,02000	0,01250	0,03750
TOTAL		0,20000	0,40000	0,40000	1,00000

TABELA 12 - Independência condicional: $\pi_{ijk} = \pi_{i.k}\pi_{.jk}/\pi_{..k}$

		F_1	F_2	F_3	TOTAL
	D_1	0,075000	0,030000	0,037500	0,142500
C_1	D_2	0,037500	0,045000	0,065625	0,148125
	D_3	0,037500	0,075000	0,046875	0,159375
	D_1	0,015000	0,040000	0,055000	0,110000
C_2	D_2	0,007500	0,060000	0,096250	0,163750
	D_3	0,007500	0,100000	0,068750	0,176250
	D_1	0,010000	0,010000	0,007500	0,027500
C_3	D_2	0,005000	0,015000	0,013125	0,033125
	D_3	0,005000	0,025000	0,009375	0,039375
TOTAL		0,200000	0,400000	0,400000	1,000000

TABELA 13 - Independência condicional: $\pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$

		F_1	F_2	F_3	TOTAL
	D_1	0,0285714	0,0228572	0,0285714	0,0800000
C_1	D_2	0,0173913	0,0417392	0,0608696	0,1200001
	D_3	0,0333333	0,1333333	0,0833333	0,2499999
	D_1	0,0535714	0,0428575	0,0535714	0,1500003
C_2	D_2	0,0289855	0,0695650	0,1014493	0,1999998
	D_3	0,0133333	0,0533333	0,0333333	0,0999999
	D_1	0,0178571	0,0142857	0,0178571	0,0499999
C_3	D_2	0,0036232	0,0086957	0,0126813	0,0250002
	D_3	0,0033333	0,0133333	0,0083333	0,0249999
TOTAL		0,1999998	0,4000002	0,4000000	1,0000000

TABELA 14 - Independência condicional: $\pi_{ijk} = \pi_{ij.}\pi_{i.k}/\pi_{i..}$

		F_1	F_2	F_3	TOTAL
	D_1	0,026667	0,026667	0,026667	0,080001
C_1	D_2	0,040000	0,040000	0,040000	0,120000
	D_3	0,083333	0,083333	0,083333	0,249999
	D_1	0,010000	0,066667	0,073333	0,150000
C_2	D_2	0,013333	0,088889	0,097778	0,200000
	D_3	0,006667	0,044444	0,048889	0,100000
	D_1	0,010000	0,025000	0,015000	0,050000
C_3	D_2	0,005000	0,012500	0,007500	0,025000
	D_3	0,005000	0,012500	0,007500	0,025000
TOTAL		0,200000	0,400000	0,400000	1,000000