MODELOS LINEARES PARA DADOS
QUALITATIVOS: REVISÃO
BIBLIOGRÁFICA E APLICAÇÕES EM
EXPERIMENTOS AGRONÔMICOS

RUI VIEIRA DE MORAES

ORIENTADOR: VIVALDO FRANCISCO DA CRUZ

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
OUTUBRO, 1988

A meus pais:

João (in memoriam)

Lázara

DEDICO

À minha esposa,
Elisadete,

e aos meus filhos:

Daniel,
Fernando,
Carolina,

OFEREÇO

AGRADECIMENTOS

Ao Dr. VIVALDO FRANCISCO DA CRUZ, Professor Assistente Doutor do Departamento de Matemática e Estatística da ESALQ-USP, pela orientação valiosa na realização deste trabalho.

Aos Professores do Departamento de Matemática e Estatística da ESALQ-USP, pelos ensinamentos, apoio e incentivo oferecidos.

Ao DEPARTAMENTO DE BIOESTATÍSTICA DO INSTITUTO DE BIOCÍÊNCIAS - UNESP, Campus de Botucatu, pela oportunidade oferecida para realização do curso.

À CAPES/PICD, pela bolsa concedida para a realização do programa de Mestrado.

Ao Dr. CARLOS ROBERTO PADOVANI, Professor Assistente Doutor do Departamento de Bioestatística do IB-UNESP, pela leitura do manuscrito e sugestões dispensadas.

À Dr^a MARTHA MARIA MISCHAN, Professora Assistente Doutora do Departamento de Bioestatística do IB-UNESP, pelas sugestões despendidas.

A Dr^a SHEILA ZAMBELLO DE PINHO, Professora Assistente Doutora do Departamento de Bioestatística do IB-UNESP, pelas sugestões despendidas.

A SECRETARIA DE AGRICULTURA DO ESTADO DE SÃO PAULO, pela cessão dos dados.

Ao Dr. ARY APPARECIDO SALIBE, Professor Titular do Departamento de Horticultura da FCA-UNESP, pela cessão dos dados.

Aos Colegas do Departamento de Bioestatística do IB-UNESP, pelo apoio, incentivo e compreensão.

Às Funcionárias do Departamento de Bioestatística do IB-UNESP, Elza Ivone de Almeida Mota, Laura Aparecida Lopes, Célia Maria Nogueira da Silva e Maria Betina Pereira, pela colaboração prestada.

Aos Funcionários do Pólo Computacional do Campus de Botucatu -Rubião Júnior, pela colaboração prestada.

A todos aqueles que, de uma forma ou outra, contribuíram para a realização deste trabalho.

I N D I C E

	Página
RESUMO	vii
SUMMARY	viii
1. INTRODUÇÃO	01
2. REVISÃO BIBLIOGRÁFICA	03
3. MATERIAL E MÉTODO	19
3.1 Material	19
3.2 Método	21
3.2.1 Introdução	21
3.2.2 Estimação dos parâmetros π_{ij}	24
3.2.3 Variância do estimador de π_{ij}	26
3.2.4 Caracterização das funções paramétricas de interesse	
3.2.5 Ajuste do modelo e testes de hipóteses lineares	31
4. RESULTADOS E DISCUSSÃO	33
4.1 Análise de dados categorizados para uma única população, baseada na produção de pares de la- ranjeiras	33

4.1.1 Tabela de contingência sem casela nula .	33
4.1.2 Tabela de contingência com casela nula .	36
4.2 Análise de dados categorizados para várias po- pulações, baseado no rendimento médio de cana de-açúcar	37
4.2.1 Tabela de contingência sem casela nula .	38
4.2.2 Tabela de contingência com casela nula .	42
5. CONCLUSÕES	43
6. REFERÊNCIAS BIBLIOGRÁFICAS	45
7. APÊNDICE : Programas para análise de dados catego- rizados pelo método GSK	52
7.1 Programa UNIPOP.BAS	53
7.2 Programa SEVPOP.BAS	59

MODELOS LINEARES PARA DADOS QUALITATIVOS:
REVISÃO BIBLIOGRÁFICA E APLICAÇÕES EM
EXPERIMENTOS AGRONÔMICOS

Autor: Rui Vieira de Moraes

Orientador: Prof. Dr. Vivaldo Francisco da Cruz

RESUMO

Em algumas situações dentro da Estatística Experimental, a variável em estudo é qualitativa e suas mensurações podem ser apresentadas na forma de uma tabela de contingência (sxr). Nesses casos o estudo estatístico dessa variável pode ser feito através de modelos lineares para dados qualitativos, utilizando a metodologia proposta por GRIZZLE, STARMER e KOCH (1969), também conhecida como método GSK.

Em nosso trabalho, discutiremos o artigo básico dessa metodologia, que será apresentado em capítulo especial, bem como os mais recentes trabalhos nessa área, a partir dos quais se programará, em linguagem BASIC, o método GSK. Finalizaremos nosso trabalho, apresentando exemplos aplicados a agronomia, relativos à produção de laranja e rendimento médio agrícola em kg/ha de cana-de-açúcar.

LINEAR MODELS FOR QUANTITATIVE DATA:
BIBLIOGRAPHICAL REVIEW AND APPLICATIONS
TO AGRONOMIC EXPERIMENTS

Author: Rui Vieira de Moraes

Adviser: Prof. Dr. Vivaldo Francisco da Cruz

SUMMARY:

In some situations in the Experimental Statistic, field the variable in study is qualitative and its measures can be presented like a contingency table ($s \times r$). In such cases, a statistic study of this variable can be made through linear models for qualitative data using the method of GRIZZLE, STARMER and KOCH (1969), well known as GSK method.

In this work, we will discuss the basic point of this methodology, that will be presented in a special chapter, as well as the update papers in this area, that will be used to program the GSK method in BASIC language. Finally, we will present the some examples in data concerning from orange production and agricultural average production in kg/ha of sugar cane.

1 INTRODUÇÃO

Na Estatística Experimental, existem situações em que, para atingir o objetivo do pesquisador, precisamos considerar variáveis qualitativas. Nos casos em que os resultados amostrais podem ser apresentados em tabela de contingência ($s \times r$), uma forma interessante de analisar esses dados categorizados é através de modelos lineares.

Nesse sentido, considerando as distribuições multinomiais, cada uma com r categorias de respostas, GRIZZLE, STARMER e KOCH (1969) desenvolveram uma metodologia baseada em modelos lineares, ajustados através dos mínimos quadrados ponderados, denominado GSK.

Em face da importância dessas situações, dentro do campo da Estatística Aplicada e da grande flexibilidade do uso dessa técnica GSK, objetivou-se, neste trabalho, reunir e apresentar os principais artigos relativos a esta metodologia, dando-se ênfase especial ao artigo original.

Usando-se os resultados desse levantamento

bibliográfico, fez-se o programa computacional em linguagem BASIC, para microcomputadores da linha "Personal Computer" com o mínimo de 256 Kbytes de memória central, para a utilização dos problemas referidos inicialmente.

Finalmente, com o objetivo de ilustrar a metodologia e evidenciar o manuseio dos programas computacionais, duas aplicações da área agronômica foram consideradas. A primeira, para o caso de uma única população, na qual se utilizam dados relativos à produção de laranja, e a segunda, para casos que envolvem várias populações, onde se utilizam dados relativos à produtividade de cana-de-açúcar.

2. REVISÃO BIBLIOGRÁFICA

Nesta revisão, será feito um levantamento cronológico dos principais artigos, envolvendo aplicações de análise de dados categorizados, usando modelos lineares.

BERKSON (1968) comparou as estimativas dos quiquadrados de máxima verossimilhança, quiquadrado (logit) mínimo e quiquadrado de Pearson para diversos problemas, salientando que os resultados são numericamente iguais. Embora não tenha sido divulgado que o quiquadrado (logit) mínimo é só um membro da família cujo teste estatístico, em grandes amostras, tem distribuição de quiquadrado, se a hipótese nula é verdadeira; e a exploração desse fato leva a uma aproximação conceitual e computacional para a análise de dados categorizados.

Esta aproximação tem vantagens porque dá ao pesquisador maior liberdade na escolha do modelo e testes de hipóteses que são precisamente determinados para um dado específico, possibilitando, assim, a elaboração de um programa computacional para a análise de dados categorizados, compa-

rável, na maioria das vezes, à análise de modelos lineares. Para Berkson (1968) o método GSK representa uma alternativa para o processo de Lewis (1968), o qual se baseava em estimativa de máxima verossimilhança.

GRIZZLE, STARMER e KOCH (1969) consideram s amostras casualizadas e independentes, de tamanhos $n_1, n_2, \dots, n_s$, extraídas de distribuições multinomiais; cada uma apresentando por hipótese, r categorias de respostas. Definimos então um conjunto de u funções de probabilidades para π_{ij} ; $i=1,2,\dots,s$; $j=1,2,\dots,r$ e $\sum_{j=1}^r \pi_{ij} = 1$, que possuem derivadas parciais contínuas até segunda ordem. Um estimador consistente da matriz de variância-covariância desse conjunto é dado por $V_F = H V(P) H'$, onde H é a matriz das derivadas parciais de primeira ordem das funções consideradas.

Um processo geral não iterativo é descrito para ajustar essas funções a um modelo linear, testar a qualidade de ajustamento desse modelo e as hipóteses concernentes aos parâmetros do modelo linear.

Os casos particulares das funções lineares e das funções logarítmicas dos π_{ij} são desenvolvidos de maneira detalhada e apresentam exemplos que utilizam vários tipos de dados qualitativos.

JOHNSON e KOCH (1971) aplicaram o método GSK nos dados de Ries-Smith, referentes a um experimento,

comparando dois tipos de detergentes. Selecionou-se uma amostra casualizada de $n=1008$ indivíduos para os quais foram fornecidos os produtos X e M, com o propósito de verificar se o uso prévio de M influenciava o uso de X, bem como a temperatura e a qualidade da água. Depois de um dado período, retornou-se aos consumidores e determinaram-se as respostas para as proposições em estudo.

Os autores tinham por objetivo obter um modelo que explicasse a maior parte das variações, estatisticamente significantes entre as frequências observadas, com o menor número de parâmetros. Mostram que qualquer tipo de análise pode ser obtido, usando-se o método de regressão com dados contínuos e similarmente por tabela de contingência, através do uso do método GSK.

KOCH e REINFURT (1971) tratam das tabelas de contingência de forma análoga à análise de variância para modelos mistos. No experimento, temos n parcelas, a cada um dos d níveis de um determinado fator classificados em uma das r categorias. Os dados resultantes são representados em tabela de contingência $r \times r \times \dots \times r$ de dimensão d . A hipótese de interesse é a igualdade das distribuições marginais. Alternativamente, se r categorias podem ser colocadas numa escala quantitativa, a análise está voltada no sentido de verificar a igualdade dos "scores" médios, das d marginais de primeira ordem. Os testes estatísticos são desenvolvidos baseados no

χ^2 mínimo de Neyman ou na análise dos quadrados mínimos ponderados o qual apresenta grande semelhança com o procedimento T^2 de Hotelling, usado em modelos mistos, com dados contínuos.

GRIZZLE e WILLIAMS (1972) apresentam métodos estatísticos para testar a independência em tabelas de contingência multivariada, baseados na correspondência entre análise de experimentos fatoriais e testes de independência das marginais. Essas relações simplificam a interpretação das análises em tabelas de contingência e simplificam os testes para tabelas que contêm probabilidade zero. Transformações log-lineares do tipo $F(P) = K \ln P$ são utilizadas para a análise de independência completa, parcial ou condicional de marginais.

KOCH et alii (1972a) estendem o método GSK em situações onde: i) a falta de dados para certos indivíduos surge aleatoriamente como ausência de resposta ou resposta errada; ii) amostras suplementares pertencentes a diversos subconjuntos de variáveis obtidas devido ao especial interesse nessas variáveis. Observam que os problemas discutidos são distintos dos que envolvem tabela de contingência incompleta contendo, "a priori", caselas vazias.

A extensão é apresentada através de uma série de exemplos, que mostram que a aproximação pode ser usada

para uma grande variedade de dados fora do padrão. Aplicações para dados categorizados de modelo linear misto e parcela subdividida são enfatizados.

WILLIAMS e GRIZZLE (1972) abordam um método de análise baseado nos postos, o qual utiliza informações contidas nas respostas ordenadas e outro método, baseado na análise GSK do modelo originariamente proposto por Thurstone (1959). Os dois métodos utilizam testes χ^2_1 mínimo modificado e os resultados das duas análises são apresentados na forma de análise de variância. A exemplificação dos métodos é feita com dados que apresentam respostas univariada e bivariada .

KOCH *et alii* (1972b) apresentam algoritmos matriciais que geram estimativas de taxas de sobrevivência em t anos para pacientes com doenças crônicas, na análise de dados categorizados.

Partindo de um exemplo teórico, estabelecidas as hipóteses de interesse, foram obtidas as estimativas do modelo linear pelo método dos quadrados mínimos ponderados.

Dando continuidade, descrevem detalhadamente, passo a passo, o uso de produto cruzado e estrutura hierárquica na formulação do modelo. Este método tem aplicação fundamental no grupamento de combinações de laudo clínico em grupos indicativos do estágio da doença. O uso seqüencial de produtos cruzados e estruturas hierárquicas

é descrito, passo a passo, em detalhes, úteis para a formulação de modelos eficientes.

FORTHOFER e KOCH (1973) estendem o método do χ^2_1 mínimo modificado para a combinação de funções logarítmicas, exponenciais e lineares. Essas funções compostas das proporções são aplicadas às seguintes situações práticas: 1) modelos de associação em tabelas de contingência quadradas, 2) coeficientes de correlação de postos, 3) "ridits", 4) associação parcial.

Tais situações podem ser analisadas, usando-se uma extensão apropriada do método GSK.

Segundo KOCH e TOLLEY (1975), um método bem conhecido de estimação da densidade de bactérias, numa solução amostral, é o procedimento conhecido como o número mais provável (MPN). Esse trabalho analisa a mudança na densidade, através do tempo, de populações de bactérias que se encontram em fase de desaparecimento. O MPN, com sua variância estimada para um tempo fixo, é o ponto básico dessa investigação e é abordado a partir de funções gerais das proporções na análise de dados categorizados por modelos lineares. Utilizam-se essas quantidades para o ajuste dos modelos exponenciais em função do tempo, pelo método dos mínimos quadrados ponderados. Quando o ajuste é significativo, é possível comparar as taxas de decrescimento das populações sob diferentes condições experimentais.

Ilustra-se o método com populações de *Leptospira autumnalis*.

KOCH *et alii* (1976) apresentam situações onde as unidades experimentais são conglomerados de elementos, e cada um deles classificados de acordo com a presença ou não de um atributo individual (resposta binária). Com esse propósito, utiliza-se o modelo probabilístico binomial agrupado, para definir um parâmetro de taxa média associado, em termos do qual os efeitos de certos tratamentos, bem como o tamanho dos conglomerados, podem ser pesquisados. A metodologia resultante envolve uma síntese de procedimentos de máxima verossimilhança e mínimos quadrados ponderados.

LANDIS *et alii* (1976) desenvolveram um programa computacional que implementa consideravelmente a metodologia geral para a análise multivariada de dados categorizados. O programa envolve essencialmente a construção de testes de hipóteses através de funções das proporções observadas, que estão diretamente relacionadas à pesquisa e às estimativas dos parâmetros do modelo. Qualquer função composta de proporções observadas que pode ser formulada como uma sequência das seguintes transformações de dados: linear, logarítmica, exponencial ou adição de constante, pode ser analisada dentro dessa estrutura geral.

A entrada dos dados pode ser através de tabela de contingência multidimensional; um vetor de funções

que estime a matriz de covariância, e por dados puros na forma de valor inteiro, associado com cada objeto.

IMREY *et alii* (1976) discutem experimentos que envolvem comparações emparelhadas no contexto de tabelas de contingência incompletas, considerando o método GSK. Estudam-se também esses dados, através do modelo de Bradley-Terry, e posteriormente, pela análise baseada na transformação logit. Embora o método esteja sujeito a restrições sobre o tamanho amostral, pode-se evitar problemas relativos a condições de simetria, suposições de independência e balanceamento dos dados.

KOCH *et alii* (1977) analisam dados categorizados multivariados, provenientes de experimentos repetidos. Uma discussão expositiva das hipóteses pertinentes a tais situações é feita, e testes estatísticos apropriados são desenvolvidos através da aplicação de mínimos quadrados ponderados. São feitas considerações especiais aos problemas computacionais associados à manipulação de tabelas de grande dimensão e aos relacionados com caselas de frequências nulas.

LANDIS e KOCH (1977a) apresentam uma metodologia geral para a análise de dados categorizados multivariados, provenientes de experimentos realizadas com o objetivo de verificar a concordância entre observadores. O procedimento envolve a construção de funções das proporções observadas, que traduzem a concordância entre os observadores e os testes

estatísticos para hipóteses associadas a essas funções. Testes de viés entre observadores são apresentados em termos de homogeneidade marginal de primeira ordem, e medidas de concordância entre os observadores são desenvolvidas, através de estatísticas generalizadas, do tipo Kappa.

LANDIS e KOCH (1977b), discutem uma metodologia geral para a análise de dados categorizados multivariados, provenientes de experimentos realizados com o objetivo de verificar a concordância entre dois observadores. Experimentos desse tipo originam tabelas de contingência de grande dimensão, apresentando, em consequência, um grande número de caselas com frequências nulas, acarretando procedimentos baseados nos indicadores de variáveis, associados a dados originais, para gerar as marginais de primeira ordem e as somas das diagonais da tabela teórica de contingência. Dessas quantidades, são geradas estimativas de modo que reflitam a extensão de uma decisão majoritária que represente o consenso entre os observadores, sobre a característica em julgamento.

STANISH *et alii* (1978) referem-se à análise estatística para o caso da ausência de dados para uma ou mais variáveis, dependentes na análise de dados categorizados multivariados. Este problema é analisado para comparar as diferenças entre as respostas de pacientes submetidos a dois tratamentos (droga vs placebo), obtidas em três períodos.

CLAUDIO SILVA (1979) apresenta a metodologia GSK em duas etapas: a) Dada uma tabela de contingência multidimensional, usam-se as proporções observadas nas diferentes caselas para a obtenção das funções paramétricas de interesse. Tais funções podem ser construídas mediante operações matriciais e/ou transformações logarítmicas e exponenciais. b) O ajuste do modelo é feito pelo método dos quadrados mínimos ponderados.

O autor mostra exemplos de aplicação da metodologia GSK em tabelas de taxa de mortalidade neo-natal, segundo a idade materna (Chile, 1972); acidentes com veículos, ocorridos na Carolina do Norte, de 1966 a 1972, levando em conta a gravidade do acidente, condições do tempo e modelo do automóvel. Estes exemplos foram desenvolvidos, utilizando o programa computacional GENCAT.

KOCH *et alii* (1982) aplicam três métodos de análise de covariância para dados categorizados em um exemplo clínico sobre artrite reumática. Os três métodos são: processo de casualização do modelo não paramétrico, regressão logística para máxima verossimilhança e análise de quadrados mínimos ponderados de funções marginais correlacionadas. As suposições e os resultados estatísticos de cada método são discutidos assim como se enfatizam diferenças entre suas origens. Observadas as apro-

ximações de cálculo, verifica-se que os métodos tendem para resultados similares em problemas análogos.

DOBSON (1983), apresenta um capítulo sobre a análise de dados em que as respostas são categorizadas complementando com três exemplos numéricos representando modelos diferentes. No primeiro exemplo os dados referem-se a um estudo de cancer de pele, no qual quatrocentos pacientes são avaliados quanto a localização do tumor e o tipo histológico. A questão de interesse é verificar se existe associação entre essas duas variáveis. No segundo exemplo os dados são fictícios e objetiva-se avaliar os níveis de uma nova droga em relação a um grupo controle. Para cada um dos três grupos experimentais consideram-se cinquenta indivíduos distribuídos em três categorias de resposta. No terceiro exemplo, avalia-se a associação entre a procedência, grupo sanguíneo e estado clínico para úlcera péptica. O estudo dessa tabela de contingência $2 \times 2 \times 3$, baseia-se em 31747 indivíduos.

AGRESTI (1983), baseando-se no fato que o teste qui-quadrado da homogeneidade marginal padrão não leva em conta a ordem sobre as classes, descreve várias estratégias que levam em conta essa dificuldade de ordem e que permite mostrar testes mais poderosos para certas alternativas frequentemente encontradas.

AGRESTI (1984) discute modelos especializados e algumas medidas descritivas para dados categorizados. Descreve detalhadamente os modelos log-linear e logit e utiliza os dados discutidos originariamente por Grizzle, Starmer e Koch (1969) na ilustração de vários tópicos abordados.

OGLIARI (1984) fez um estudo exploratório dos problemas causados pela aproximação assintótica utilizada nos testes de hipóteses para tabelas de contingência. Nesse estudo foram analisadas tabelas de contingência do tipo $C \times D \times F$ obtidas através da simulação de dados de população com tamanhos variados, culminando com a elaboração de um programa em linguagem BASIC para o ajustamento de modelos log-lineares e a determinação das estruturas de independência para os dados dessas tabelas.

MEHTA (1984) apresenta o problema do estabelecimento de equivalência terapêutica entre dois tratamentos, que são inicialmente baseados nos dados categorizados ordenados. O problema é formulado como um teste de significância, em que a hipótese de nulidade especifica a diferença entre tratamentos e complementado com um algoritmo para calcular o nível de significância assintótico.

Para HARVILLE e MEE (1984), agrônomos e biólogos necessitam de técnicas estatísticas para fazer previsões

partindo de um conjunto de informações anteriores. Henderson (1975) descreve um processo de previsão, conhecido como BLUP "best linear unbiased prediction", que é adequado para dados quantitativos. Em alguns problemas de previsão, a resposta é mais categórica que quantitativa. Harville e Mee apresentam um processo semelhante ao BLUP, para respostas categorizadas ordenadas. Essa aproximação é baseada em um modelo no qual se supõe que a categoria é determinada por um valor não observado e que segue modelo linear misto. Os resultados são ilustrados com um exemplo que prevê o grau de dificuldade no parto de vaca leiteira.

BONETT et alii (1985) estendem o modelo linear GSK com as inclusões do modelo log-linear tradicional e o ajuste das distribuições Poisson condicional e incondicional. As estimativas dos parâmetros do modelo são definidas sob os contrastes lineares exatos e estocásticos.

BEITHER e LANDIS (1985) propõem um modelo misto para dados categorizados de experimentos não balanceados, que é análogo ao da análise de variância para classificação dupla de dados quantitativos. Desenvolvem uma extensão da metodologia do ajustamento de constantes, baseado na redução adequada das somas de quadrados, para estimar os componentes de variância do modelo. As estimativas resultantes dos componentes de variância são

incorporadas na estrutura de covariância do modelo linear geral, para testar tratamentos diferentes. O processo é ilustrado com dados obtidos de um laboratório clínico, no qual os tratamentos são considerados como efeitos fixos e as clínicas são consideradas como efeitos aleatórios.

SNAPINN e SMALL (1986) afirmam que modelos de regressão do tipo proporcional aditivo proposto por McCullagh (1980) são, em geral, métodos poderosos para analisar dados categorizados ordenados, supondo-os contínuos de uma distribuição especificada. Testes de significância para esses modelos são geralmente baseados na razão de verossimilhança, cuja estatística resultante se distribui assintoticamente como qui-quadrado. Em um estudo de Monte Carlo, os testes de significância do modelo ordinal são considerados poderosos, entretanto, é proposto um teste modificado, utilizando a distribuição F com um grau de liberdade para o denominador, seu grau de significância empírico está mais próximo do grau nominal, no caso de pequenas amostras. Discutem também as semelhanças entre um modelo ordinal de regressão, suposto normal, e a regressão múltipla convencional. Alguns modelos são apresentados; o primeiro estuda a relação entre fósforo contido na terra e fósforo necessário para a planta e o segundo descreve um ensaio terapêutico, comparando diferentes analgésicos.

GRAUBARD e KORN (1987) referem-se aos métodos estatísticos que testam a não associação entre respostas binárias (linhas) e k grupos de categorias (colunas), normalmente em duas classes: as que necessitam de uma coluna de pontos e as que não necessitam. Um exemplo antigo seria a análise de regressão logística e um exemplo posterior seria o teste de Wilcoxon com soma dos postos. Os autores demonstram que a perceptível vantagem da não predeterminação dos pontos é ilusória. É mostrado um exemplo de experiência com consulta no qual são claramente impróprios os pontos no posto médio usados pelo teste de postos que não necessitam de pontos predeterminados. Suas recomendações são para determinadas colunas de pontos sempre que possível e para considerar pontos igualmente espaçados, quando a opção não é clara. A conveniência do ponto médio como ponto poderá ser examinada antes da aplicação de um teste de postos.

HERRERA (1987) considerou o problema de estimar as frequências esperadas de uma tabela de contingência $R \times S \times T$, quando algumas frequências marginais são consideradas fixas, para dados são provenientes de uma única população multinomial.

Usou o princípio da informação mínima de discriminação para gerar hipóteses de interesse concernentes a interação e efeitos na análise de uma tabela de contingência tridimensional e desenvolveu um programa computacional em PASCAL para tabela de contingência tridimensional.

3. MATERIAL E MÉTODO

3.1. MATERIAL

Os dados utilizados para a exemplificação que envolve uma população são produções (kg) totais de duas laranjeiras, nos anos de 1971 e 1972, de dois pomares experimentais SALIBE (1974), um na área Lageado e outro na Edgárdia, da Fazenda Experimental de Botucatu. Em cada pomar, foram realizados cinco experimentos de porta-enxertos e cada experimento com uma variedade copa. As variedades de copa utilizadas para os cinco experimentos, foram: Hamlim, Baianinha, Westin, Rubi e Itaborai. No experimento, foram utilizados cinco porta-enxertos com seis repetições, dando um total de trezentos dados de produção anual nos dois locais.

Através do programa PRLAR.BAS, a produção total de duas laranjeiras foi classificada numa tabela de dupla entrada, relacionando-a com os anos de 1971 e 1972, sem levar em consideração local, variedade, porta-enxerto e repetições.

Os programas elaborados tiveram por referência o programa MOLIMU (GODOY *et alii*, 1987).

Os dados utilizados para a exemplificação que

envolve várias populações foram obtidos a partir do rendimento médio agrícola municipal de cana-de-açúcar em kg/ha no Estado de São Paulo IBGE(1978-1980), cedidos pela Secretaria de Agricultura do Estado de São Paulo.

Os rendimentos médios agrícolas municipais estão agrupados em trinta e duas microrregiões homogêneas.

Através do programa PRCAN.BAS, os anos pares foram classificados por: grupo de Divisão Regional Agrícola (DIRA) a que pertence a microrregião e faixa de rendimento, conforme exemplo junto aos resultados.

Foram consideradas trinta e duas microrregiões homogêneas nas dez Divisões Regionais Agrícolas do Estado de São Paulo designados por Alta Mojiana (1), Alta Noroeste de Araçatuba (2), São José do Rio Preto (3), Média Araraquarense (4), Serra de Jaboticabal (5), Ribeirão Preto (6), Serra de Batatais (7), Alta Noroeste de Penápolis (8), Bauru (9), Araraquara (10), Depressão Periférica Setentrional (11), Encosta Ocidental Mantiqueira Paulista (12), Alta Paulista (13), Jaú (14), Rio Claro (15), Campinas (16), Estâncias Hidrominerais Paulistas (17), Alta Sorocabana de Presidente Prudente (18), Alta Sorocabana de Assis (19), Ourinhos (20), Serra de Botucatu (21), Açucareira de Piracicaba (22), Tatuí (23), Sorocaba (24), Jundiaí (25), Bragança Paulista (26), Vale do Paraíba Paulista (27), Campos de Itapetininga (28), Grande São Paulo (29), Alto Paraíba (30), Baixada Santista (31) e Costa Norte Paulista (32).

O grupo A foi constituído das DIRAS de Campinas e Vale do Paraíba que envolve as microrregiões 11, 12, 15,

16, 17, 22, 25, 26, 27, 29 e 30. O grupo B foi constituído das DIRAS de Sorocaba e Registro que envolve as microrregiões 21, 23, 24, 28, 31 e 32. O grupo C foi constituído das DIRAS de Ribeirão Preto e São José do Rio Preto que envolve as microrregiões: 1, 3, 4, 5, 6, 7 e 10. O grupo D foi constituído das DIRAS de Bauru, Araçatuba, Presidente Prudente e Marília que envolve as microrregiões: 2, 8, 9, 13, 14, 18, 19 e 20.

3.2. MÉTODO

3.2.1. INTRODUÇÃO

Considerem-se s amostras casualizadas e independentes, de tamanhos $n_1, n_2, \dots, n_s$, extraídas de s distribuições multinomiais, cada uma dessas distribuições apresentando, por hipótese, r categorias de respostas.

As frequências de cada uma dessas caselas, resultantes da amostragem, podem ser resumidas numa tabela de contingência $s \times r$ (Tabela 1), onde cada n_{ij} denota a frequência de respostas da j -ésima categoria, na amostra da i -ésima subpopulação.

Tabela 1 - Distribuição das frequências de respostas das subpopulações

Subpopulação	Categoria de Resposta				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}	$n_{s.}$

O vetor $\tilde{n}_i = (n_{i1}, n_{i2}, \dots, n_{ir})'$ tem, por hipótese, distribuição multinomial com parâmetros $n_{i.}$ e

$$\tilde{\pi}_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{ir})'$$

onde π_{ij} representa a probabilidade teórica de um indivíduo selecionado aleatoriamente da i -ésima subpopulação ser classificado na j -ésima categoria de resposta.

A Tabela 2 apresenta a distribuição das probabilidades teóricas, nas várias subpopulações.

Tabela 2 - Distribuição das probabilidades teóricas nas várias subpopulações

Subpopulação	Categoria de resposta				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 modelo probabilístico associado à Tabela 1, é obtido através do produto das distribuições multinomiais independentes,

$$P(X_1=n_{11}, X_2=n_{12}, \dots, X_r=n_{1r}) = \frac{n_{1.}!}{n_{11}!n_{12}!\dots n_{1r}!} \pi_{11}^{n_{11}} \dots \pi_{1r}^{n_{1r}}$$

$$\text{com } \sum_{j=1}^r n_{ij} = n_{i.} \text{ e } i = 1, 2, \dots, s$$

originando a seguinte distribuição conjunta;

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

$$\text{com as restrições } \sum_{j=1}^r \pi_{ij} = 1 \text{ para } i=1, 2, \dots, s$$

A tabela de contingência das proporções observadas correspondente à Tabela 2. é apresentada a seguir:

Tabela 3. Proporções observadas nas subpopulações

Subpopulação	Categoria de resposta				Total
	1	2	. . .	r	
1	p_{11}	p_{12}	. . .	p_{1r}	1
2	p_{21}	p_{22}	. . .	p_{2r}	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
s	p_{s1}	p_{s2}	. . .	p_{sr}	1

onde $p_{ij} = \frac{n_{ij}}{n_{i.}}$ para $i = 1, \dots, s$ e $j = 1, \dots, r$.

3.2.2. ESTIMAÇÃO DOS PARÂMETROS π_{ij}

Mostraremos a seguir que $\tilde{p}_i = (p_{i1}, \dots, p_{ir})'$, definido anteriormente, trata-se do estimador de máxima verossimilhança de π_i . Por tratar-se de um vetor genérico de subpopulações, omitiremos seu índice na demonstração, considerando então apenas o índice relativo à classe.

Sabemos que $\pi_1 + \pi_2 + \dots + \pi_{r-1} + \pi_r = 1$ e, portanto, a função de verossimilhança é

$$L(\theta) = \frac{n!}{n_1! \dots n_r!} \pi_1^{n_1} \dots \pi_r^{n_r}, \quad \text{onde } \theta = (\pi_1, \dots, \pi_r)'$$

$$L(\theta) = \frac{n!}{n_1! \dots n_r!} \pi_1^{n_1} \dots \pi_{r-1}^{n_{r-1}} (1 - \pi_1 - \dots - \pi_{r-1})^{n_r}$$

Como $L(\theta)$ é monotônica crescente, podemos obter seu máximo, através do máximo de $\ln L(\theta)$

$$\ln L(\theta) = \ln \frac{n!}{n_1! \dots n_r!} + \ln \pi_1^{n_1} + \dots + \ln \pi_{r-1}^{n_{r-1}} + \ln(1 - \pi_1 - \dots - \pi_{r-1})^{n_r}$$

Logo:

$$\frac{\partial \ln L(\theta)}{\partial \pi_1} = n_1 \frac{1}{\pi_1} + n_r \frac{-1}{1 - \pi_1 - \dots - \pi_{r-1}}$$

$$\frac{\partial \ln L(\theta)}{\partial \pi_2} = n_2 \frac{1}{\pi_2} + n_r \frac{-1}{1 - \pi_1 - \dots - \pi_{r-1}}$$

$$\frac{\partial \ln L(\theta)}{\partial \pi_{r-1}} = n_{r-1} \frac{1}{\pi_{r-1}} + n_r \frac{-1}{1 - \pi_1 - \dots - \pi_{r-1}}$$

Igualando essas derivadas parciais a zero e desde

que $\sum_{j=1}^r \hat{\pi}_j = 1$, temos $\frac{n_1}{\pi_1} = \frac{n_2}{\pi_2} = \dots = \frac{n_r}{\pi_r}$ ou equivalentemente:

$$\frac{n_1 + \dots + n_r}{\hat{\pi}_1 + \dots + \hat{\pi}_r} = \frac{n_j}{\hat{\pi}_j} \quad \text{para } j = 1, 2, \dots, r$$

resultando;

$$\frac{n}{1} = \frac{n_j}{\hat{\pi}_j} \quad \text{para } j = 1, 2, \dots, r$$

$$\hat{\pi}_j = \frac{n_j}{n} \quad \text{para } j = 1, 2, \dots, r$$

Dessa forma, para um $\hat{\pi}_{ij}$ genérico, teremos então:

$$\hat{\pi}_{ij} = p_{ij} = \frac{n_{ij}}{n_{i.}} \quad \text{com } i = 1, \dots, s \text{ e } j = 1, \dots, r. \quad (3.1)$$

3.2.3. VARIANCIA DO ESTIMADOR DE π_{ij}

Para obtenção da variância de $\hat{\pi}_{ij}$ podemos fazer a mesma consideração anterior, sem perda de generalidade.

Sabemos que a variância de uma variável aleatória X é dada por $\text{Var}(X) = E(X^2) - [E(X)]^2$.

Dessa forma, temos:

$$f(x_1, \dots, x_r) = \frac{n!}{x_1! \dots x_r!} \pi_1^{x_1} \dots \pi_r^{x_r} \quad \text{com } 0 < \pi_i < 1$$

$$x_1 + \dots + x_r = n;$$

$$\text{e } x_i = 0, \dots, n$$

Logo

$$E(X_i) = \sum_{x_1} \dots \sum_{x_r} x_i \frac{n!}{x_1! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{x_i} \dots \pi_r^{x_r}$$

$$E(X_i) = \sum_{x_1} \dots \sum_{x_r} x_i \frac{n!}{x_1! \dots x_i! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{x_i} \dots \pi_r^{x_r}$$

$$E(X_i) = \sum_{x_1} \dots \sum_{x_r} \frac{n!}{x_1! \dots (x_i-1)! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{x_i-1} \dots \pi_r^{x_r}$$

$$E(X_i) = n \pi_i \sum_{x_1} \dots \sum_{x_r} \frac{(n-1)!}{x_1! \dots (x_i-1)! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{x_i-1} \dots \pi_r^{x_r}$$

Fazendo

$$y_l = \begin{cases} x_l & \text{para } l \neq i \\ x_{l-1} & \text{para } l = i \end{cases}, \text{ temos}$$

$$E(X_i) = n \pi_i \sum_{y_1} \dots \sum_{y_r} \frac{(n-1)!}{y_1! \dots y_i! \dots y_r!} \pi_1^{y_1} \dots \pi_i^{y_i} \dots \pi_r^{y_r} = n \pi_i$$

e ainda

$$E(X_i^2) = \sum_{x_1} \dots \sum_{x_r} x_i^2 \frac{n!}{x_1! \dots x_i! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{x_i} \dots \pi_r^{x_r}$$

$$E(X_i^2) = \sum_{x_1} \dots \sum_{x_r} x_i \frac{n!}{x_1! \dots (x_i-1)! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{x_i} \dots \pi_r^{x_r}$$

$$E(X_i^2) = n \pi_i \sum_{x_1} \dots \sum_{x_r} x_i \frac{(n-1)!}{x_1! \dots (x_i-1)! \dots x_r!} \pi_1^{x_1} \dots \pi_i^{(x_i-1)} \dots \pi_r^{x_r}$$

Considerando a definição anterior dos y_i , temos

$$E(X_i^2) = n \pi_i \sum_{y_1} \dots \sum_{y_r} (y_i+1) \frac{(n-1)!}{y_1! \dots y_i! \dots y_r!} \pi_1^{y_1} \dots \pi_i^{y_i} \dots \pi_r^{y_r}$$

$$\begin{aligned} E(X_i^2) &= n \pi_i \sum_{y_1} \dots \sum_{y_r} y_i \frac{(n-1)!}{y_1! \dots y_i! \dots y_r!} \pi_1^{y_1} \dots \pi_i^{y_i} \dots \pi_r^{y_r} + \\ &\quad + n \pi_i \sum_{y_1} \dots \sum_{y_r} \frac{(n-1)!}{y_1! \dots y_i! \dots y_r!} \pi_1^{y_1} \dots \pi_i^{y_i} \dots \pi_r^{y_r} \end{aligned}$$

$$E(X_i^2) = n \pi_i [(n-1) \pi_i + 1] = n^2 \pi_i^2 + n \pi_i (1 - \pi_i)$$

Dessa forma,

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2$$

$$\text{Var}(X_i) = n_i^2 \pi_i^2 + n_i \pi_i (1 - \pi_i) - (n_i \pi_i)^2$$

$$\text{Var}(X_i) = n_i \pi_i (1 - \pi_i)$$

Generalizando,

$$E(X_{ij}) = n_{i.} \pi_{ij} \quad (3.2)$$

$$\text{Var}(X_{ij}) = n_{i.} \pi_{ij} (1 - \pi_{ij}) \quad (3.3)$$

Do resultado obtido em 3.1, temos então

$$\text{Var}(p_{ij}) = \text{Var}\left(\frac{n_{ij}}{n_{i.}}\right) = \frac{1}{n_{i.}^2} \text{Var}(n_{ij}) = \frac{1}{n_{i.}^2} n_{i.} \pi_{ij} (1 - \pi_{ij})$$

$$\text{Var}(p_{ij}) = \frac{1}{n_{i.}} \pi_{ij} (1 - \pi_{ij}) \quad (3.4)$$

Além disso,

$$\text{Var}(p_{ij} + p_{ij'}) = \text{Var}\left(\frac{n_{ij} + n_{ij'}}{n_{i.}}\right) \quad \text{para } j \neq j'$$

$$\text{Var}(p_{ij} + p_{ij'}) = \frac{1}{n_{i.}^2} \text{Var}(n_{ij} + n_{ij'})$$

que por 3.3 resulta,

$$\text{Var}(p_{ij} + p_{ij'}) = \frac{1}{n_{i.}^2} n_{i.} (\pi_{ij} + \pi_{ij'}) (1 - \pi_{ij} - \pi_{ij'})$$

$$\text{Var}(p_{ij} + p_{ij'}) = \frac{1}{n_{i.}} (\pi_{ij} + \pi_{ij'}) (1 - \pi_{ij} - \pi_{ij'}) \quad (3.5)$$

Sabemos que

$$\text{Cov}(p_{ij}, p_{ij, \cdot}) = - \frac{1}{2} [\text{Var}(p_{ij}) + \text{Var}(p_{ij, \cdot}) - \text{Var}(p_{ij} + p_{ij, \cdot})]$$

Por 3.4 e 3.5, temos

$$\begin{aligned} \text{Cov}(p_{ij}, p_{ij, \cdot}) &= - \frac{1}{2 n_{i.}} [\pi_{ij} (1 - \pi_{ij}) + \pi_{ij, \cdot} (1 - \pi_{ij, \cdot}) - \\ &\quad - (\pi_{ij} + \pi_{ij, \cdot}) (1 - \pi_{ij} - \pi_{ij, \cdot})] \end{aligned}$$

$$\text{Cov}(p_{ij}, p_{ij, \cdot}) = - \frac{1}{n_{i.}} \pi_{ij} \pi_{ij, \cdot} \quad (3.6)$$

3.2.4. CARACTERIZAÇÃO DAS FUNÇÕES PARAMÉTRICAS DE INTERESSE

Neste item, definiremos funções dos parâmetros do vetor π , que possibilitem verificar relações de interesse entre a distribuição das respostas e a natureza das subpopulações.

Dessa forma, sejam $F_1(p), F_2(p), \dots, F_u(p)$ um conjunto genérico de $u \leq (r - 1)$ s funções de π , que, por hipótese, tem derivadas parciais contínuas até segunda ordem com relação a π_{ij} , numa região aberta, contendo $\pi = E(p)$. Considerando o vetor $F = F(p)$, de ordem $u \times 1$, definido por:

$$F' = [F_1(p), F_2(p), \dots, F_u(p)],$$

um estimador consistente da matriz das variâncias e covariâncias de F é uma matriz quadrada de ordem $u \times u$, dada

por

$$\underset{\sim}{V}_F = \underset{\sim}{H} \left[\underset{\sim}{V}(P) \right] \underset{\sim}{H}' \quad \text{onde}$$

$$\underset{\sim}{H} = \left[\frac{dF(\pi)}{d\pi} \mid \pi=p \right] \quad \text{e} \quad \underset{\sim}{V}(p) \text{ é a matriz das derivadas}$$

$$\underset{\sim}{u} \times \underset{\sim}{r} \times \underset{\sim}{s} \quad \underset{\sim}{r} \times \underset{\sim}{s}$$
 parciais de primeira ordem, relativas às u funções paramétricas de interesse calculadas no vetor P . Lembramos também que na determinação da matriz $\underset{\sim}{V}_F$, $\underset{\sim}{V}(p)$ é a matriz bloco diagonal contendo $\underset{\sim}{V}(p_i)$ (estimador consistente de $\underset{\sim}{V}(\pi_i)$) na diagonal principal.

Para a aplicação dos resultados obtidos, as funções de interesse que descrevem F , além de possuírem derivadas parciais até 2ª ordem devem ser tais que $\underset{\sim}{V}_F$ é assintoticamente não singular.

Observamos também que, quando as funções $F_1(\pi)$, $F_2(\pi)$, ..., $F_u(\pi)$ são funções lineares de π , o vetor $F(\pi)$ pode ser escrito da forma $F(\pi) = A\pi$, então $F(p) = Ap$ e, portanto,

$$\underset{\sim}{H} = \underset{\sim}{A} \text{ e } \underset{\sim}{V}_F = \underset{\sim}{A} \underset{\sim}{V}(p) \underset{\sim}{A}'.$$

Para alguns tipos de dados, quando ocorrem $n_{ij}=0$, $\underset{\sim}{V}_F$ será de posto menor que u . Nestes casos, GRIZZLE, STARMER e KOCH (1969), utilizam a proposta de Berkson (1955), a qual sugere substituir os valores nulos por $1/r$ e estimar π_{ij} na i -ésima subpopulação, por $1/(rn_{i.})$. Essa nova

caracterização das probabilidades nas caselas nulas farão que as probabilidades das outras caselas sejam corrigidas, já que a probabilidade total de cada subpopulação é unitária.

Neste sentido, considerando que t das caselas da i -ésima subpopulação sejam nulas e que a distribuição do total das probabilidades determinado pelo processo de Berkson seja equitativamente atribuído entre $r-t$ caselas não nulas, as estimativas das novas probabilidades em relação às originais dada pela fórmula (3.1) são obtidas por

$$p_{ij}^* = p_{ij} - \frac{t}{n_i \cdot r(r-t)} \quad (3.7)$$

3.2.5. AJUSTE DO MODELO E TESTES DE HIPÓTESES LINEARES

Consideremos que se pretende estudar a variação entre os elementos de $F(\pi)$ através de modelos de regressão linear, isto é, $E(F(\pi)) = F(\pi) = X\beta$, onde β é um vetor $(u \times 1)$ e X uma matriz $(u \times u)$.

Utilizando o método dos mínimos quadrados ponderados, encontramos o vetor b $(u \times 1)$ que é um estimador assintótico de β , ou seja, obtido pela minimização da forma quadrática

$$(F - X\beta)' V_F^{-1} (F - X\beta)$$

A expressão desse estimador é dada por

$$\underset{\sim}{b} = (\underset{\sim}{X}' \underset{\sim}{V}_F^{-1} \underset{\sim}{X})^{-1} \underset{\sim}{X}' \underset{\sim}{V}_F^{-1} \underset{\sim}{F}$$

O teste estatístico, referente ao ajuste do modelo, é feito através da estatística

$$SQ[\underset{\sim}{F}(\underset{\sim}{\pi})=\underset{\sim}{X}\underset{\sim}{\beta}] = \underset{\sim}{F}' \underset{\sim}{V}_F^{-1} \underset{\sim}{F} - \underset{\sim}{b}' (\underset{\sim}{X}' \underset{\sim}{V}_F^{-1} \underset{\sim}{X}) \underset{\sim}{b}$$

a qual se distribui assintoticamente como uma distribuição de χ^2 central com $u-v$ graus de liberdade sob a veracidade de H_0 .

Nos casos onde o modelo proposto se ajusta aos dados, hipóteses do tipo $H_0: \underset{\sim}{C}\underset{\sim}{\beta}=0$ relativas a funções lineares dos parâmetros, onde $\underset{\sim}{C}$ ($c \times v$) é uma matriz de constantes conhecidas de posto $c < v$, podem ser testadas através de

$$SQ[\underset{\sim}{C}\underset{\sim}{\beta}=0] = \underset{\sim}{b}' \underset{\sim}{C}' [\underset{\sim}{C} (\underset{\sim}{X}' \underset{\sim}{V}_F^{-1} \underset{\sim}{X})^{-1} \underset{\sim}{C}']^{-1} \underset{\sim}{C} \underset{\sim}{b},$$

que, sob a veracidade de H_0 , tem distribuição χ^2 central com c graus de liberdade.

4. RESULTADOS E DISCUSSÃO

A aplicação da metodologia discutida neste trabalho é reportada, baseando-se em dois exemplos.

4.1. ANÁLISE DE DADOS CATEGORIZADOS PARA UMA ÚNICA POPULAÇÃO, BASEADA NA PRODUÇÃO DE PARES DE LARANJEIRAS

Abordaremos, em primeiro lugar, o exemplo relativo à produção do par de laranjeiras nos anos de 1971 e 1972, para ilustrar a análise de dados categóricos, envolvendo uma única população, realizada em duas etapas. Na primeira etapa, consideraremos o caso em que a tabela de contingência não possui casela nula, enquanto que, na segunda etapa, consideraremos a casela relativa à primeira linha e quarta coluna como nula.

4.1.1. TABELA DE CONTINGÊNCIA SEM CASELA NULA

Através do programa PRLAR.BAS, temos a

seguinte tabela para o primeiro caso:

Tabela 4. Distribuição do número de pares de laranjeiras, nos anos de 1971 e 1972, segundo as classes de produção em kg consideradas:

1971 \ 1972	FAIXAS DE PRODUÇÃO				TOTAL
	(1) 1 - 40	(2) 41 - 80	(3) 81 - 120	(4) 121 - 200	
(1) 1 - 40	21	49	18	2	90
(2) 41 - 80	8	29	37	12	86
(3) 81 - 120	5	9	29	10	53
(4) 121 - 200	6	5	37	23	71
TOTAL	40	92	121	47	300

Neste caso, o interesse é verificar se a distribuição da produção de 1971 e 1972 ocorreu de forma semelhante nas classes de produção consideradas. Isto é, verificar a hipótese:

$$H_0 : \pi_{1.} = \pi_{.1} ; \pi_{2.} = \pi_{.2} ; \pi_{3.} = \pi_{.3} ; \pi_{4.} = \pi_{.4}$$

versus

$$H_1 : \text{existe pelo menos um } \pi_{i.} \neq \pi_{.i}, i=1,2,3,4.$$

Em notação matricial, a hipótese corresponde a

$$H_0 : F(\pi) = A \pi = 0$$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ \bullet & \bullet & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \vdots \\ \pi_{44} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Desta forma, utilizando o programa UNIPOP.BAS com as matrizes abaixo:

MATRIZ DE DADOS

$$\underset{\sim}{N}_{4 \times 4} = \begin{bmatrix} 21 & 49 & 18 & 2 \\ 8 & 29 & 37 & 12 \\ 5 & 9 & 29 & 10 \\ 6 & 5 & 37 & 23 \end{bmatrix}$$

MATRIZ A

$$\underset{\sim}{A}_{8 \times 16} = \begin{bmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix}$$

obtemos os seguintes resultados:

Tabela 5. Proporções observadas nos anos de 1971(linha) e 1972 (coluna).

LINHA	COLUNA	PROPORÇÕES OBSERVADAS
1	1	0.070000
1	2	0.163333
1	3	0.060000
1	4	0.006667
2	1	0.026667
2	2	0.096667
2	3	0.123333
2	4	0.040000
3	1	0.016667
3	2	0.030000
3	3	0.096667
3	4	0.033333
4	1	0.020000
4	2	0.016667
4	3	0.123333
4	4	0.076667

O valor obtido para a estatística $\chi^2 = 65,02$, permite concluir ao nível de 5% de significância que a distribuição da produção, nos dois anos considerados, não ocorre de forma semelhante entre as classes consideradas.

4.1.2. TABELA DE CONTINGÊNCIA COM CASELA NULA

A aplicação para o segundo caso foi feita considerando a mesma distribuição do caso anterior, com exceção da casela relativa à primeira linha e quarta coluna, a qual foi propositadamente zerada.

Nessas condições, obtiveram-se, através do programa UNIPOP.BAS, os seguintes resultados:

Tabela 6. Proporções observadas e proporções corrigidas nos anos de 1971 (linha) e 1972 (coluna).

LINHA	COLUNA	PROPORÇÕES OBSERVADAS	PROPORÇÕES CORRIGIDAS
1	1	0.070470	0.070456
1	2	0.164430	0.164416
1	3	0.060403	0.060389
1	4	0.000000	0.000210
2	1	0.026846	0.026832
2	2	0.097315	0.097301
2	3	0.124161	0.124147
2	4	0.040268	0.040545
3	1	0.016778	0.016764
3	2	0.030201	0.030187
3	3	0.097315	0.097301
3	4	0.033557	0.033543
4	1	0.020134	0.020120
4	2	0.016778	0.016764
4	3	0.124161	0.124147
4	4	0.077181	0.077167

O valor da estatística $\chi^2 = 63,52$, é significativo, o que permite concluir, ao nível de 5%, que a distribuição da produção, nos dois anos considerados, não ocorre de forma semelhante entre as classes consideradas.

4.2. ANÁLISE DE DADOS CATEGORIZADOS PARA VÁRIAS POPULAÇÕES, BASEADO NO RENDIMENTO MÉDIO DE CANA-DE-AÇÚCAR

Para ilustrar a análise de dados categorizados, envolvendo mais de uma população, consideraremos dados relativos a rendimento médio de cana-de-açúcar no Estado de São Paulo. Estes dados constam na Tabela 7, abaixo:

4.2.1. TABELA DE CONTINGÊNCIA SEM CASELA NULA

Através do programa PRCAN.BAS, temos a seguinte tabela para o segundo caso:

Tabela 7. Distribuição das frequências do rendimento médio agrícola de cana-de-açúcar em kg/ha, segundo ano e grupo de DIRA.

GRUPO DE DIRAS	ANO	CLASSE DE RENDIMENTO	
		1 + 55000	55000 + 110000
A	1978	38	49
	1980	34	53
B	1978	14	12
	1980	12	14
C	1978	8	57
	1980	1	64
D	1978	17	27
	1980	12	32

Neste caso, o objetivo é verificar o ajuste de um modelo linear aos dados de rendimento médio agrícola de cana-de-açúcar e testar a influência dos GRUPOS DE DIRAS e ANO nesse rendimento.

Considerando a taxa da classe de rendimento médio entre 1 e 55000 kg/ha, como nosso interesse, temos a seguinte matriz da transformação linear,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, e a$$

equivalente matriz de planejamento

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$$

Desta forma, utilizando o programa SEVPOP.BAS, com o conjunto de dados da tabela 7, obtemos os resultados apresentados a seguir:

Tabela 8. Proporções observadas para o rendimento médio de cana-de-açúcar em kg/ha

GRUPO DE DIRAS	ANO	CLASSE DE RENDIMENTO	PROPORÇÕES OBSERVADAS
1	1	1	0.436782
1	1	2	0.563218
1	2	1	0.538462
1	2	2	0.461538
2	1	1	0.123077
2	1	2	0.876923
2	2	1	0.386364
2	2	2	0.613636
3	1	1	0.390805
3	1	2	0.609196
3	2	1	0.461538
3	2	2	0.538462
4	1	1	0.015385
4	1	2	0.984615
4	2	1	0.272727
4	2	2	0.727273

Para os testes relativo aos efeitos de ANO, GRUPO DE DIRAS e a interação ANO x GRUPO DE DIRAS, temos as seguintes hipóteses estatística:

$$H_{O_A} : \alpha_1 = 0$$

$$\text{ou } H_{O_A} : C_A \beta = 0$$

$$\text{com } C_A = [0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0];$$

$$H_{O_G} : \beta_1 = 0$$

$$\beta_2 = 0$$

$$\beta_3 = 0$$

$$\text{ou } H_{O_G} : C_G \beta = 0$$

$$\text{com } C_G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix};$$

$$H_{O_{AxG}} : (\alpha\beta)_{11} = 0$$

$$(\alpha\beta)_{12} = 0$$

$$(\alpha\beta)_{13} = 0$$

$$\text{ou } H_{O_{AxG}} : C_{AxG} \beta = 0$$

$$\text{com } C_{AxG} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

cujo quadro de resultados apresentamos a seguir.

Quadro 1. Resultados das estatísticas relativas aos testes de hipóteses anteriores

C. V.	GL	χ^2	$\Pr(\chi^2 > \text{ou} =)$
ANO (A)	1	3.248985	0.071468
GRUPO DE DIRAS (G)	3	94.582180	$\cong 0.000000$
A $\times$ G	3	0.567600	0.903808

Como a interação ANO $\times$ GRUPO DE DIRAS não foi significativa, ao nível de 5% , um novo modelo linear é sugerido, utilizando-se a seguinte matriz de planejamento:

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

Quanto ao ajuste do modelo linear categorizado aos dados do rendimento médio de cana-de-açúcar desprezando o efeito da interação AxG, temos o seguinte teste de hipóteses $H_0 : E(y_{ijk}) = \mu + \alpha_i + \beta_j$. O resultado desse teste estatístico é apresentado no Quadro 2.

Quadro 2. Resultados das estatísticas relativas aos testes do ajuste do modelo

C. V.	GL	χ^2	$\Pr(\chi^2 > \text{ou} =)$
ANO (A)	1	7.55171	0.005995
GRUPO DE DIRAS (G)	3	107.36390	$\cong 0.000000$
AJUSTE DO MODELO	3	0.56778	0.903767

Verificamos que ANO e GRUPO DE DIRAS são significativos, enquanto que o ajuste do modelo não o é, caracterizando assim o ajuste dos dados ao modelo proposto.

4.2.2. TABELA DE CONTINGÊNCIA COM CASELA NULA

Da mesma forma do primeiro exemplo, consideraremos neste segundo momento, a casela relativa à sétima linha e primeira coluna como sendo nula. Nestas condições, obtemos os seguintes resultados:

Quadro 3. Resultados das estatísticas relativas aos testes de hipóteses anteriores do ajuste do modelo com casela nula

C. V.	GL	χ^2	$\Pr(\chi^2 > \text{ou} =)$
ANO (A)	1	9.639874	0.001904
GRUPO DE DIRAS (G)	3	119.229700	$\cong$ 0.000000
AJUSTE DO MODELO	3	0.857086	0.835770

Em face dos resultados obtidos para as estatísticas utilizadas, verificamos que ANO e GRUPO DE DIRAS são significativos, enquanto que o ajuste do modelo não o é, caracterizando assim o ajuste dos dados ao modelo proposto.

5. CONCLUSÕES

- 1^o) Através da consulta bibliográfica exigida para o desenvolvimento do nosso trabalho, podemos constatar a existência de uma gama bem variada de publicações em modelos lineares para dados qualitativos, tanto no que diz respeito ao desenvolvimento metodológico quanto à aplicabilidade da metodologia em biologia, de maneira geral.
- 2^o) No que diz respeito a tabelas de contingência com frequências nulas, fato que geralmente ocorre na prática, destaca-se somente o artigo de BERKSON(1955), ao qual, em nosso estudo, demos uma ênfase especial, culminando com a proposição de uma fórmula geral, utilizada nos dois programas computacionais, desenvolvidos no sentido da correção das proporções para as caselas nulas.
- 3^o) Através da revisão bibliográfica elaborou-se dois pro-

gramas computacionais, UNIPOP.BAS e SEVPOP.BAS, em linguagem BASIC. A utilização do programa UNIPOP.BAS presta-se ao caso de uma única população enquanto que o programa SEVPOP.BAS para várias populações.

- 4^o) Quanto à aplicação da metodologia em problemas agronômicos, verificamos que, para a análise de uma população baseado em dados da produção de pares de laranjeiras e para análise de dados envolvendo várias populações, baseada no rendimento médio de cana-de-açúcar, tanto para caso de tabelas de contingência sem caselas nula e com caselas nulas, os resultados obtidos nos testes estatísticos foram os mesmos.
- 5^o) No caso de várias populações, observamos ainda que, para os dois tipos de problemas abordados, houve ajuste do modelo, sem o efeito da interação ANO x DIRAS.

6. REFERENCIAS BIBLIOGRAFICAS

AGRESTI, A., 1983. Testing Marginal Homogeneity for Ordinal Categorical Variables. Biometrics. North Carolina, 39: 505-510.

AGRESTI, A., 1984. Analysis of Ordinal Categorical Data. New York, John Wiley & Sons. 287 p.

BEITHER, P. J. e J. R. LANDIS, 1985. A Mixed-Effects Model for Categorical Data. Biometrics. North Carolina, 41:991-1000.

BERKSON, J., 1955. Maximum likelihood and minimum chi-square estimates of the logistic function. Journal of the American Statistical Association. Washington, 50: 130-162.

BERKSON, J., 1968. Application of minimum logit chi-square estimate to a problem of Grizzle with a notation on the problem of no interaction. Biometrics. North Carolina, 24: 75-95.

BONETT, D. G., J. A. WOODWARD e P. M. BENTHER, 1985. Some Extensions of a Linear Model for Categorical Variables. Biometrics. North Carolina, 41: 745-750.

BRASIL. Fundação Instituto Brasileiro de Geografia e Estatística, 1978-1980. Anuário de produção agrícola municipal (culturas temporais).

CLAUDIO SILVA, Z., 1979. Modelos Lineares para Dados Cua-
litativos. Journal of the Inter-American Statistical-
Institute, 121: 161-176.

DOBSON, A. J., 1983. An Introduction to Statistical Modeling.
Great Britain, University Press, Cambridge. 125 p.

FORTHOFER, R. N. e G. G. KOCH, 1973. An Analysis for Com-
pounded Functions of Categorical Data. Biometrics. North
Carolina, 29: 143-157.

- GODOI, C. R., C. T. SANTOS DIAS e C. H. MATIOLLI, 1987. Sistema de Análise de Modelos Lineares Multidimensionais e Algebra de Matrizes -MOLIMU. Segundo Simpósio de Estatística Aplicada à Experimentação Agronômica. Londrina - Paraná. 72p.
- GRAUBARD, B. I. e E. L. KORN, 1987. Choice of Column Scores for Testing Independence in Ordered $2 \times k$ Contingency Tables. Biometrics. North Carolina, 43: 471-476.
- GRIZZLE, J. E., C. F. STARMER, e G. G. KOCH, 1969. Analysis of Categorical Data by Linear Models. Biometrics. North Carolina, 25: 489-504.
- GRIZZLE, J. E. e O. DALE WILLIAMS, 1972. Log Linear Models and Tests of Independence for Contingency Tables. Biometrics. North Carolina, 28: 137-156.
- HARVILLE, D. A. e R. W. MEE, 1984. A Mixed-Model Procedure for Analyzing Ordered Categorical Data. Biometrics. North Carolina, 40: 393-408.
- HENDERSON, C. R., 1975. Best Linear Unbiased Estimation and Prediction Under a Selection Model. Biometrics. North Carolina, 31: 423-447.

- HERRERA, L. A., 1987. Análise de Tabelas de Contingência Tridimensionais: Uso da Informação Mínima de Discriminação. Piracicaba, ESALQ/USP, 180 p. (Tese de Doutorado).
- IMREY, P. B., W. D. JOHNSON e G. G. KOCH, 1976. An Incomplete Contingency Table Approach to Paired-Comparison Experiments. Journal of the American Statistical Association. Washington, 71: 614-623.
- JOHNSON, W. D. e G. G. KOCH, 1971. A Note on the Weighted Least Squares Analysis of the Ries-Smith Contingency Table Data. Technometrics. North Carolina, 13: 438-447.
- KOCH, G. G. e D. W. REINFURT, 1971. The Analysis of Categorical Data from Mixed Models. Biometrics. North Carolina, 27: 157-173.
- KOCH G. G. e H. DENNIS TOLLEY, 1975. A Generalized Modified Chi-square Analysis of Categorical Bacteria Survival Data from a Complex Dilution Experiment. Biometrics. North Carolina, 31: 59-92.
- KOCH, G. G., H. DENNIS TOLLEY e J. L. FREEMAN, 1976. An Application of the Clumped Binomial Model to the Analysis of Clustered Attribute Data. Biometrics. North Carolina,

32:337-354.

KOCH, G. G., I. A. AMARA, G. W. DAVIS e D. B. GILLINGS, 1982.

A Review of Some Statistical Methods for Covariance Analysis of Categorical Data. Biometrics. North Carolina, 38: 563-595.

KOCH, G. G., J. R. LANDIS, J. L. FREEMAN, D. H. FREEMAN e R.

G. LEHNEN, 1977. A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data. Biometrics. North Carolina, 33: 133-158.

KOCH, G. G., P. B. IMREY e D. W. REINFURT, 1972a. Linear Model Analysis of Categorical Data with Incomplete Response Vectors. Biometrics. North Carolina, 28:663-692.

KOCH, G. G., W. D. JOHNSON e H. DENNIS TOLLEY, 1972b. A Linear Models Approach to the Analysis of Survival and Extent of Disease in Multidimensional Contingency Tables. Journal of the American Statistical Association. Washington, 67: 783-796.

LANDIS, J. R. e G. G. KOCH, 1977b. The measurement of Observer Agreement for Categorical Data. Biometrics. North Carolina, 33: 159-174.

- LANDIS, J. R. e G. G. KOCH, 1977a. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics. North Carolina, 33: 363-374.
- 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. North Carolina, 6: 196-231.
- LEWIS, J. A., 1968. A program to fit constants to multiway tables of quantitative and quantal data. Appl. Statist., 17: 33 - 42.
- MCCULLAGH, P., 1980. Regression models for ordinal data. Journal of the Royal Statistical Society, Series B 42: 109 - 142.
- MEHTA, C. R., N. R. PATEL e A.A. TSIATIS, 1984. Exact Significance Testing to Establish Treatment equivalence with Ordered Categorical Data. Biometrics. North Carolina, 40: 819-825.
- OGLIARI, P. J., 1984. Análise de Dados Categorizados: Modelos Log-Lineares e Independência em Tabelas de Contingência.

Piracicaba, ESALQ/USP, 96 p. (Tese de Mestrado).

RIES, P. N. e H. SMITH, 1963. The use of chi-square for preference testing in multidimensional problems. Chemical Engineering Progress, 52: 39-43.

SALIBE, A. A., 1974. Efeito do porta-enxerto e da localidade no vigor e produção de laranjeiras doces, Citrus Sinensis (L.) Osbeck. Botucatu, FCMBB, 226 p. Tese de Livre Docência.

SNAPINN, S. M. e R. D. SMALL, 1986. Test of Significance Using Regression Models for Ordered Categorical Data. Biometrics. North Carolina, 42: 583-592.

STANISH, W. M., D. B. GILLINGS e G. G. KOCH, 1978. An Application of Multivariate Ratio Methods for the Analysis of a Longitudinal Clinical Trial with Missing Data. Biometrics. North Carolina, 34: 305-317.

THURSTONE, L. L., 1959. The Measurement of Values. University of Chicago Press, Chicago.

WILLIAMS, O. D. e J. E. GRIZZLE, 1972. Analysis of Contingency Tables Having Ordered Response Categories. Journal of the American Statistical Association. Washington, 67: 55-63.

7 . A P [^] E N D I C E : Programas para análise de dados
categorizados pelo método G S K.

7.1 Programa UNIPOP.BAS

```

10 REM UNIPOP.BAS - TRANSFORMACOES LINEARES
20 CLS
30 INPUT "NOME DO ARQUIVO DE DADOS ";AQDA$
40 OPEN "I",#1,AQDA$
50 INPUT "NOME DO ARQUIVO DE RESULTADOS";AQRE$
60 OPEN "O",#2,AQRE$
70 PRINT #2,"ANALISE DE DADOS CATEGORIZADOS P/ UMA UNICA POPULACAO"
80 PRINT
90 INPUT#1,CAM$
100 PRINT #2,CAM$
110 PRINT #2,
120 INPUT #1,NL,NC
130 PRINT #2,"NL = ";NL,"NC = ";NC
140 DIM X(NL,NC),N(NL)
150 PRINT #2,
160 LC=NL*NC
170 PRINT #2," MATRIZ DE DADOS"
180 T=0
190 FOR I=1 TO NL
200 FOR J=1 TO NC
210 INPUT #1,X(I,J)
220 PRINT #2,X(I,J);
230 N(I)=N(I)+X(I,J)
240 NT=NT+X(I,J)
250 IF X(I,J)=0 THEN T=T+1
260 NEXT J
270 PRINT #2,
280 NEXT I
290 PRINT #2,"MATRIZ DE HIPOTESIS"
300 INPUT #1,ML,MC
310 NC1=NC-1
320 IL=ML*ML
330 PRINT #2," ML = ";ML,"MC = ";MC
340 DIM PL(LC),PA(ML),PAVA(ML),AIA(ML,ML),P(NL,NC)
350 DIM V(MC,MC),AP(MC,MC),A(ML,MC),AVA(ML,ML)
360 DIM AV(IL),LV(IL),MV(IL)
370 FOR I=1 TO ML
380 FOR J= 1 TO MC
390 INPUT #1,A(I,J)
400 PRINT #2,A(I,J);
410 NEXT J
420 PRINT #2,
430 NEXT I
440 PRINT #2,
450 PRINT #2,"PROPORCOES OBSERVADAS E PROPORCOES CORRIGIDAS"
460 PRINT #2,
470 FOR I=1 TO NL
480 FOR J=1 TO NC
490 P(I,J)=X(I,J)/NT
500 PRINT #2,I,J,P(I,J);

```

```

510 IF P(I,J)=0 THEN P(I,J)=1/LC/NT: GOTO 540
520 IF T=0 THEN PRINT #2,:GOTO 550
530 P(I,J)=P(I,J)-T/(NT*LC*(LC-T))
540 PRINT #2,TAB(45);P(I,J)
550 NEXT J
560 NEXT I
570 FOR I=0 TO NC1
580 FOR J=1 TO NL
590 FOR K=1 TO NC
600 IF J=K THEN V(J+NL*I,K+NL*I)=P(I+1,J)*(1-P(I+1,J))/NT:GOTO 620
610 V(J+NL*I,K+NL*I)=-P(I+1,J)*P(I+1,K)/NT
620 NEXT K
630 NEXT J
640 NEXT I
650 FOR I=1 TO LC
660 FOR J=1 TO LC
670 PRINT I;J;V(I,J)
680 NEXT J
690 NEXT I
700 REM A*V(P)
710 FOR I=1 TO ML
720 FOR K=1 TO LC
730 FOR J=1 TO LC
740 AP(I,K)=AP(I,K)+A(I,J)*V(J,K)
750 NEXT J
760 NEXT K
770 NEXT I
780 PRINT
790 REM A*V(P)*A'
800 FOR I=1 TO ML
810 FOR K= 1 TO LC
820 FOR J= 1 TO ML
830 AVA(I,J)=AVA(I,J)+AP(I,K)*A(J,K)
840 NEXT J
850 NEXT K
860 NEXT I
870 IJ=0
880 FOR I=1 TO ML
890 FOR J=1 TO ML
900 IJ=IJ+1
910 PRINT AVA(I,J)
920 AV(IJ)=AVA(I,J)
930 NEXT J
940 NEXT I
950 GOSUB 1320
960 IJ=0
970 FOR I= 1 TO NL
980 FOR J=1 TO NC
990 IJ=IJ+1
1000 PL(IJ)=P(I,J)

```

```

1010 NEXT J
1020 NEXT I
1030 REM P'A'
1040 IJ=0
1050 FOR I=1 TO ML
1060 PA(I)=0
1070 FOR J= 1 TO LC
1080 IJ=IJ+1
1090 PA(I)=PA(I)+PL(J)*A(I,J)
1100 NEXT J
1110 NEXT I
1120 REM P'A'[AVA]-'
1130 IJ=0
1140 FOR I=1 TO ML
1150 PAVA(I)=0
1160 FOR J=1 TO ML
1170 IJ=IJ+1
1180 AIA(I,J)=AV(IJ)
1190 PAVA(I)=PAVA(I)+PA(J)*AIA(I,J)
1200 NEXT J
1210 PRINT PAVA(I)
1220 NEXT I
1230 REM QUIQUADRADO
1240 QUI2=0
1250 FOR I=1 TO ML
1260 QUI2=QUI2+PAVA(I)*PA(I)
1270 NEXT I
1280 PRINT #2,"QUIQUADRADO = ";QUI2, " COM ";ML;" GL."
1290 SQ=QUI2
1300 GOSUB 2420
1310 END
1320 REM SUBROTIA PARA INVERTER MATRIZ
1330 D=1
1340 NK=-ML
1350 FOR K=1 TO ML
1360 NK=NK+ML
1370 LV(K)=K
1380 MV(K)=K
1390 KK=NK+K
1400 BIGA=AV(KK)
1410 FOR J=K TO ML
1420 IZ=ML*(J-1)
1430 FOR I= K TO ML
1440 IJ=IZ+I
1450 IF(ABS(BIGA)-ABS(AV(IJ)))>=0 THEN GOTO 1490
1460 BIGA=AV(IJ)
1470 LV(K)=I
1480 MV(K)=J
1490 NEXT I
1500 NEXT J
1510 REM MUDANCA DE LINHA

```

```
1520 J=LV(K)
1530 IF (J-K)<=0 THEN GOTO 1630
1540 KI=K-ML
1550 FOR I=1 TO ML
1560 KI=KI+ML
1570 HOLD=-AV(KI)
1580 JI=KI-K+J
1590 AV(KI)=AV(JI)
1600 AV(JI)=HOLD
1610 NEXT I
1620 REM MUDANCA DE COLUNA
1630 I=MV(K)
1640 IF(I-K)<=0 THEN GOTO 1740
1650 JP=ML*(I-1)
1660 FOR J=1 TO ML
1670 JK=NK+J
1680 JI=JP+J
1690 HOLD =-AV(JK)
1700 AV(JK)=AV(JI)
1710 AV(JI)=HOLD
1720 NEXT J
1730 REM DIVISAO DA COLUNA P/ MENOR PIVO
1740 IF(ABS(BIGA)-9.999999E-21)>0 THEN GOTO 1770
1750 D=0
1760 GOTO 2330
1770 FOR I=1 TO ML
1780 IF(I-K)=0 THEN GOTO 1810
1790 IK=NK+I
1800 AV(IK)=AV(IK)/(-BIGA)
1810 NEXT I
1820 REM MATRIZ REDUZIDA
1830 FOR I=1 TO ML
1840 IK=NK+I
1850 HOLD=AV(IK)
1860 IJ=I-ML
1870 FOR J=1 TO ML
1880 IJ=IJ+ML
1890 IF(I-K)=0 THEN GOTO 1940
1900 IF(J-K)=0 THEN GOTO 1940
1910 KJ=IJ-I+K
1920 AV(IJ)=HOLD*AV(KJ)+AV(IJ)
1930 REM DIVIDE LINHA PELO PIVO
1940 NEXT J
1950 NEXT I
1960 KJ=K-ML
1970 FOR J=1 TO ML
1980 KJ=KJ+ML
1990 IF(J-K)=0 THEN GOTO 2010
2000 AV(KJ)=AV(KJ)/BIGA
2010 NEXT J
2020 REM PRODUTO DOS PIVOS
```



```

2030 D=D*BIGA
2040 REM RECOLOCA PIVO PELO RECIPROCO
2050 AV(KK)=1/BIGA
2060 NEXT K
2070 REM LINHA FINAL E COLUNA DE TROCA
2080 K=ML
2090 K=K-1
2100 IF K<=0 THEN GOTO 2330
2110 I=LV(K)
2120 IF(I-K)<=0 THEN GOTO 2220
2130 JQ=ML*(K-1)
2140 JR=ML*(I-1)
2150 FOR J=1 TO ML
2160 JK=JQ+J
2170 HOLD=AV(JK)
2180 JI=JR+J
2190 AV(JK)=-AV(JI)
2200 AV(JI)=HOLD
2210 NEXT J
2220 J=MV(K)
2230 IF(J-K)<=0 THEN GOTO 2090
2240 KI=K-ML
2250 FOR I=1 TO ML
2260 KI=KI+ML
2270 HOLD=AV(KI)
2280 JI=KI-K+J
2290 AV(KI)=-AV(JI)
2300 AV(JI)=HOLD
2310 NEXT I
2320 GOTO 2090
2330 REM IMPRESSAO DA MATRIZ INVERSA
2340 PRINT #2,
2350 PRINT #2," (AV(P)A`)-1"
2360 PRINT #2,
2370 FOR I= 1 TO IL
2380 PRINT #2,"A(";I;")= ";AV(I)
2390 NEXT I
2400 PRINT #2,
2410 RETURN
2420 VQ=ML
2430 IF VQ= 0 THEN 2650
2440 RQ=1
2450 FOR IQ=VQ TO 2 STEP -2
2460 RQ=RQ*IQ
2470 NEXT IQ
2480 KQ= SQ^(INT((VQ+1)/2))*EXP(-SQ/2)/RQ
2490 IF INT(VQ/2)=VQ/2 THEN 2520
2500 JQ =SQR(2/SQ/3.14159265399#)
2510 GOTO 2530
2520 JQ=1
2530 LQ=1

```

```
2540 MQ=1
2550 VQ=VQ+2
2560 MQ=MQ*SQ/VQ
2570 IF MQ < .0000001 THEN 2600
2580 LQ=LQ+MQ
2590 GOTO 2550
2600 PRINT #2,"Pr(QUI-QUADRADO > ou = ) = ";1-JQ*KQ*LQ
2610 IF 1-JQ*KQ*LQ >= .05 THEN PRINT #2,"N.S." :GOTO 2660
2620 IF 1-JQ*KQ*LQ >= .01 THEN PRINT #2,"* " :GOTO 2660
2630 PRINT #2,"**"
2640 GOTO 2660
2650 PRINT #2,"ERRO GL=0 "
2660 RETURN
```

7.2 Programa SEVPOP.BAS

```

10 REM SEVPOP.BAS (EX 2.3 TRANSFORMACOES LINEARES): CLS
20 PRINT "ANALISE DE DADOS CATEGORIZADOS OBTIDOS ATRAVES DE
    VARIAS POPULACOES"
30 INPUT "NOME DO ARQUIVO DE DADOS";AQDA$
40 INPUT "NOME DO ARQUIVO DE RESULTADOS";AQRE$
50 OPEN "I",#1,AQDA$
60 OPEN "O",#2,AQRE$
70 INPUT #1,ENUN$
80 LPRINT CHR$(12)
90 PRINT #2,"ANALISE DE DADOS CATEGORIZADOS PARA VARIAS POPULACOES"
100 PRINT #2,
110 PRINT #2,ENUN$
120 INPUT #1,NG,NT,NP
130 PRINT #2,"NG = ";NG,"NT = ";NT,"NP = ";NP
140 DIM A(NG,NT,NP),P(NG,NT,NP),S(NG,NT),PD(NG,NT,NP,NP),
    V(NG*NT*NP,NG*NT*NP)
150 DIM HV(NG*NT,NG*NT*NP),HVV(NG*NT),PP(NP),TZ(NG,NT),HVV2(NG*NT,NG*NT)
160 DIM H(NG*NT,NG*NT*NP)
170 REM CALCULO DA VARIANCIA E DA MATRIZ S
180 FOR I=1 TO NP
190 INPUT #1,PP(I)
200 NEXT I
210 N=0
220 TS=0
230 PRINT #2,"GRUPO          TRAT          MATRIZ DE DADOS"
240 FOR I=1 TO NG
250 FOR J=1 TO NT
260 TZ(I,J)=0
270 PRINT #2,I,J,
280 S(I,J)=0
290 N=N+1
300 FOR K=1 TO NP
310 INPUT #1,A(I,J,K)
320 PRINT #2,A(I,J,K);
330 S(I,J)=S(I,J)+A(I,J,K)
340 IF A(I,J,K)=0 THEN TZ(I,J)=TZ(I,J)+1: TS=TS+1
350 NEXT K
360 PRINT #2,
370 PRINT "S(I,J)";S(I,J)
380 NEXT J
390 NEXT I
400 PRINT #2,
410 PRINT #2,"PROPORCOES OBSERVADAS E PROPORCOES CORRIGIDAS"
420 PRINT #2,
430 FOR I=1 TO NG
440 FOR J=1 TO NT
450 FOR K=1 TO NP
460 P(I,J,K)=A(I,J,K)/S(I,J)
470 PRINT #2,I,J,K:P(I,J,K);
480 IF P(I,J,K)= 0 THEN P(I,J,K)=1/S(I,J)/NP:GOTO 510
490 IF TS=0 THEN PRINT #2,: GOTO 520

```

```

500 P(I,J,K)=P(I,J,K) - TZ(I,J)/(S(I,J)*NP*(NP-TZ(I,J)))
510 PRINT #2,TAB(40);P(I,J,K)
520 PRINT I,J,K,P(I,J,K)
530 NEXT K
540 NEXT J
550 NEXT I
560 FOR I=1 TO NG*NT*NP
570 FOR J=1 TO NG*NT*NP
580 V(I,J)=0
590 NEXT J
600 NEXT I
610 REM PRODUTO H*VAR(P)*H'
620 FOR I=1 TO NG
630 FOR J=1 TO NT
640 FOR K=1 TO NP
650 FOR L=1 TO NP
660 IF L=K THEN PD(I,J,K,L)=P(I,J,K)*(1-P(I,J,K)):GOTO 680
670 PD(I,J,K,L)=-P(I,J,K)*P(I,J,L)
680 PD(I,J,K,L)=PD(I,J,K,L)/S(I,J)
690 PRINT I,J,K,L,PD(I,J,K,L)
700 NEXT L
710 NEXT K
720 NEXT J
730 NEXT I
740 SI=0
750 FOR I=1 TO NG
760 FOR J=1 TO NT
770 FOR K=1 TO NP
780 FOR L=1 TO NP
790 NPL=L+SI
800 NPC=K+SI
810 V(NPL,NPC)=PD(I,J,K,L)
820 NEXT L
830 NEXT K
840 SI=SI+NP
850 NEXT J
860 NEXT I
870 SNP=0
880 FOR I=1 TO NG*NT
890 FOR K=1 TO NP
900 H(I,K+SNP)=PP(K)
910 NEXT K
920 SNP=SNP+NP
930 NEXT I
940 PRINT #2,
950 PRINT #2,"MATRIZ DA TRANSFORMACAO LINEAR"
960 PRINT #2,
970 FOR I= 1 TO NG*NT
980 FOR J=1 TO NG*NT*NP
990 PRINT #2,H(I,J);
1000 FOR K=1 TO NG*NT*NP

```

```

1010 HV(I,J)=HV(I,J)+H(I,K)*V(K,J)
1020 NEXT K
1030 NEXT J
1040 PRINT #2, .
1050 NEXT I
1060 FOR I=1 TO NG*NT
1070 FOR J= 1 TO NG*NT
1080 FOR K=1 TO NG*NT*NP
1090 HVH2(I,J)=HVH2(I,J)+HV(I,K)*H(J,K)
1100 NEXT K
1110 IF I=J THEN HVH(I)=HVH2(I,J):PRINT "S(";I;I;") =";HVH(I)
1120 NEXT J
1130 NEXT I
1140 PRINT #2,
1150 PRINT #2,"DIMENSAO DA MATRIZ DE PLANEJAMENTO"
1160 INPUT #1,NL,NC
1170 PRINT #2,
1180 PRINT #2,"NL = ";NL,"NC = ";NC
1190 PRINT #2,
1200 NQ=NC
1210 DIM SM1(NL),X(NL,NC),XSM(NC,NL),XSMX(NC,NC),YSMY(NC,NC)
1220 DIM AV(NQ,NQ),BV(NQ,NQ),AS(NQ,NQ)
1230 DIM BEX(NC,NG*NT),F(NG*NT),BE(NC),BQ(NC)
1240 DIM C(NQ,NC),CHIN(NQ,NC),ABC(NQ),CB(NQ)
1250 REM CALCULO DA INVERSA DE S
1260 FOR I= 1 TO NG*NT
1270 SM1(I)=1/HVH(I)
1280 NEXT I
1290 FOR I=1 TO NL
1300 FOR J=1 TO NC
1310 INPUT #1,X(I,J)
1320 PRINT #2,X(I,J);
1330 PRINT X(I,J);
1340 NEXT J
1350 PRINT #2,
1360 NEXT I
1370 REM PRODUTO X'SM1X
1380 FOR I=1 TO NL
1390 FOR J=1 TO NC
1400 XSM(J,I)=SM1(I)*X(I,J)
1410 NEXT J
1420 NEXT I
1430 R=NC
1440 FOR J=1 TO NC
1450 FOR K=1 TO NC
1460 XSMX(J,K)=0
1470 FOR I=1 TO NL
1480 XSMX(J,K)=XSMX(J,K)+X(I,J)*XSM(K,I)
1490 NEXT I
1500 PRINT J,K,XSMX(J,K)
1510 AV(J,K)=XSMX(J,K)

```

```

1520 YSMY(J,K)=XSMX(J,K)
1530 NEXT K
1540 NEXT J
1550 GOSUB 2660
1560 REM CALCULO DE BETA ESTIMADO
1570 PRINT #2," ESTIMADOR ASSINTOTICO DE BETA"
1580 PRINT #2,
1590 FOR I=1 TO NL
1600 FOR J=1 TO NC
1610 BEX(J,I)=0
1620 FOR K=1 TO NC
1630 XSMX(J,K)=BV(J,K)
1640 BEX(J,I)=BEX(J,I)+XSMX(J,K)*XSM(K,I)
1650 NEXT K
1660 NEXT J
1670 NEXT I
1680 REM CALCULO DE F
1690 IJ=0
1700 FVF=0
1710 FOR I=1 TO NG
1720 FOR J=1 TO NT
1730 IJ=IJ+1
1740 FOR K=1 TO NP
1750 F(IJ)=F(IJ)+A(I,J,K)*PP(K)
1760 NEXT K
1770 F(IJ)=F(IJ)/S(I,J)
1780 FVF=FVF+F(IJ)*SM1(IJ)*F(IJ)
1790 NEXT J
1800 NEXT I
1810 FOR J=1 TO NC
1820 BE(J)=0
1830 FOR I=1 TO NL
1840 BE(J)=BE(J)+BEX(J,I)*F(I)
1850 NEXT I
1860 PRINT #2, "BE (":J;" ) = ",BE(J)
1870 NEXT J
1880 PRINT #2,
1890 INPUT #1,NSQ
1900 PRINT #2, "NUMERO DE EFEITOS = ":NSQ
1910 PRINT #2,
1920 FOR Z=1 TO NSQ
1930 INPUT #1,IDE1$
1940 PRINT #2," EFEITO (":Z;" ) ":IDE1$
1950 INPUT #1,HL,HC
1960 PRINT #2,"HL = ":HL,"HC = ":HC
1970 PRINT #2,
1980 FOR I=1 TO HL
1990 FOR J=1 TO HC
2000 INPUT #1,C(I,J)
2010 PRINT #2,C(I,J);
2020 NEXT J

```

```

2030 PRINT #2,
2040 NEXT I
2050 PRINT #2,
2060 REM CALCULO DA SQ
2070 FOR I=1 TO HL
2080 FOR J=1 TO HC
2090 CHIN(I,J)=0
2100 FOR K=1 TO HC
2110 CHIN(I,J)=CHIN(I,J)+C(I,K)*XSMX(J,K)
2120 NEXT K
2130 NEXT J
2140 NEXT I
2150 FOR I=1 TO HL
2160 FOR J=1 TO HL
2170 AV(I,J)=0
2180 FOR K=1 TO HC
2190 AV(I,J)=AV(I,J)+CHIN(I,K)*C(J,K)
2200 NEXT K
2210 NEXT J
2220 NEXT I
2230 R=HL
2240 GOSUB 2660
2250 FOR I=1 TO HL
2260 CB(I)=0
2270 FOR K=1 TO HC
2280 CB(I)=CB(I)+C(I,K)*BE(K)
2290 NEXT K
2300 NEXT I
2310 FOR I=1 TO HL
2320 ABC(I)=0
2330 FOR J=1 TO HL
2340 ABC(I)=ABC(I)+EV(I,J)*CB(J)
2350 NEXT J
2360 NEXT I
2370 SQ=0
2380 FOR I= 1 TO HL
2390 SQ=SQ+CB(I)*ABC(I)
2400 NEXT I
2410 PRINT #2,
2420 PRINT #2,"QUI-QUADRADO PARA ";I DE1$;" = ";SQ
2430 PRINT #2,
2440 GOSUB 3150
2450 PRINT #2,
2460 NEXT Z
2470 HL=NL-NC
2480 IF HL=0 THEN 2650
2490 BVB=0
2500 FOR J=1 TO NC
2510 BQ(J)=0
2520 FOR I= 1 TO NC
2530 BQ(J)=BQ(J)+BE(I)*YSMY(I,J)

```

```
2540 NEXT I
2550 PRINT BQ(J)
2560 BVB=BVB+BQ(J)*BE(J)
2570 NEXT J
2580 PRINT #2,
2590 AMOD=FVF-BVB
2600 PRINT #2,"AJUSTE DO MODELO =";AMOD
2610 PRINT #2,
2630 SQ=AMOD
2640 GOSUB 3150
2650 END
2660 REM SUBROTIA PARA INVERTER MATRIZ
2670 FOR I=1 TO R-1
2680 FOR J=I+1 TO R
2690 IF AV(I,J)-AV(J,I)>.00001 THEN PRINT "MATRIZ ASSIMETRICA":END
2700 AV(I,J)=AV(J,I)
2710 NEXT J
2720 NEXT I
2730 FOR J=1 TO R
2740 FOR I=1 TO R
2750 BV(I,J)=0
2760 NEXT I
2770 BV(J,J)=1
2780 NEXT J
2790 REM INVERSAO DA MATRIZ
2800 FOR J=1 TO R
2810 FOR I=J TO R
2820 IF AV(I,J)<>0 THEN 2860
2830 NEXT I
2840 PRINT #2,"MATRIZ SINGULAR"
2850 GOTO 2650
2860 FOR K=1 TO R
2870 U=AV(J,K)
2880 AV(J,K)=AV(I,K)
2890 AV(I,K)=U
2900 U=BV(J,K)
2910 BV(J,K)=BV(I,K)
2920 BV(I,K)=U
2930 NEXT K
2940 T=1/AV(J,J)
2950 FOR K=1 TO R
2960 AV(J,K)=T*AV(J,K)
2970 BV(J,K)=T*Bv(J,K)
2980 NEXT K
2990 FOR L=1 TO R
3000 IF L=J THEN 3060
3010 T=-AV(L,J)
3020 FOR K=1 TO R
3030 AV(L,K)=AV(L,K)+T*AV(J,K)
3040 BV(L,K)=BV(L,K)+T*Bv(J,K)
3050 NEXT K
```



```

3060 NEXT L
3070 NEXT J
3080 LPRINT
3090 FOR I=1 TO R
3100 FOR J=1 TO R
3110 PRINT I,J,BV(I,J)
3120 NEXT J
3130 NEXT I
3140 RETURN
3150 VQ=HL
3160 IF VQ= 0 THEN 3380
3170 RQ=1
3180 FOR IQ=VQ TO 2 STEP -2
3190 RQ=RQ*IQ
3200 NEXT IQ
3210 KQ= SQ^(INT((VQ+1)/2))*EXP(-SQ/2)/RQ
3220 IF INT(VQ/2)=VQ/2 THEN 3250
3230 JQ =SQR(2/SQ/3.14159265399#)
3240 GOTO 3260
3250 JQ=1
3260 LQ=1
3270 MQ=1
3280 VQ=VQ+2
3290 MQ=MQ*SQ/VQ
3300 IF MQ < .0000001 THEN 3330
3310 LQ=LQ+MQ
3320 GOTO 3280
3330 PRINT #2,"Pr(QUI-QUADRADO > ou = ) = ";1-JQ*KQ*LQ
3340 IF 1-JQ*KQ*LQ >= .05 THEN PRINT #2,"N.S." :GOTO 3390
3350 IF 1-JQ*KQ*LQ >= .01 THEN PRINT #2,"* " :GOTO 3390
3360 PRINT #2,"**"
3370 GOTO 3390
3380 PRINT #2,"ERRO GL=0 "
3390 RETURN

```