ALGUNS ASPECTOS DA TÉCNICA AID
PARA VARIÁVEL RESPOSTA
CATEGORIZADA

LINDA LEE HO

DISSERTAÇÃO APRESENTADA AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA

UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE MESTRE
EM

ESTATÍSTICA

ÁREA DE CONCENTRAÇÃO: ESTATÍSTICA

ORIENTADOR: PROF. DR. WILTON DE OLIVEIRA BUSSAB

SÃO PAULO, OUTUBRO DE 1987

AGRADECIMENTOS

Quero agradecer a todos os que direta ou indiretamente colaboraram para a realização deste trabalho. Em especial, quero destacar:

- Ao Prof. Dr. Wilton de Oliveira Bussab, pela orientação segura e paciente neste trabalho.

- Ao Prof. Dr. Júlio M. Singer, pelas sugestões e críticas durante o decorrer do trabalho.

- Ao Antonio Carlos Pedroso Lima pela parte de programação computacional.

- A Inês Aschar pela leitura minuciosa.

- A Marlene Santana pelo excelente trabalho de datilografia.

- A Fundação SEADE pela cessão dos dados.

- Aos amigos do Senai e Mavibel pelo apoio constante.

INDICE

	P.
Capítulo 1 - Apresentação	1
1.1 - Introdução	1
1.2 - Um exemplo	3
1.3 - Histórico	7
1.4 - Objetivos do trabalho	9
 Capítulo 2 - Tipos de partição	
2.1 - Introdução	13
2.2 - Método hierárquico divisivo	16
2.3 - Método hierárquico aglomerativo	22
 Capítulo 3 - Medidas de variação para partições fixas	
3.1 - Introdução	25
3.2 - Medidas de variação para partição fixa com a estatística C	29
3.3 - Medidas de variação para partição fixa com a estatísticas de Pearson e da razão de Veros- similhança	41
3.4 - Medidas de variação para partição fixa com a estatística Theta (θ)	43
3.5 - Medidas de variação para partição fixa com a estatística Delta (δ)	52

Capítulo 4 - Medidas de variação para partições ótimas	
4.1 - Introdução	54
4.2 - Seleção de partição ótima	54
4.3 - Critérios para avaliar a significância de partição ótima	55
4.4 - Um exemplo	58
Capítulo 5 - Simulações	
5.0 - Introdução	61
5.1 - Alguns resultados teóricos	62
5.2 - Simulações para partições fixas	63
5.3 - Simulações para partições ótimas	72
5.3.1 - Introdução	72
5.3.2 - Comportamento das estatísticas para valores fixos de K e I	73
5.3.3 - Comportamento das estatísticas para valores variados de I e K	75
5.3.4 - Distribuição empírica das estatísticas ...	81
5.3.5 - Valores críticos obtidos de vários cri- térios	85
5.3.6 - Coincidência da partição ótima escolhida através de critérios diferentes	88
5.4 - Propriedades amostrais de θE_0	89
Capítulo 6 - Aplicações práticas	
6.1 - Exemplo com uma variável preditora	92
6.2 - Exemplos com mais de uma variável preditora	101

6.2.1 - Primeiro exemplo	101
6.2.2 - Segundo exemplo.....	107
Capítulo 7 - Conclusão	118
Apêndice	122
Bibliografia	131

1. APRESENTAÇÃO

1.1 Introdução

Uma situação muito comum encontrada pelo analista de dados é aquela de examinar inter-relações entre um conjunto de variáveis preditoras e uma variável resposta. Uma das maneiras de investigar essas relações, no contexto de modelos lineares, pode ser feita através de regressão linear ou análise de variância quando a variável resposta é do tipo intervalar (contínua), ou de modelos log-lineares ou probites (por exemplo) quando a variável resposta é categorizada. Atualmente, procura-se fazer um tratamento unificado para estes modelos, utilizando a designação de Modelos Lineares Generalizados (McCullagh e Nelder, 1983)./

Principalmente na área das ciências humanas, situações com variável resposta categorizada e variáveis preditoras do tipo categorizada são muito comuns. Note que as variáveis categorizadas podem ser obtidas através de agrupamento de variáveis contínuas em classes. Considere, por exemplo, o consumo de gasolina semanal (em litros) de um dado veículo - esta variável poderia ser agrupada em 3 classes:

- consumo baixo - até 20 litros semanais;
- consumo médio - + 20 a 40 litros semanais;
- consumo alto - + 40 litros semanais

ou em outras classes mais convenientes, conforme situação.

Quando o número de variáveis preditoras é muito grande e, especialmente, quando existe interação, a escolha do modelo adequa-

do e sua conseqüente avaliação podem tornar um trabalho muito árduo e, na maioria dos casos, de difícil interpretação. Assim, podem-se usar técnicas alternativas baseadas em análise de conglomerados, conceitualmente mais intuitivos, e cujos resultados são mais simples de interpretar. Uma dessas técnicas é o Método AID - Automatic Interaction Detection (Morgan e Sonquist, 1963).

Esta técnica consiste em ir dividindo o conjunto de dados sucessivamente em grupos, combinando ou não as classes das variáveis preditoras, de modo a produzir grupos que se diferenciem ao máximo em relação à média da variável resposta. O processo continua enquanto as divisões obtidas contribuem para explicar a variação dos dados. No final do processo, tem-se um conjunto de grupos identificados por algumas variáveis preditoras e suas interações mais importantes; esses grupos apresentam o maior poder explicativo da variável resposta, que será atribuído às variáveis preditoras que os identificam.

Na técnica AID, o efeito da variável preditora é avaliado em cada grupo. Isto também é possível na Análise Regressão, com a introdução de variáveis dummy (Draper & Smith, 1966) e na análise de variância, através de modelos hierárquicos porém, ambas, produzem modelos mais complicados e de difícil interpretação.

Quando a variável resposta é do tipo categorizado, os mesmos argumentos citados podem ser usados para sugerir técnicas alternativas aos modelos log-lineares ou probites, etc. Tal técnica, considerada extensão da técnica AID, pode e deve ser usada como técnica exploratória ou complementar àquelas análises. É uma técnica útil, principalmente quando envolve, simultaneamente, um con-

junto de dados grande e um conjunto de muitas variáveis preditoras. Este trabalho pretende abordar alguns aspectos da técnica AID com variável resposta categorizada.

1.2 Um Exemplo

Suponha que um novo produto está para ser lançado no mercado e que foi realizada uma pesquisa para caracterizar o tipo de comprador. A variável resposta, neste caso, é dicotômica (compra ou não compra o produto) e as variáveis preditoras são: local de residência, sexo, idade, renda familiar, escolaridade, tamanho da família e ocupação principal. Na Tabela 1.2.1 estão resumidas as informações referentes às variáveis preditoras e resposta. A coluna denominada tipo de partição será explicada no Capítulo 2.

TABELA 1.2.1 - Descrição das variáveis preditoras e resposta referentes à pesquisa para caracterização dos compradores do produto em questão.

Nome da variável	Tipo de variável	Nº de classes	Tipo de partição	Rótulo das classes
compra de um produto novo	resposta	2		Sim, Não
local de residência	preditora	2	livre	Grande São Paulo, Interior
Sexo	preditora	2	livre	Masculino, Feminino
Ocupação principal	preditora	5	livre	Empregado, empregador, autônomo, prof. liberais e outros.
Idade	preditora	4	Monotônica	<24, 24 a 35, 36 a 45, > 45 anos
Renda familiar	preditora	4	Monotônica	até 1 S.M., + 1 a 3 S.M., + 3 a 5 S.M., + 5 S.M.
Tamanho da família	preditora	3	Monotônica	até 2 p, 3 a 4 p, 5 ou + p.
Escolaridade	preditora	4	Flutuante	até 1 g, + 1 a 2 g, + 2 grau, s/ declaração.

S.M. = salário mínimo, p = pessoas.

A Figura 1.2.1 apresenta um resultado típico da aplicação da técnica AID a um conjunto de dados com características indicadas na Tabela 1.2.1.

FIGURA 1.2.1 - Resultado da aplicação da técnica AID ao conjunto de dados referente ao teste do novo produto.

Legenda:

prop. de compradores	prop. de não compradores
----------------------	--------------------------

○ → identificação do grupo

O primeiro passo na aplicação da técnica consiste em analisar a variável resposta para o grupo todo. Pode-se observar na casela 1 da Figura 1.2.1 que a proporção de compradores é de 0.45. A seguir, respeitando algum critério conveniente, procura-se a variável preditora que mais contribua para explicar a variabilidade da variável resposta. No exemplo, essa variável é local de residência, que separa o grupo todo em dois subgrupos: um contendo os moradores da Grande São Paulo e o outro os moradores do Interior. Observando a Figura 1.2.1, nota-se que estes grupos apresentam comportamentos diferentes com relação à variável resposta; a estimativa da proporção de compradores na Grande São Paulo decresceu de 0.45 para 0.35 (casela 2), enquanto que, no Interior, essa proporção estimada cresceu de 0.45 para 0.72 (casela 3).

No passo seguinte, executa-se a mesma operação em um dos 2 grupos resultantes, repetindo-se o procedimento até que se atinja algum critério de parada pré-fixado. No caso, conforme a Figura 1.2.1, obtiveram-se seis grupos finais, cujas descrições estão na Tabela 1.2.2.

TABELA 1.2.2 - Descrição dos grupos obtidos através da Técnica AID.

Grupo	Características do Grupo	prop. de comprar o prod.	identificação do grupo
1	Morador do Interior e idade ≤ 35 anos	0.80	7
2	Morador da Gde. S.P. do sexo masculino	0.75	5
3	Morador do Interior, +35 anos com renda familiar igual ou superior a 5 s.m.	0.60	11
4	Morador da Gde. S.P. do sexo feminino com instrução superior ao 2.grau	0.56	9
5	Morador do interior, +35 anos com renda familiar inferior a 5 s.m.	0.33	10
6	Morador da Gde. S.P. do sexo feminino com instrução até 2. grau	0.09	8

Pode-se constatar que algumas variáveis preditoras, apesar de terem sido incluídas na análise, não foram utilizadas nas partições. Isso se deve ao fato de que as mesmas não contribuíram significativamente para explicação da variabilidade da variável resposta ou estão altamente correlacionadas com as variáveis preditoras utilizadas nas partições.

1.3 Histórico

Os algoritmos referentes à técnica AID com variável resposta categorizada foram inicialmente obtidos como modificações do algoritmo THAID (Messenger e Mandell, 1972; Morgan e Messenger, 1973), cujas partições são obtidas a partir de critérios de seleção não-

paramétricos (Theta (θ) e Delta (δ)).

O primeiro critério utilizado (Theta) foi baseado em estratégias de previsões ótimas, e a estatística utilizada neste critério tem relação com a medida de associação Lambda desenvolvida por Goodman e Kruskall (1954).

O segundo critério (Delta) foi baseado na simples noção que se deveria procurar partições cujas distribuições de probabilidade se diferenciasssem ao máximo daquela do grupo original. Para medir esta diferença, foi sugerida uma estatística baseada no critério da distância em valor absoluto. O critério Delta foi anteriormente chamado de "deslocamento máximo" por Belson (1968) num contexto mais simples, com uma variável resposta dicotômica.

A estatística de Pearson
$$(X = \sum_i \frac{(O_i - E_i)^2}{E_i})$$
, como

critério para obter as partições, foi utilizada no método ELISEE (Cellard, Labbe e Savistky, 1967; Bouruche e Tennehaus, 1972).

Um outro critério utilizado no contexto de análise de variância para dados categóricos (CATANOVA) com uma variável preditora foi introduzido por Light e Margolin (1971). A estatística C (que será vista no capítulo 3), utilizada para verificar o efeito da variável preditora, é baseada numa medida de variabilidade para dados categorizados definida por Gini (1938).

As partições obtidas com os critérios já citados baseiam-se em métodos iterativos divisivos dicotômicos e são avaliadas através de critério de parada empírico. Tais métodos consideram, inicialmente, o conjunto de dados como um único grupo. Em seguida,

obedecendo a algum critério de seleção, ele é dividido em 2 grupos ótimos. O passo seguinte é dividir novamente apenas um dos 2 grupos resultantes em outros 2 novos. E assim, sucessivamente, até satisfazer um critério de parada (empírico) previamente estabelecido.

Kass (1980) introduziu um novo método, utilizando a estatística de Pearson como critério de seleção das partições ótimas, através de um método aglomerativo. Este algoritmo recebeu o nome de CHAID (Chi-Square Automatic Interactive Detection). Além disso, Kass apresentou também alguns critérios de parada não empíricos.

Além destes critérios de seleção citados, outros poderiam ser utilizados para obter partições ótimas, como aqueles empregados nos testes não-paramétricos ou, ainda, qualquer outro que sirva para medir a diferença entre amostras provenientes de populações multinomiais.

1.4 Objetivo do trabalho

O objetivo do trabalho é abordar alguns aspectos da técnica AID numa situação com variável resposta categorizada.

Para aplicar esta técnica, devem-se levar em conta os seguintes aspectos:

- 1) As combinações das classes das variáveis preditoras devem ser interpretáveis. Assim, devem-se tomar certas precauções nas diversas formas alternativas de divisão; no caso de variáveis ordinais, por exemplo, os grupos só terão interpretação adequada se forem

compostos por classes contíguas. No exemplo da Seção 1.2, a variável Idade (ordinal) apresentava inicialmente 4 classes e, de todas as possíveis maneiras de dividir estas classes em 2 grupos, haveria apenas 3 alternativas interpretáveis, que são:

Grupo A	Grupo B
< 24	≥ 24 anos
≤ 35	> 35 anos
≤ 45	> 45 anos

As partições que obedecem a este tipo de restrição recebem o nome de **partições monotônicas**. Este e outros tipos de restrições nas partições serão objeto de discussão no Capítulo 2.

2) A divisão em grupos distintos requer a definição de uma função que permita selecionar a partição "ótima". No exemplo anterior, para escolher a alternativa mais adequada para dividir as classes das variáveis preditoras ou a variável preditora mais importante para a divisão, usou-se a estatística de Pearson (X). Outras estatísticas paramétricas poderiam ser usadas, tais como: A Razão de Verossimilhança (V), a estatística C de Light e Margolin, as estatísticas Theta (θ) e Delta (δ), que serão descritas para o caso de uma partição fixa no Capítulo 3.

3) A estatística é calculada para cada uma das partições possíveis e é selecionada a melhor partição segundo algum critério estabelecido. O mais usual é o critério empírico que consiste em escolher a partição que apresentar maior valor da estatística. Há

também outro critério sugerido por Kass (1980) que será abordado no capítulo 6. No exemplo anterior, é calculada a estatística de Pearson para cada uma das partições possíveis e estabelece-se que a melhor partição é a que apresentar maior valor desta estatística. Uma das maiores dificuldades encontradas nestes métodos iterativos é a de se decidir até onde continuar as divisões sucessivas. Além dos critérios mais usados, que são os critérios empíricos, outros avaliam o "ganho" introduzido numa nova partição através da significância da estatística associada à partição ótima. A maior dificuldade da utilização deste último critério reside na identificação da distribuição da estatística utilizada no teste. No exemplo anterior, embora a estatística usada seja a estatística de Pearson, sua distribuição não é conhecida pois, na realidade, escolhe-se aquela combinação que produz o maior valor da estatística de Pearson. A estatística de Pearson tem distribuição assintótica qui-quadrado sob hipótese nula de homogeneidade com u graus de liberdade onde $u = (I-1)*(K-1)$, porém a distribuição do max das estatística de Pearson não tem a mesma distribuição assintótica.

No capítulo 4 serão descritos:

- estatísticas utilizadas para escolher as partições ótimas e suas propriedades;
- modos de obter as partições ótimas;
- modos de avaliar a significância da estatística associada à partição ótima.

Salientando que a técnica AID é iterativa, ou seja, ela é aplicada sucessivamente até que alguma regra de parada esteja

satisfeita, neste trabalho será privilegiada a forma de se obter a primeira partição ótima dicotômica com apenas uma variável preditora. A generalização do processo iterativo de escolha de partições ótimas com mais de uma variável preditora será abordada no capítulo 6, através de algumas aplicações práticas .

No capítulo 5 serão comparados, através de simulação, os comportamentos de algumas estatísticas utilizadas na escolha das partições fixas e ótimas.

Nenhum dos pacotes computacionais mais usuais como SPSS, BMDP, SAS apresentam procedimentos aplicativos diretos da Técnica AID. A obtenção de resultados através destes pacotes é quase artesanal, havendo necessidade de se utilizar procedimentos como FREQ (SAS), CROSSTABS (SPSS) e 4D (BMDP) diversas vezes. Com relação às estatísticas utilizadas na escolha das partições ótimas, apenas a estatística de Pearson e Razão de Verossimilhança estão disponíveis nesses pacotes. Outra alternativa é o pacote denominado OSIRIS (Sonquist, 1963), com a opção de utilizar apenas a estatística de Pearson como critério de seleção da partição ótima. Além desse, há um outro programa denominado CHAID (Kass, 1980) que utiliza o método aglomerativo para obter as partições ótimas, selecionadas segundo a estatística de Pearson.

2 TIPOS DE PARTIÇÃO

2.1 Introdução

O objetivo deste capítulo é abordar as várias maneiras de dividir um conjunto de dados em grupos segundo a técnica AID. Uma solução seria examinar todas as partições possíveis do conjunto de dados e escolher aquela que é "ótima", segundo algum critério. Na prática, isto pode ser inviável devido ao grande número de alternativas que deveriam ser examinadas. Nas situações em que o número final de grupos é conhecido, seriam examinadas apenas as partições com este número de grupos, diminuindo bastante o número de alternativas a serem investigadas. Esta solução, no entanto, só é viável se o número de classes das variáveis preditoras for pequeno.

Por exemplo, considere uma situação com as variáveis preditoras A (A_1, A_2) e B (B_1, B_2), ambas com 2 classes onde se quer dividir um conjunto de dados em 2 grupos finais. As 7 alternativas, com 2 grupos, que devem ser examinadas para escolher a partição ótima são:

ALTERNATIVA	GRUPO 1	GRUPO 2
1	A1 B1	A1 B2, A2 B1, A2 B2
2	A1 B2	A1 B1, A2 B1, A2 B2
3	A2 B1	A1 B1, A1 B2, A2 B2
4	A2 B2	A1 B1, A1 B2, A2 B1
5	A1 B1, A1 B2	A2 B1, A2 B2
6	A1 B1, A2 B1	A2 B2, A1 B2
7	A1 B1, A2 B2	A1 B2, A2 B1

Uma outra situação é aquela em que algumas partições são eliminadas do conjunto das soluções possíveis com base em conhecimento prévio, teoria ou condições de conveniência gerais. Suponha uma situação em que as observações são tomadas sucessivamente durante um período de tempo. Deseja-se dividir o conjunto de dados respeitando a ordem cronológica. Tal restrição elimina um grande número de partições possíveis.

Assim, cada problema tem associado um conjunto de partições admissíveis a ser examinado.

As partições podem ser classificadas como:

- a) **livres**: nenhuma condição é imposta às partições;
- b) **restritas**: uma ou mais condições são impostas às partições;

b.1) **monotônicas**: os grupos devem ser formados por combinações de classes consecutivas da variável preditora com respeito a alguma ordem fixada. O caso onde as divisões respeitam a ordem cronológica é um exemplo de partição monotônica.

b.2) **flutuantes**: considere uma situação em que os grupos devem ser formados por combinações das classes da variável preditora, onde em apenas algumas classes, são impostas restrições nas partições; as classes em que não são impostas restrições são chamadas de classes flutuantes. Por exemplo, suponha que informações sobre renda mensal foram categorizadas em:

- Não informou/recusa/ausência
- até 5 SM
- + 5 a 10 SM SM = salário mínimo
- + 10 SM

e se quer dividir o conjunto de dados, impondo restrições nas 3 últimas classes. Este tipo de partição foi introduzido por Kass (1980) e é útil, principalmente, para caracterizar o comportamento dos dados omissos. Por exemplo, na situação acima, pode-se verificar qual das classes de renda apresenta o mesmo "comportamento" da classe flutuante.

b.3) complexos: outros tipos de restrições (circulares, parcialmente restritos). São casos onde são impostas restrições especiais às partições. Suponha que alguma variável geológica é coletada segundo as classes categorizadas:

$\begin{array}{l} \circ \\ \text{até } 90^\circ \text{ da direção Norte} \\ \circ \quad \circ \\ + 90^\circ \text{ a } 180^\circ \text{ da direção Norte} \\ \circ \quad \circ \\ + 180^\circ \text{ a } 270^\circ \text{ da direção Norte} \\ \circ \quad \circ \\ + 270^\circ \text{ a } 360^\circ \text{ da direção Norte} \end{array}$

A combinação das classes deve obedecer a uma restrição especial: deve ser consecutiva e considerar a primeira classe como o consecutivo da última classe. Este caso é um exemplo de partição circular.

De modo geral, as restrições utilizadas para definir as partições são chamadas de Regra (s) Gama (Bussab, 1976).

Neste trabalho, a solução alternativa para encontrar as par-

tições ótimas será feita através dos métodos hierárquicos: divisivo e aglomerativo. As características desses métodos serão discutidas a seguir, sendo que será dada maior ênfase ao primeiro, enquanto que o método aglomerativo será brevemente descrito.

2.2 Método hierárquico divisivo

Este método foi inicialmente apresentado por Sonquist et al (1963) para variáveis respostas contínuas. Sua extensão para variáveis respostas categorizadas foi apresentada por Morgan e Messenger (1973), podendo ser resumida nas seguintes etapas:

1) No início, todas as classes da variável preditora são consideradas como um único grupo, que será dividido em 2 grupos, obedecendo a um critério de otimização.

2) Em seguida, obedecendo a critério de seleção, escolhe-se um dos dois grupos resultantes para ser novamente divididas em 2.

3) Esta operação é repetida até que o número pré-fixado de grupos finais seja alcançado ou algum outro critério de parada esteja satisfeito.

Definição 2.2.1 - Uma partição é chamada gama-dicotômica se for obtida a partir de uma regra Gama, que pode ser: Livre (L), Monotônica (M), Flutuante (F), segundo o método hierárquico divisivo.

Seja $J = \{1, 2, \dots, K\}$ o conjunto de índices das classes da variável preditora. Então pode-se considerar a seguinte.

Definição 2.2.2 - $\Phi(J, \text{gama})$ é o conjunto de todas possíveis partições gama-dicotômicas do conjunto J . Se $d \in \Phi(J, \text{gama})$, então d é formada por 2 subconjuntos: $A(d)$ e $B(d)$, tais que:

$$A(d) \cup B(d) = J$$

$$A(d) \cap B(d) = \emptyset \quad (2.2.1)$$

Exemplo 2.2.1

Suponha $J = \{1, 2, 3, 4\}$. Se a regra for livre, então $\Phi(J, L)$ conterá 7 elementos, ou seja: $\Phi(J, L) = (d_1, d_2, d_3, \dots, d_7)$, onde as partições definidas pelos $d(j)$, $j = 1, 2, \dots, 7$ estão indicadas na tabela 2.2.1.

TABELA 2.2.1 - Partições dicotômicas de $J = \{1, 2, 3, 4\}$ obedecendo a regra Livre. Exemplo 2.2.1.

d	A(d)	B(d)
d_1	{1}	{2, 3, 4}
d_2	{2}	{1, 3, 4}
d_3	{3}	{1, 2, 4}
d_4	{4}	{1, 2, 3}
d_5	{1, 2}	{3, 4}
d_6	{1, 3}	{2, 4}
d_7	{2, 3}	{1, 4}

Se a regra gama for monotônica (M) e respeitar a ordem natural, então $\Phi(J, M) = (d_1, d_4, d_5)$, onde d_1, d_4 e d_5 estão

definidos na Tabela 2.2.1.

Se a regra gama for flutuante (F) e tal que $\{1, 2, 3\}$ só permite partições monotônicas na ordem natural, então $\Phi(J, F) = (d_1, d_5, d_4, d_7, d_3)$, onde d_1, d_5, d_4, d_7, d_3 estão indicados na Tabela 2.2.1.

Exemplo 2.2.2 - Considere agora o conjunto J esquematizado na Tabela 2.2.2, correspondendo às caselas de uma classificação cruzada de 2 variáveis preditoras (A e B), onde a variável preditora A tem 2 classes (M, F) e a variável preditora B tem 4 classes (B_1, B_2, B_3, B_4).

Os números do corpo da tabela identificam a classificação cruzada das 2 variáveis preditoras.

Índice	Variável A	Variável B
1	M	B ₁
2		B ₂
3		B ₃
4		B ₄
5	F	B ₁
6		B ₂
7		B ₃
8		B ₄

TABELA 2.2.2 - Esquema da classificação cruzada das variáveis preditoras A e B. Exemplo 2.2.2.

		Classes da variável preditora B			
Classes da variável preditora A	A \ B	B ₁	B ₂	B ₃	B ₄
	M	1	2	3	4
	F	5	6	7	8

Considere uma regra gama que:

a) permite partições baseadas apenas nos agrupamentos dos marginais;

b) com relação à variável preditora B só permite partições monotônicas na ordem natural; então ter-se-ia o seguinte conjunto $\Phi(J, \text{Gama})$, esquematizando na Tabela 2.2.3.

TABELA 2.2.3 - Partições dicotômicas de J correspondentes ao exemplo 2.2.3.

d	A(d)	B(d)
d_1	(1, 2, 3, 4)	(5, 6, 7, 8)
d_2	(1, 5)	(2, 6, 3, 7, 4, 8)
d_3	(1, 2, 5, 6)	(3, 7, 4, 8)
d_4	(1, 2, 3, 5, 6, 7)	(4, 8)

Considere, agora, o conjunto $\Phi[B(d), \text{Gama}]$. Cada um dos

seus elementos d' é novamente formado por 2 subconjuntos de $B(d)$: $C(d')$ e $D(d')$, tais que:

$$C(d') \cup D(d') = B(d),$$

$$C(d') \cap D(d') = \emptyset$$

Pode-se constatar que o conjunto $\{A(d), C(d'), D(d')\}$ constituirá uma tri-partição do conjunto J , obtida através de 2 partições gama-dicotômicas sucessivas.

Definem-se as tri-partições do conjunto J formadas, desta maneira, como resultado de uma operação especial dada por

$$A(d) * \text{Phi}[B(d), \text{gama}] = ([A(d), C(d'), D(d'), d' \in (B(d), \text{Gama})]) \quad 2.2.2$$

Pode-se definir a classe $\Pi(3, J, \text{Gama})$ de todas as tri-partições do conjunto J , obtida por 2 partições sucessivas gama-dicotômicas, onde:

$$\Pi(3, J, \text{Gama}) = \bigcup_{d \in \text{Phi}(J, \text{Gama})} ([\text{Phi}(A(d), \text{Gama}) * B(d)] \cup [A(d), \text{Gama}] * \text{Phi}[B(d), \text{Gama}]) \quad 2.2.3$$

Seja j um elemento de $\Pi(3, J, \text{Gama})$, então j é formado por 3 subconjuntos de J :

$$j = [A(j), B(j), C(j)], \text{ tais que: } A(j) \cup B(j) \cup C(j) = J$$

$$A(j) \cap B(j) \cap C(j) = \emptyset$$

$$A(j) \cap B(j) = \emptyset$$

$$A(j) \cap C(j) = \emptyset$$

$$B(j) \cap C(j) = \emptyset$$

Aplicando a operação $*$ nos elementos da classe $\Pi(3, J, \text{Gama})$,

pode-se construir a classe $\pi(4, J, \text{Gama})$, constituída de tetra-partições do conjunto J , obtida por 3 divisões gama-dicotômicas sucessivas, onde:

$$\begin{aligned} \pi(4, J, \text{Gama}) = & \bigcup_{J \in \pi(3, J, \text{Gama})} \{ [A(J), \text{gama}] * B(J) * C(J) \} \cup \\ & [A(J) * (B(J), \text{Gama}) * C(J)] \cup \\ & [A(J) * B(J) * \text{Phi}(C(J), \text{Gama})] \} \end{aligned} \quad (2.2.4)$$

Este procedimento pode ser repetido até se alcançar a partição onde cada subconjunto tem apenas um elemento, constituindo a classe $\pi(K, J, \text{Gama})$ onde:

$$\pi(K, J, \text{Gama}) = \{ (1), (2), \dots, (K) \} \quad (2.2.5)$$

→ **Definição 2.2.3** - $\text{Tau}(J, \text{Gama})$ é a classe de todas as partições possíveis do conjunto J , obtidas por aplicações sucessivas da divisão gama dicotômica, ou seja,

$$\text{Tau}(J, \text{Gama}) = \bigcup_{k=1}^K \pi(k, J, \text{Gama}), \quad (2.2.6)$$

lembrando que $\pi(1, J, \text{gama}) = J$ e

$$\pi(2, J, \text{gama}) = \text{Phi}(J, \text{gama})$$

Exemplo 2.2.3

Seja $J = \{1, 2, 3, 4\}$. A classe $\text{Tau}(J, L)$ é formada por 15 partições, que são:

$$i(1) = \{(1, 2, 3, 4)\}$$

$$i(10) = \{(1, 3), (2), (4)\}$$

$$i(2) = \{(1), (2, 3, 4)\}$$

$$i(11) = \{(1, 4), (2), (3)\}$$

$$i(3) = \{(2), (1, 3, 4)\}$$

$$i(12) = \{(1), (2), (3, 4)\}$$

$$i(4) = \{(3), (1, 2, 4)\}$$

$$i(13) = \{(1), (3), (2, 4)\}$$

$$i(5) = \{(4), (1, 2, 3)\}$$

$$i(14) = \{(1), (4), (2, 3)\}$$

$$i(6) = \{(1, 2), (3, 4)\}$$

$$i(15) = \{(1), (2), (3), (4)\}$$

$$i(7) = \{(1, 3), (2, 4)\}$$

$$i(8) = \{(1, 4), (2, 3)\}$$

$$i(9) = \{(1, 2), (3), (4)\}$$

Se a regra Gama for monotônica (M), a classe Tau (J, M) será constituída de 8 partições com :

$\text{Tau}(J, M) = \{i(1), i(2), i(6), i(5), i(9), i(12), i(14), i(15),$
onde os $i(j)$ são os mesmos indicados na classe Tau(J, L).

Ressalta-se que, neste trabalho, apenas a classe $\pi(2, J, \text{Gama})$ será focalizada detalhadamente para investigar as propriedades estatísticas, visto que as classes $\pi(2+j, J, \text{Gama})$, $j = 1, \dots, K-2$, podem ser obtidas através de $(j-1)$ partições gama-dicotômicas sucessivas de J.

2.3 Método hierárquico aglomerativo

Este método foi apresentado por Kass (1980) para variáveis respostas categorizadas, cujas etapas estão resumidas a seguir:

1) Inicialmente, as classes são consideradas como grupos individuais, ou seja, se fossem K classes, seriam K grupos.

2) Em seguida, obedecendo a algum critério, dois grupos são selecionados para compor um único grupo.

3) O passo 2 é repetido até que o número pré-fixado de grupos finais seja alcançado ou algum outro critério de parada esteja satisfeito.

Como no método divisivo, conhecimento prévio ou condições de conveniência podem eliminar algumas partições das soluções possíveis; assim, cada problema tem associado um conjunto de regras que determinam aglomerações admissíveis ou não, obedecendo a todos os procedimentos descritos para o método divisivo dicotômico.

Definição 2.3.1 - $W(J, \text{Gama})$ é o conjunto das aglomerações possíveis do conjunto J , obedecendo a regra gama, obtidas segundo o método aglomerativo (na 1-^a iteração).

Exemplo 2.3.1 - Suponha $J = \{1, 2, 3, 4\}$. Se a regra gama = L, $W(J, L) = (h_1, h_2, \dots, h_6)$ onde as partições definidas pelos $h(j)$ indicadas na Tabela 2.3.1.

TABELA 2.3.1 - Partições de $J = \{1, 2, 3, 4\}$, segundo método aglomerativo.

h	Partições Possíveis
h_1	(1, 2), (3), (4)
h_2	(1, 3), (2), (4)
h_3	(1, 4), (2), (3)
h_4	(2, 3), (1), (4)
h_5	(2, 4), (1), (3)
h_6	(3, 4), (1), (2)

Se a regra $Gama = M$, então $W(J, Gama) = (h_1, h_4, h_6)$, onde h_1, h_4 e h_6 estão indicados na Tabela 2.3.1.

3 MEDIDAS DE VARIAÇÃO PARA PARTIÇÕES FIXAS

3.1 Introdução

Numa situação de análise de variância com um fator fixo (com K classes), as medidas de variação da variável resposta são calculadas através da soma de quadrados, segundo as seguintes fontes de variação:

- entre as K classes;
- dentro das K classes;
- total das K classes.

Dado que a hipótese nula de igualdade das K médias tenha sido rejeitada, quer-se estudar o comportamento das mesmas através de sucessivas divisões dicotômicas, segundo a Técnica AID. Suponha que as K classes tenham sido divididas segundo uma partição d , composta de 2 grupos: A e B. Analogamente, as seguintes medidas de variação podem ser calculadas:

- soma de quadrados entre os grupos da partição d ;
- soma de quadrados dentro da partição d ;
- soma de quadrados total da partição d .

Além disso, com relação ao grupo A e ou B, podem ser obtidas:

- soma de quadrados entre as classe alocadas no grupo A ou B;
- soma de quadrados dentro de A ou B;
- soma de quadrado total de A ou B.

No caso de variável resposta categorizada, a situação análoga à análise de variância é o teste de homogeneidade entre amostras multinomiais independentes.

Considere uma tabela $K \times I$, onde cada linha representa o resultado de N_k , $k = 1, \dots, K$ ensaios multinomiais independentes de $\underline{X}$ com parâmetros $\underline{\pi}_k = (\pi_{k1} \ \pi_{k2} \ \dots \ \pi_{kI})$, onde $\underline{X} = (X_1, \dots, X_I)$ com $X_i = \begin{cases} 1 & \text{se observar resposta da categoria } i \\ 0 & \text{caso contrário} \end{cases} \quad i = 1, \dots, I$

Na tabela 3.1.1 está um exemplo típico de uma situação como a descrita.

TABELA 3.1.1 - Exemplo de uma situação de teste de homogeneidade de amostras multinomiais independentes.

		níveis da variável resposta				
		1	2	...	I	Total
classes da variável preditora	1	n_{11}	n_{12}	...	n_{1I}	N_1
	2	n_{21}	n_{22}	...	n_{2I}	N_2
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	K	n_{K1}	n_{K2}	...	n_{KI}	N_K
	Total	$n_{.1}$	$n_{.2}$	...	$n_{.I}$	N

n_{ki} = frequência observada do i -ésimo nível de resposta na k -ésima classe em N_k ensaios multinomiais.

$\underline{N}_k = (n_{k1} \ \dots \ n_{kI})$ tem distribuição multinomial com parâmetros

$$(N_k; \pi_{k1} \ \dots \ \pi_{kI}), \quad k = 1, \dots, K.$$

Semelhante às somas de quadrado na análise de variância, algumas medidas de variação da variável resposta podem ser calculadas como:

- variação entre as K classes;
- variação dentro das K classes;
- variação total das K classes.

Como no caso contínuo, a medida de variação serve para quantificar o quanto são homogêneas as observações, ou seja, valores "grandes" da medida de variação indicam que as respostas são heterogêneas e valores "pequenos" indicam que os dados são homogêneos.

Considere que a hipótese nula de homogeneidade entre as K classes tenha sido rejeitada e se quer estudar o comportamento das classes através de divisões dicotômicas, segundo a técnica AID.

Suponha que as K classes foram divididas segundo uma partição fixa d, composta de 2 grupos: A e B, tais que as classes numeradas de 1 a k' fazem parte do grupo A e as classes $k' + 1$ a K estão no grupo B. Nas Tabelas 3.1.2 e 3.1.3 estão ilustrados os grupos A e B compostos desta maneira.

TABELA 3.1.2 - Frequências observadas das K classes.

grupo	classes alocadas	níveis da variável resposta				Total
		1	2	...	I	
A	1	n_{11}	n_{12}	...	n_{1I}	N_1
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	k'	$n_{k'1}$	$n_{k'2}$	...	$n_{k'I}$	$N_{k'}$
	SOMA	n_{A1}	n_{A2}	...	n_{AI}	N_A
B	$k'+1$	$n_{k'+1,1}$	$n_{k'+1,2}$	...	$n_{k'+1,I}$	$N_{k'+1}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	K	n_{K1}	n_{K2}	...	n_{KI}	N_K
	SOMA	n_{B1}	n_{B2}	...	n_{BI}	N_B
Total		n_1	n_2	...	n_I	N

TABELA 3.1.3 - Frequências observadas após divisão dos grupos A e B.

níveis da variável resposta					
grupo	1	2	...	I	Total
A	n_{A1}	n_{A2}	...	n_{AI}	N_A
B	n_{B1}	n_{B2}	...	n_{BI}	N_B
Total	n_1	n_2	...	n_I	N

Nesta situação:

$$\begin{aligned} \tilde{N}_A &= (n_{A1}, \dots, n_{AI}) \text{ onde } N_A \sim M(N_A, \pi_{A1}, \dots, \pi_{AI}) \\ \tilde{N}_B &= (n_{B1}, \dots, n_{BI}) \text{ onde } N_B \sim M(N_B, \pi_{B1}, \dots, \pi_{BI}) \end{aligned}$$

Analogamente ao caso contínuo, as seguintes medidas de variação podem ser obtidas com relação à partição d:

- variação total da partição d;
- variação dentro da partição d;
- variação entre os grupos da partição d.

Além disso, outras medidas poderiam expressar com relação ao Grupo A ou B.

- variação total do grupo A ou B;
- variação dentro do grupo A ou B;
- variação entre as classes alocadas ao grupo A ou B.

Neste capítulo serão descritas as medidas de variação para partições fixas através das estatísticas de Light e Margolin (C), de Pearson (X), da razão de Máxima Verossimilhança (V), Theta (θ) e Delta (δ), nas situações anteriores.

3.2 Medidas de variação numa partição fixa com a estatística C.

Light e Margolin (1971) apresentaram uma análise de variância com um fator fixo para dados categorizados (CATANOVA). A hipótese de homogeneidade de resposta nas classes do fator é verificada através da estatística C que será introduzida neste item.

A construção das medidas de variação necessárias para CATANOVA fundamenta-se nos seguintes resultados apresentados por Gini (1938):

Teorema 3.2.1 :

A soma dos quadrados dos desvios da variável aleatória contínua X em relação à média de uma amostra de tamanho N pode ser expressa como uma função composta dos quadrados das diferenças dos N pares de todas as $\binom{N}{2}$ combinações possíveis, ou seja:

$$SST = \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^2 \quad \text{onde } d_{ij} = X_i - X_j$$

(3.2.1A)

Demonstração:

$$\frac{1}{2N} \sum_i \sum_j d_{ij}^2 = \frac{1}{2N} \sum_i \sum_j (X_i - X_j)^2 = \frac{1}{2N} \sum_i \sum_j (X_i^2 + X_j^2 - 2X_i X_j) =$$

$$\frac{1}{2N} \left(\sum_i \sum_j X_i^2 + \sum_i \sum_j X_j^2 - 2 \sum_i \sum_j X_i X_j \right) = \frac{1}{2N} \left(\sum_i N X_i^2 + \sum_j N X_j^2 - 2 \sum_i X_i \sum_j X_j \right) =$$

$$\frac{\sum_i X_i^2}{2} + \frac{\sum_j X_j^2}{2} - \frac{\sum_i X_i \sum_j X_j}{N}. \quad \text{Como } \sum_{i=1}^N X_i = \sum_{j=1}^N X_j, \text{ então:}$$

$$\frac{\sum_i X_i^2}{2} + \frac{\sum_j X_j^2}{2} - \frac{\sum_i X_i \sum_j X_j}{N} = \frac{\sum_{i=1}^N X_i^2}{2} - \frac{(\sum_i X_i)^2}{N} = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{2}$$

De modo análogo, Gini definiu uma função d_{ij} para dados categorizados do seguinte modo:

$$d_{ij} = \begin{cases} 1 & \text{se } \underline{X}_i \neq \underline{X}_j \\ 0 & \text{caso contrário} \end{cases} \quad i, j = 1, \dots, N \quad (3.2.1.B)$$

onde $\underline{X}_i$ e $\underline{X}_j$ são, respectivamente, os resultados do i -ésimo e j -ésimo ensaio multinomial de $\underline{X} = (X_1, \dots, X_1)$.

Considerando N ensaios multinomiais e utilizando a expressão 3.2.1.B em 3.2.1.A, obtém-se:

$$SST = \sum_{i=1}^N \sum_{j=1}^N \frac{1 d_{ij}^2}{2N} = \frac{N}{2} \left[1 - \sum_{i=1}^I p_i^2 \right] \quad (3.2.2)$$

onde n_i e n_j são as frequências observadas da i -ésima e j -ésima categoria de resposta em N ensaios multinomiais.

Demonstração:

$$SST = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^2 = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^N d_{ij} = \frac{1}{2N} \sum_{i \neq j} n_i n_j =$$

$$\frac{1}{2N} n_1(N-n_1) + n_2(N-n_2) + \dots + n_I(N-n_I) =$$

$$\frac{1}{2N} (N^2 - \sum_{i=1}^I n_i^2) = \frac{N}{2} - \frac{1}{2N} \sum_{i=1}^I n_i^2 = \frac{N}{2} \left[1 - \sum_{i=1}^I p_i^2 \right] =$$

onde $p_i = \frac{n_i}{N}$

Esta medida de variação para dados categorizados possui as seguintes propriedades:

1) Tem valor mínimo igual a zero e somente se todos os ensaios N ensaios multinomiais apresentarem um único nível de resposta, ou seja,

$$SST = 0 \Leftrightarrow \tilde{X}_i = \tilde{X}_j, \text{ qualquer } i, j = 1, \dots, N \quad (3.2.3)$$

2) O valor máximo de SST ocorre quando as frequências de todos os I níveis forem iguais, ou seja,

$$n_i = N/I \text{ qualquer } i = 1, \dots, I$$

Então:

$$\begin{aligned} SST &= \frac{N}{2} \left[1 - \sum_{i=1}^I p_i^2 \right] = \frac{N}{2} \left[1 - \sum_{i=1}^I \left(\frac{1}{I} \right)^2 \right] = \frac{N}{2} \left[1 - \frac{1}{I} \right] \\ &= \frac{N}{2} \left[\frac{I-1}{I} \right] \end{aligned} \quad (3.2.4)$$

Utilizando as propriedades 1 e 2, pode-se definir a estatística:

$$R^2 = SST / \max SST \quad (3.2.5)$$

que expressa o quanto a variabilidade do conjunto das respostas está próximo do seu valor máximo. Quanto mais próximo de 1 for o valor de R^2 , mais heterogêneas são as respostas e quanto mais próximo de zero for o valor de R^2 , mais homogêneas são as respostas.

Exemplo 3.2.1

Numa pesquisa sobre a preferência do consumidor por uma entre três marcas de sabonete, uma amostra de 600 pessoas apresentou os seguintes resultados:

Números de pessoas que preferem			
Marca A	Marca B	Marca C	Total
270	120	210	600

Neste caso:

$$SST = 600/2 - 1/1200 (270^2 + 120^2 + 210^2) = 190,80$$

O valor máximo que SST pode atingir neste caso é:

$$\max SST = \frac{600}{2} \times \frac{(2)}{3} = 200$$

Então $R^2 = 190,50/200 = 0,95$, o que sugere que o conjunto é bastante heterogêneo.

Suponha que, ao invés de se observar os resultados anteriores, tivessem sido observados os seguintes

Número de pessoas que preferem			
Marca A	Marca B	Marca C	Total
190	200	210	600

Nesta situação, teria

$$SST = \frac{600^2}{2} - \frac{1}{1200} (190^2 + 200^2 + 210^2) = 199,83334$$

O valor máximo que SST pode atingir, neste caso, é idêntico ao caso anterior.

Então $R^2 \cong 1,0$, o que sugere, neste caso, maior heterogeneidade das respostas do que no caso anterior.

Considere uma análise de variância clássica (com variável contínua) com um fator fixo (uma variável preditora) e usando o teorema de Gini, pode-se escrever:

$$SST(K) = \sum_{k=1}^K \sum_{j=1}^{N_k} (Y_{kj} - \bar{Y})^2 = \frac{1}{2N} \sum_{k=1}^K \sum_{j=1}^{N_k} \sum_{k'=1}^K \sum_{j'=1}^{N_{k'}} (Y_{kj} - Y_{k'j'})^2 =$$

$$= \frac{1}{2N} \sum_{k=1}^K \sum_{j=1}^{N_k} \sum_{k'=1}^K \sum_{j'=1}^{N_{k'}} d_{kjk'j'}^2 \quad \text{onde}$$

$$d_{kjk'j'} = Y_{kj} - Y_{k'j'}$$

(3.2.6)

Considere, agora, a situação ilustrada na Tabela 3.1.1. A expressão 3.2.1.B pode novamente ser adaptada, neste caso, para construir uma medida de variação total para dados categorizados, referente às respostas nas K classes, do seguinte modo:

$$SST(K) = \frac{1}{2N} \sum_{k=1}^K \sum_{j=1}^{N_k} \sum_{k'=1}^K \sum_{j'=1}^{N_{k'}} d_{kjk'j'}^2 = \frac{1}{2N} \sum_{k=1}^K \sum_{j=1}^{N_k} \sum_{k'=1}^K \sum_{j'=1}^{N_{k'}} \frac{d_{kjk'j'}}{2N}$$

$$= \frac{1}{2N} \sum_{k=1}^K \sum_{k'=1}^K \sum_{i \neq j} \sum_{i \neq j'} n_{ki} n_{k'j'}$$

$$\text{onde } d_{kjk'j'} = \begin{cases} 1 & \text{se } \tilde{x}_{kj} \neq \tilde{x}_{k'j'} \\ 0 & \text{se } \tilde{x}_{kj} = \tilde{x}_{k'j'} \end{cases} \quad \begin{matrix} j \text{ e } j' = 1, \dots, N_k \\ k \text{ e } k' = 1, \dots, K \end{matrix}$$

(3.2.7)

$\underline{X}_{kj}$ = é o j-ensaio multinomial da k-ésima classe da variável preditora.

$$\begin{aligned} SST(K) &= \frac{1}{2N} \left[n_{11}(N_1 - n_{11}) + n_{12}(N_2 - n_{21}) + \dots + n_{1I}(N_K - n_{K1}) + \right. \\ &\quad n_{12}(N_1 - n_{12}) + n_{12}(N_2 - n_{22}) + \dots + n_{2I}(N_1 - n_{11}) + \\ &\quad n_{21}(N_2 - n_{21}) + \dots + n_{2I}(N_K - n_{KI}) + \dots + n_{KI}(N_K - n_{KI}) \left. \right] = \\ &= \frac{1}{2N} \left[n_{11}(N - n_1) + n_{12}(N - n_2) + \dots + n_{KI}(N - n_I) \right] = \\ &= \frac{1}{2N} \left[n_1(N - n_1) + n_2(N - n_2) + \dots + n_I(N - n_I) \right] = \\ &= \frac{1}{2N} (N^2 - \sum_{i=1}^I n_i^2) = \frac{N}{2} - \sum_{i=1}^I \frac{n_i^2}{2N} = \frac{N}{2} \left[1 - \sum_{i=1}^I p_i^2 \right] \end{aligned}$$

$$\text{onde } p_i = \frac{n_i}{N}$$

Nota-se que o símbolo (K) agregado ao SST indica que a medida de variação está relacionada com todas as K classes da variável preditora. Este tipo de notação será utilizado para as demais medidas de variação.

Considerando agora uma determinada classe k; a expressão 3.2.1.B pode ser novamente adaptada, obtendo uma medida de variação dentro de uma classe k.

$$SSd(k) = \frac{1}{2N_k} \sum_{j=1}^{N_k} d_{kjkj'} \quad \text{onde } d_{kjkj'} = \begin{cases} 1 & \text{se } \underline{X}_{kj} \neq \underline{X}_{kj'} \\ 0 & \text{se } \underline{X}_{kj} = \underline{X}_{kj'} \end{cases} \quad (3.2.8)$$

$j \text{ e } j' = 1, \dots, N_k ; k, \text{ fixo}$

A medida de variação dentro das K classes, semelhantemente à da análise de variância clássica, pode ser expressa como:

$$SSD(K) = \sum_{k=1}^K SSd(k)$$

Então, tem-se:

$$\begin{aligned} SSD(K) &= \sum_{k=1}^K SSd(k) = \sum_{k=1}^K \frac{1}{2N_k} \left[N_k^2 - \sum_{i=1}^I n_{ki}^2 \right] = \frac{N}{2} - \frac{1}{2} \left(\sum_{k=1}^K \frac{1}{N_k} \sum_{i=1}^I n_{ki}^2 \right) \\ &= \frac{N}{2} - \frac{1}{2} \sum_{k=1}^K \sum_{i=1}^I n_{ki} p_i / k = \frac{N}{2} \left[1 - \sum_{k=1}^K w_k \sum_{i=1}^I p_i^2 / k \right] \end{aligned} \quad (3.2.9)$$

$$\text{onde } w_k = \frac{N_k}{N} \text{ e } p_i / k = \frac{n_{ki}}{N_k}$$

A medida de variação entre as K classes, pode ser obtida como a análise de variância clássica através de:

$$SSE(K) = SST(K) - SSD(K)$$

$$\begin{aligned} SSE(K) &= \frac{N}{2} - \frac{1}{2N} \left[\sum_{i=1}^I n_i^2 \right] + \frac{1}{2} \left[\sum_{k=1}^K \frac{1}{N_k} \sum_{i=1}^I n_{ki}^2 \right] - \frac{N}{2} = \\ &= \frac{1}{2} \left[\sum_{k=1}^K \sum_{i=1}^I n_{ki} p_i / k - \sum_{i=1}^I n_i p_i \right] \end{aligned} \quad (3.2.10)$$

$$SSE(K) = \frac{N}{2} \left[\sum_{k=1}^K w_k \sum_{i=1}^I (p_i^2 / k - p_i^2) \right]$$

A estatística C(K) utilizada para testar a hipótese de homogeneidade entre as K classes foi definida por Light e Margolin (1971) como:

$$C(K) = \frac{SSE(K) / (I-1)}{SST(K) / (N-1)} \quad (3.2.11)$$

que tem distribuição assintótica qui-quadrado com $(I-1) \times (K-1)$

graus de liberdade sob a hipótese nula.

Suponha agora a situação ilustrada na Tabela 3.1.3. Adaptando a expressão 3.2.1.B, pode-se obter as seguintes medidas de variação referentes à partição d:

1) variação total da partição d,

$$\begin{aligned} SST(d) &= \frac{N}{2} - \frac{1}{2N} \sum_{i=1}^I n_i^2 \\ &= \frac{N}{2} \left[1 - \sum_{i=1}^I p_i^2 \right] \end{aligned} \quad (3.2.12)$$

2) variação dentro da partição d,

$$\begin{aligned} SSD(d) &= \frac{N_A}{2} - \frac{1}{2N_A} \sum_{i=1}^I n_{Ai}^2 + \frac{N_B}{2} - \frac{1}{2N_B} \sum_{i=1}^I n_{Bi}^2 = \\ &= \frac{N_A + N_B}{2} - \frac{1}{2} \left[\frac{\sum n_{Ai}^2}{N_A} + \frac{\sum n_{Bi}^2}{N_B} \right] = \\ &= \frac{N}{2} \left[1 - \sum_{i=1}^I w_A p_i^2/A - \sum_{i=1}^I (1-w_A) p_i^2/B \right] \end{aligned} \quad (3.2.13)$$

onde $w_A = \frac{N_A}{N}$, $w_B = \frac{N_B}{N}$, $p_i/A = \frac{n_{Ai}}{N_A}$ e $p_i/B = \frac{n_{Bi}}{N_B}$

3) De modo semelhante àquele utilizado para obtenção de $SSE(K)$, a medida de variação entre os grupos da partição d, pode ser obtida calculando-se a diferença entre $SST(d)$ e $SSD(d)$.

$$\begin{aligned}
SSE(d) &= SST(d) - SSD(d) = \\
&= \frac{N}{2} \left[1 - \sum_{i=1}^I p_i^2 \right] - \frac{N}{2} \left[1 - \sum_{i=1}^I w_A p_i^2/A - \sum_{i=1}^I (1-w_A) p_i^2/B \right] \\
&= \frac{N}{2} \left[\sum_{i=1}^I w_A p_i^2/A + \sum_{i=1}^I (1-w_A) p_i^2/B - \sum_{i=1}^I p_i^2 \right]
\end{aligned} \tag{3.2.14}$$

Para testar a hipótese de homogeneidade dos 2 grupos (A e B), componentes da partição d, pode-se definir a estatística

$$C(d) = \frac{(I-1)SSE(d)}{SST(d)/(N-1)} \tag{3.2.15}$$

que tem distribuição assintótica $\chi^2_{(I-1)}$ sob a hipótese nula.

Além disso, pode-se obter informações sobre a variabilidade das respostas das classes que fazem parte do grupo A ou B, adaptando novamente a expressão 3.2.1.B. Por exemplo, para o grupo A têm-se

1) variação total do grupo A,

$$SSt(A) = \frac{N_A}{2} \left[1 - \sum_{i=1}^I p_i^2/A \right] \tag{3.2.16}$$

2) variação dentro do grupo A,

$$\begin{aligned}
SSd(A) &= \frac{N_A}{2} \left[1 - \sum_{k \in A} w_{k/A} \sum_{i=1}^I p_{i/k.A} \right] \quad \text{onde } w_{k/A} = \frac{N_k}{N_A} \quad k \in A \\
&\quad p_{i/k.A} = \frac{n_{ki}}{N_A}
\end{aligned} \tag{3.2.17}$$

3) como nos casos anteriores, a variação entre classes do grupo A pode ser obtida através da diferença entre $SSt(A)$ e $SSd(A)$

$$SSe(A) = SSt(A) - SSd(A)$$

(3.2.18)

O teste para verificar a homogeneidade das classes da variável preditora pertencentes ao grupo A pode utilizar a estatística

$$C(A) = \frac{(I-1)SSe(A)}{SSt(A)/(N-1)} \quad (3.2.19)$$

que tem distribuição assintótica χ^2 $(I-1)(\sum_{k \in A} I(k)-1)$

$$\text{onde } I(k) = \begin{cases} 1 & \text{se } k \in A \\ 0 & \text{caso contrário} \end{cases}$$

Medidas de variação e testes semelhantes podem ser deduzidos para o grupo B.

Uma estatística muito utilizada para avaliar a adequação da partição é o coeficiente de explicação

$$R^2 = SSE(d)/SST(K) \quad (3.2.20)$$

onde R^2 vai indicar a proporção da variabilidade total explicada através da partição d. Comparando 3.2.15, 3.2.20, verifica-se que as estatísticas $C(d)$ e R^2 se relacionam através da expressão:

$$C(d) = (I-1)(N-1)R^2, \quad (3.2.21)$$

que, em situações práticas, facilita os cálculos das estimativas.

Examinando-se as expressões 3.2.7, 3.2.9, 3.2.12, 3.2.13, 3.2.16 e 3.2.17, podem-se constatar as seguintes relações:

$$1) SST(K) = SST(d) \quad (3.2.22)$$

$$2) SSD(d) = SSt(A) + SSt(B) \quad (3.2.23)$$

$$3) SSD(K) = SSd(A) + SSd(B) \quad (3.2.24)$$

Das relações 3.2.23 e 3.2.24, pode-se verificar que:

$$\begin{aligned} SSD(d) &= SSt(A) + SSt(B) = \\ &SSd(A) + SSd(B) + SSe(A) + SSe(B) = \\ &SSD(K) + SSe(A) + SSe(B) \end{aligned} \quad (3.2.25)$$

Chamando $SSe(A) + SSe(B)$ de $SSR(d)$, tem-se que:

$$4) SSD(d) = SSD(K) + SSR(d) \quad (3.2.26)$$

De 3.2.22 e 3.2.26, tem-se:

$$\begin{aligned} SST(K) &= SST(d) \\ SSD(K) + SSE(K) &= SSD(d) + SSE(d) = \\ SSd(K) + SSE(K) &= SSD(K) + SSR(d) + SSE(d) \Rightarrow \\ SSE(K) &= SSR(d) + SSE(d) \end{aligned} \quad (3.2.27)$$

que relaciona as medidas de variação entre os grupos nas 2 situações.

Estas relações mostram que as medidas de variação através da estatística C se decompõem como na análise de variância clássica.

3.3 Medidas de variação numa partição fixa com as estatísticas de Pearson e da razão de verossimilhança

Considerando a situação ilustrada na tabela 3.1.1, as estatísticas de Pearson e da razão de verossimilhança para testar a hipótese de homogeneidade, tem-se as seguintes expressões, respectivamente:

$$XSE(K) = \sum_{k=1}^K \sum_{i=1}^I \frac{\left[n_{ki} - \frac{n_{ki} N_k}{N} \right]^2}{\frac{n_{ki} N_k}{N}} = \sum_{k=1}^K \sum_{i=1}^I N \frac{(p_{i/k} - p_i)^2}{p_i} \quad (3.3.1)$$

$$VSE(K) = 2 \sum_{k=1}^K \sum_{i=1}^I n_{ki} \ln \frac{p_{i/k}}{p_i} \quad (3.3.2)$$

$$\text{onde } p_i = \frac{N_i}{N} \quad \text{e} \quad p_{i/k} = \frac{n_{ki}}{N_k}$$

Pode-se verificar que ambas têm distribuição assintótica χ^2 com $(I-1)(K-1)$ graus de liberdade sob a hipótese de homogeneidade. Além disso, pode-se constatar que:

$$0 < XSE(K) < (T-1)N, \quad \text{onde } T = \min(K, I)$$

(Kendall & Stuart, vol.2, pg.588)

Supondo a situação ilustrada na Tabela 3.1.3, pode-se testar a homogeneidade entre os grupos obtidos segundo a partição d através das estatísticas de Pearson e da Razão de Verossimilhança,

cujas expressões são , respectivamente

$$XSE(d) = \sum_{i=1}^I \left[N_A \frac{(p_{i/A} - p_i)^2}{p_i} + N_B \frac{(p_{i/B} - p_i)^2}{p_i} \right] \quad (3.3.3)$$

$$VSE(d) = 2 \sum_{i=1}^I \left[n_{Ai} \ln \frac{p_{i/A}}{p_i} + n_{Bi} \ln \frac{p_{i/B}}{p_i} \right] \quad (3.3.4)$$

$$\text{onde } p_{i/A} = \frac{n_{Ai}}{N_A}, \quad p_{i/B} = \frac{n_{Bi}}{N_B}, \quad p_i = \frac{N_i}{N}$$

Pode-se comprovar que ambas têm distribuição assintótica χ^2 com $(I-1)$ graus de liberdade sob H_0 .

Ainda, nesta situação, pode-se avaliar a homogeneidade das classes dentro dos grupos A ou B. No caso do grupo A, resultaria:

$$XSe(A) = \sum_{k \in A} \sum_{i=1}^I N_k \frac{(p_{i/k} - p_{i/A})^2}{p_{i/A}} \quad (3.3.5)$$

se fosse utilizada a estatística de Pearson, e

$$VSe(A) = 2 \sum_{k \in A} \sum_{i=1}^I n_{ki} \ln \frac{p_{i/k}}{p_{i/A}} \quad p_{i/k} = \frac{n_{ki}}{N_k} \quad (3.3.6)$$

para a estatística de razão de verossimilhança.

Pode-se verificar que ambas têm distribuição assintótica χ^2 com $(\sum_{k \in A} I(k) - 1)(I - 1)$ graus de liberdade sob H_0 .

$$\text{onde } I(k) \begin{cases} = 1 & \text{se } k \in A \\ = 0 & \text{se } k \notin A \end{cases}$$

Testes semelhantes podem ser aplicados para o grupo B.

Chamando

$$XSR(d) = XSE(K) - XSE(d) \quad (3.3.7)$$

$$\text{e } VSR(d) = VSE(K) - VSE(d) \quad (3.3.8)$$

Constata-se que:

$$1) XSR(d) \neq XSe(A) + XSe(B)$$

$$2) VSR(d) = 2 \sum_{k=1}^K \left[\sum_{i=1}^I n_{ki} \ln \frac{p_i/k}{p_i} - \sum_{i=1}^I n_{Ai} \ln \frac{p_i/A}{p_i} - \sum_{i=1}^I n_{Bi} \ln \frac{p_i/B}{p_i} \right] =$$

$$= 2 \left[\sum_{k \in A} \sum_{i=1}^I n_{ki} \ln \frac{p_i/k}{p_i/A} + \sum_{k \in B} \sum_{i=1}^I n_{ki} \ln \frac{p_i/k}{p_i/B} \right]$$

$$VSR(d) = VSe(A) + VSe(B)$$

Estas expressões mostram que as estatísticas da razão de verossimilhança sob diferentes situações se relacionam, apresentando uma semelhança com a análise de variância clássica não se observando este comportamento para a estatística de Pearson.

3.4 Medidas de variação numa partição fixa com a estatística Theta (θ)

Cronologicamente, os primeiros métodos de procura de grupos ótimos por divisões dicotômicas em dados categorizados utilizaram

a estatística Theta (θ) que foi proposta por Messenger e Mendel (1972) e Morgan e Messenger (1973). A estatística está baseada na medida de associação Lambda (λ) apresentada por Kruskal e Goodman (1954) e suas características e propriedades serão discutidas em seguida.

Suponha que $\underline{N} = (n_1, n_2, \dots, n_I)$ seja o resultado observado de N ensaios multinomiais de parâmetros $(\pi_1, \dots, \pi_I)$ e que se deseje prever qual deve ser a resposta no próximo ensaio, dada a informação das frequências anteriores. Neste caso, a melhor previsão consiste em escolher o nível m , tal que:

$$n_m = \max \{n_1, n_2, \dots, n_I\} \quad (3.4.1)$$

ou seja, a melhor previsão corresponde à escolha do nível modal da variável multinomial $\underline{N}$.

$$\text{Seja } \theta_m = n_m / N \quad (3.4.2)$$

a frequência relativa do nível modal.

Pode-se verificar que vale a seguinte desigualdade para

$$\frac{1}{I} \leq \theta \leq 1 \quad (3.4.3)$$

Quando mais próximo θ estiver de 1, mais homogêneas são as respostas, e quando mais próximo estiver de $1/I$, mais heterogêneas são as respostas.

$$\text{Seja } \theta_T = 1 - \theta \quad (3.4.4)$$

Há duas denominações para θ_T . A primeira, erro de previsão,

é atribuída a Goodman e Kruskall (1954) e a segunda, medida de heterogeneidade total, por Fielding (1977). Intuitivamente, valores "grandes" de θ_T podem ser interpretados como "grandes erros de previsão" no enfoque de Goodman e Kruskall (1954), ou como "dados não homogêneos" no enfoque de Fielding (1977).

Exemplo 3.4.1 - Utilizando-se os mesmos dados do exemplo 3.2.1 onde numa amostra de 600 pessoas, 270 preferiram marca A, 120 a marca B e 210 a marca C, os valores de θ e θ_T são obtidos.

$$\theta = 270/600 = 9/20 \text{ e } \theta_T = 11/20 = 0,55$$

Se fosse realizado mais um ensaio nos mesmos moldes, "provavelmente" a marca preferida seria a marca A (nível modal), com 55% de "erro de previsão". Pode-se interpretar este resultado como sendo um indicador de que os dados não são "muito homogêneos".

Considere, agora, a situação ilustrada na tabela 3.1.1 e suponha que se deseja prever qual deva ser a resposta de um novo ensaio multinomial, dadas as informações contidas na tabela. Focalizando o problema para uma determinada classe k da variável preditora, a melhor previsão nesta classe corresponde à escolha do nível m' , tal que:

$$n_{km'} = \max \{ n_{k1} \ n_{k2} \ \dots \ n_{kl} \} \quad (3.4.5)$$

ou seja, a melhor previsão corresponde à escolha do nível modal em cada classe k .

$$\text{Seja } \theta_k = n_{km'} / N_x \quad (3.4.6)$$

a frequência relativa do nível modal da classe k . Intuitivamente, para que a variável preditora "auxilie" na realização de uma "boa" previsão; os valores θ_k individuais, $k = 1, \dots, K$ devem ser grandes, ou ainda, os θ_k 's sendo "grandes" implicam que as respostas são "homogêneas" dentro das K classes.

$$\text{Seja } \theta(K) = \sum_{k=1}^K w_k \theta_k \quad \text{onde } w_k = \frac{N_k}{N} \quad (3.4.7)$$

a frequência relativa ponderada dos níveis modais da variável preditora, onde pesos w_k se referem ao tamanho das classes.

$$\text{Seja } \theta_D(K) = 1 - \theta(K) \quad (3.4.8)$$

Como no caso de θ_T , há duas denominações $\theta_D(K)$:

- erro de previsão dada a variável preditora é atribuída a Goodman e Kruskal (1954);

- medida de heterogeneidade dentro das K classes - por Fielding (1977). Similarmente, valores grandes de $\theta_D(K)$ podem ser interpretados como dados não homogêneos dentro das classes da variável preditora.

A segunda denominação utiliza a linguagem de análise de variância. Fazendo analogia, pode-se definir uma medida de heterogeneidade entre as classes através da diferença:

$$\theta_E(K) = \theta_T - \theta_D(K) \quad (3.4.9)$$

onde θ_T é calculada com as frequências marginais, conforme a expressão (3.4.4).

Goodman e Kruskal (1954) definiram uma medida de associação denominada Lambda, como:

$$\lambda(K) = \theta E(K) / \theta T \quad (3.4.10)$$

e a interpreta como sendo a redução relativa do erro de previsão dada a variável preditora, ou melhor, mede a capacidade preditiva da variável preditora. Além disso, verificaram que

$$\sqrt{N} |\hat{\lambda}(K) - \lambda(K)| \sim N(0, V) \quad (3.4.11)$$

$$\text{onde } V = \frac{\theta D(K) * [\theta(K) + \theta - 2\rho]}{(\theta T)^3}$$

$$\rho = \sum_{k=1}^K \bar{w}_k C_k \quad \text{com} \quad c_k \begin{cases} = \theta_k & \text{se } m' = m \\ = 0 & \text{caso contrário} \end{cases}$$

Na linguagem de análise de variância, $\lambda(K)$ corresponde ao ² coeficiente de explicação (R).

Exemplo 3.4.2 - Utilizando os dados do Exemplo 3.4.1, suponha que a preferência de 3 marcas de sabonete (A, B e C) foi controlada por região geográfica (SP, GSP e Interior), obtendo-se os resultados apresentados na Tabela 3.4.1.

TABELA 3.4.1 - Preferência de marcas de sabonete por região geográfica.

Região \ Marca	A	B	C	Total
SP	100	10	90	200
GSP	80	100	20	200
Interior	90	10	100	200
Total	270	120	120	600

Pode-se calcular:

$$\theta(K) = \frac{200 \times \frac{100}{200} + 200 \times \frac{100}{200} + 200 \times \frac{100}{200}}{600} = 0,5 \Rightarrow$$

$$\theta_D(K) = 1 - \theta(K) = 0,5$$

$$\theta = 270/600 \Rightarrow \theta_T = 330/600$$

$$\text{Como } \theta_E(K) = \theta_T - \theta_D(K) = 330/600 - 0,5 = 30/600$$

$$\lambda(K) = 30/330 = 1/11 \approx 0.09$$

A variável preditora reduz em 9% o erro de previsão.

Suponha agora a situação ilustrada na Tabela 3.1.3, e que se deseja prever a resposta no novo ensaio multinomial, dadas as informações apresentadas na tabela citada. A melhor previsão para o grupo A corresponde à escolha do nível m' , tal que

$$n_{Am'} = \max (n_{A1}, n_{A2}, \dots, n_{AI}) \quad (3.4.12)$$

ou seja, selecionando o nível modal do grupo A.

Sejam $\theta_A = n_{Am}/N_A$ e $\theta_B = n_{Bm}/N_B$
as frequências relativas dos níveis modais dos grupos A e B,
respectivamente.

$$\text{Seja } \theta(d) = w_A \theta_A + w_B \theta_B \text{ onde } \begin{cases} w_A = N_A/N \\ w_B = N_B/N \end{cases} \text{ e}$$

a frequência relativa ponderada dos níveis modais da partição fixa d, que é composta dos grupos A e B.

$$\text{Seja } \theta_D(d) = 1 - \theta(d) \quad (3.4.14)$$

Segundo Fielding (1977), a expressão 3.4.14 mede a heterogeneidade dos grupos componentes da partição fixa d. Utilizando os princípios da análise de variância clássica, a medida de heterogeneidade entre os grupos pertencentes à partição d pode ser definida como:

$$\theta_E(d) = \theta_T - \theta_D(d) \quad (3.4.15)$$

e como consequência a estatística

$$\lambda(d) = \frac{\theta_E(d)}{\theta_T} \quad (3.4.16)$$

mede a redução relativa do erro de previsão dada a partição d.

Além destas medidas, outras podem ser obtidas para avaliar as classes alocadas no grupo A ou B. Se fosse para avaliar o Grupo A, as seguintes medidas poderiam ser definidas:

$$a) \theta_T(A) = 1 - \theta_A \quad (3.4.17)$$

mede a heterogeneidade total do grupo A.

$$b) \theta_D(A) = 1 - \theta(A) \quad (3.4.18)$$

mede a heterogeneidade dentro do Grupo A com

$$c) \theta_E(A) = \theta_T(A) - \theta_D(A) \quad \left[\begin{array}{l} \theta(A) = \sum_{k \in A} w_{k/A} \theta_k \\ w_{k/A} = \frac{N_k}{N_A} \end{array} \right. \quad (3.4.19)$$

mede a heterogeneidade entre as classes alocadas no Grupo A.

$$d) \lambda(A) = \frac{\theta_E(A)}{\theta_T(A)} \quad (3.4.20)$$

mede a redução relativa do erro de previsão do grupo A, dadas as classes alocadas neste grupo.

De modo semelhante, podem ser obtidas expressões para avaliar a heterogeneidade do grupo B.

Das expressões 3.4.17, 3.4.18, 3.4.14, 3.4.8, 3.4.9, podem-se verificar algumas relações:

$$1) \theta_D(A) + \theta_D(B) = 1 + \theta_D(K) \quad (3.4.21)$$

Demonstração:

$$\begin{aligned} \theta_D(A) + \theta_D(B) &= 2 - \sum_{k \in A} w_{k/A} \theta_k - \sum_{k \in B} w_{k/B} \theta_k \\ &= 2 - \sum_{k=1}^K w_k \theta_k = 1 + 1 - \sum_{k=1}^K w_k \theta_k \\ &= 1 + \theta_D(K) \\ 2) \theta_D(d) &= w_A \theta_T(A) + w_B \theta_T(B) \\ \theta_D(d) &= 1 - \theta(d) = 1 - w_A \theta_A - w_B \theta_B \quad (c) \end{aligned} \quad (3.4.22)$$

$$\text{Como } \theta_T(A) = 1 - \theta_A \Rightarrow \theta_A = 1 - \theta_T(A) \quad (a)$$

$$\text{Do mesmo modo } \theta_B = 1 - \theta_T(B) \quad (b)$$

Substituindo (a), (b) em (c), obtém-se:

$$\Theta D(d) = 1 - W_A [1 - \Theta T(A)] - W_B [1 - \Theta T(B)] =$$

$$1 - W_A - W_B + W_A \Theta T(A) + W_B \Theta T(B) = W_A \Theta T(A) + W_B \Theta T(B)$$

$$3) \Theta E(K) = \Theta E(d) + \Theta R(d) \text{ onde}$$

$$\Theta R(d) = W_A \Theta E(A) + W_B \Theta E(B) \quad (3.4.23)$$

Demonstração:

$$\Theta T = \Theta D(K) + \Theta E(K) = \Theta D(d) + \Theta E(d)$$

Substituindo $\Theta D(d)$ por 3.4.22, resulta

$$\Theta T = \Theta E(d) + W_A \Theta T(A) + W_B \Theta T(B) \quad (f)$$

$$\text{Sabe-se que } \Theta T(A) = \Theta D(A) + \Theta E(A) \text{ e} \quad (d)$$

$$\Theta T(B) = \Theta D(B) + \Theta E(B) \quad (e)$$

Substituindo (d), (e) em (f), resulta

$$\Theta T = \Theta E(d) + W_A \Theta D(A) + W_B \Theta D(B) + W_A \Theta E(A) + W_B \Theta E(B) \quad (g)$$

$$\text{Como } \Theta D(A) = 1 - \sum_{k \in A} W_{k/A} \theta_k \text{ e} \quad (h)$$

$$\Theta D(B) = 1 - \sum_{k \in B} W_{k/B} \theta_k \quad (i)$$

$$\text{Substituindo (h) e (i) em (g), resulta } \Theta T = \Theta E(d) + W_A \Theta E(A)$$

$$+ W_B \Theta E(B) + \Theta D(K) = \Theta E(K) + \Theta D(K) \Rightarrow$$

$$\Theta E(K) = \Theta E(d) + W_A \Theta E(A) + W_B \Theta E(B) \Rightarrow$$

$$\Theta E(K) = \Theta E(d) + \Theta R(d)$$

Estas relações mostram que as medidas definidas a partir da estatística θ se comportam como as somas de quadrados na análise de variância clássica.

3.5 Medidas de variação numa partição fixa com a estatística δ

Esta estatística é uma das opções para medir variações em dados categorizados. Foi apresentada por Morgan e Messenger (1973) e baseia-se no critério da distância em valor absoluto. De modo genérico $\delta = \sum_i |O_i - E_i|$, onde O_i = valor observado e E_i = valor esperado segundo a hipótese de homogeneidade das K classes.

Considere agora a situação ilustrada na Tabela 3.1.1.

A estatística correspondente, neste caso, será dada por:

$$\delta(K) = \sum_{k=1}^K \sum_{i=1}^I N_k |p_{i/k} - p_i| \quad (3.5.1)$$

$$\text{onde } p_i = \frac{n_i}{N}$$

$$p_{i/k} = \frac{n_{ki}}{N_k}$$

que possui a seguinte propriedade:

$$0 < \delta(K) < 2N * \left[1 - \sum_{i=1}^I p_i^2 \right] \quad (3.5.2)$$

$$\text{onde } p_i = \frac{n_i}{N}$$

$\delta(K)$ atinge seu valor mínimo se $P_{i/k} = P_i$, qualquer K , implicando em homogeneidade total entre as K classes. Por outro lado, se $P_{i/k} = 1$ para algum $i = 1, \dots, I$, $\delta(K)$ vai alcançar seu valor máximo, indicando total discrepância entre as K classes (Morgan e Messenger, 1973). Dividindo a desigualdade expressa em 3.5.2 pelo seu máximo, pode-se definir a estatística $\delta'(K)$, que expressa o quanto a variação dos dados está próxima do seu valor máximo.

$$0 < \delta'(K) < 1 \quad (3.5.3)$$

Suponha agora a situação ilustrada na Tabela 3.1.3. A medida de homogeneidade da partição d é dada pela expressão:

$$\delta(d) = \sum_{j=A,B} \sum_{i=1}^I N_j \left| p_{i/j} - p_i \right| \quad (3.5.4)$$

Além disso, pode-se avaliar a homogeneidade das classes alocadas nos grupos A ou B através da estatística δ . Para o grupo A, a expressão correspondente é:

$$\delta(A) = \sum_{k \in A} \sum_{i=1}^I N_k \left| p_{i/k} - p_{i/A} \right| \quad (3.5.5)$$

Expressão semelhante pode ser escrita para o grupo B.

4 MEDIDAS DE VARIAÇÕES PARA PARTIÇÕES ÓTIMAS

4.1 Introdução

Neste capítulo, será apresentado o modo de escolher a partição ótima entre as alternativas contidas em $\Phi(J, \text{gama})$ (no item 4.2) segundo as seguintes medidas: C, X, V, θ e δ descritas no capítulo 3. No item 4.3 serão discutidos os critérios para avaliar se a divisão, conforme a partição ótima, é "lucrativa". E, finalmente, no item 4.4, será apresentada uma aplicação do procedimento da escolha da partição ótima e sua respectiva avaliação de significância num exemplo simples.

4.2 Seleção da partição ótima

Conforme foi mencionado no capítulo 2, cada problema tem associado um conjunto de partições gama-dicotômicas $\Phi(J, \text{gama})$. Assim, cada elemento d deste conjunto tem associado um valor $E(d)$, que pode ser obtido segundo uma das seguintes estatísticas: $C(d)$, $XSE(d)$, $VSE(d)$, $\theta E(d)$ e $\delta(d)$ expressos em 3.2.15, 3.3.3, 3.3.4, 3.4.15 e 3.5.4, respectivamente.

$$\text{Seja } B(J, \text{gama}) = \max \{E(d), d \in \Phi(J, \text{gama})\} \quad (4.2.1)$$

A melhor partição segundo o critério de seleção pré-fixado será a partição d_0 tal que satisfaça a igualdade:

$$E(d_0) = B(J, \text{gama}) \quad (4.2.2)$$

Selecionada a partição ótima, o próximo passo é avaliar se

ela é "vantajosa" ou não, o que será descrito no próximo item.

4.3 Critérios para avaliar a significância da partição ótima

Avaliar a significância de $B(J, \gamma)$ é verificar o quanto a partição é "vantajosa", (o quanto "compensa" dividir), equivalendo a testar a seguinte hipótese:

$$H_0: \pi_A = \pi_B = \pi$$

$$H_1: \begin{cases} \pi_k = \pi_A, & k \in A \text{ tal que } A \cup B = J \\ & A \cap B = \emptyset \\ \pi_k = \pi_B, & k \in B \end{cases}$$

A e B desconhecidos, porém "ótimos"

Como $E(d_0)$ é uma estatística de ordem de $E(d)$, $d \in \Phi(J, \gamma)$, sua distribuição sob hipótese nula não deve ser a mesma que a de $E(d)$ sob hipótese nula, embora esse critério possa ser empregado como um "limite inferior", de $E(d_0)$.

A distribuição de $E(d_0)$ não é conhecida e sua significância pode ser avaliada através de aproximações, sendo que as mais comuns são:

1) Desigualdade de Bonferroni

Seja α_0 e α_j os níveis de confiança para tentar $E(d_0)$ e $E(d_j)$, $j = 1, \dots, D$, respectivamente, utilizando-se a desigualdade de Bonferroni (Hawkins e Kass, 1980), prova-se que:

$$\text{nível de confiança de } E(d_0) < \sum_{j=1}^D \text{nível de confiança de } E(d_j) \Rightarrow$$

$$\alpha < D * \alpha_0 \Rightarrow \alpha/D < \alpha_0 \quad (4.3.1)$$

Da desigualdade 4.3.1, pode-se estabelecer a seguinte regra de decisão:

Se $E(d)$ for maior que os valores críticos da distribuição de $E(d)$ sob H_0 , a nível α/D , então rejeita-se a hipótese nula.

Os valores de D são conhecidos como Multiplicadores de Bonferroni e dependem da regra de partição. Na Tabela 4.3.1, estão as expressões para obtenção desse valor.

TABELA 4.3.1 - Multiplicador de Bonferroni.

Regra de partição	D
Monotônica	$K - 1$
Flutuante	$2K - 3$
Livre	$K - 1$
	$2 - 1$

2) Comparações Múltiplas

Seja $E(K)$ o valor associado à medida de variação entre as K classes, para um critério de seleção pré-estabelecido; no capítulo 3, está demonstrado que vale a seguinte desigualdade:

$$E(d) < E(K) \quad \text{para todo } d \quad (4.3.2)$$

conforme expressões 3.2.27, 3.3.7, 3.3.8, 3.4.23, 3.5.4

Logo:

$$E(d) < E(d_0) < E(K) \quad (4.3.3)$$

Através da desigualdade 4.3.2, estabelece-se a seguinte regra de decisão:

Se $E(d_0)$ for maior que os valores críticos da distribuição de $E(K)$ sob H_0 , a nível α então rejeita-se a hipótese nula.

Ressalta-se que caso se utilize a estatística θ para seleccionar partição ótima, utilizam-se as distribuições assintóticas de $\lambda(d)$ e de $\lambda(K)$, respectivamente, para os critérios de Desigualdade de Bonferroni e Comparações Múltiplas, visto não ser conhecida a distribuição de $\theta E(d)$ e $\theta E(K)$.

3) Coeficiente de explicação

É um critério empírico utilizado para verificar a significância de $E(d_0)$. Fixa-se uma percentagem arbitrária do coeficiente de explicação $R^2_* = E(d_0)/\text{medida de variação total}$.

Se a partição d_0 seleccionada apresentar um valor de R^2_* superior ao valor fixado, ela é considerada significativa.

Ressalta-se que, em algumas estatísticas, não são definidas medidas de variação total, como X , V e δ . No caso de se utilizar este critério empírico, pode-se construir um outro coeficiente de explicação, do seguinte modo:

$$R^2_* = E(d_0)/\text{medida de variação entre } K \text{ classes}$$

Esse coeficiente pode ser utilizado como o coeficiente anterior.

4.4 - Um exemplo

Será selecionada a melhor partição dicotômica do exemplo 3.4.2, cujos dados estão reproduzidos na Tabela 4.4.1:

TABELA 4.4.1 - Preferência de marca de sabonete por região geográfica.

Região \ Marca	Marca			Total
	A	B	C	
SP	100	10	90	200
GSP	80	100	20	200
Interior	90	10	100	200
	270	120	210	600

Na Tabela 4.4.2 estão as partições dicotômicas possíveis, obedecendo a uma regra livre, bem como algumas das estatísticas discutidas acima.

TABELA 4.4.2 - Partições possíveis do exemplo 4.4.1 segundo uma regra livre e estatísticas associadas.

Partição	Componentes da partição d		Estatísticas					
			de Light e Margolin	de Pearson	do teste de Verossim.	Theta	Delta	
	A	B	C(d)	XSE(d)	VSE(d)	$\theta E(d)$	$\lambda(d)$	$\delta(d)$
d ₁	SP + GSP	Int.	42,45	53,04	68,60	10/600	1/33	120
d ₂	SP + Int.	GSP	146,21*	70,24*	195,44*	20/600*	2/33*	180*
d ₃	GPS + Int.	SP	33,05	43,99	52,21	0,00	0	120

A melhor partição dicotômica segundo qualquer um dos critérios estudados (C, X, V, θ e δ), é a partição d₂, visto que corresponde ao máximo de E(d). (Na Tabela 4.4.2, estão assinalados com *).

Para avaliar se a partição d₂ é "vantajosa" ou não, serão utilizados os critérios citados: Desigualdade de Bonferroni e Comparações Múltiplas. Como as distribuições assintóticas de C(d), XSE(d) e VSE(d) sob H₀ são as mesmas, no caso de se utilizar qualquer um dos 3 critérios o valor crítico será o mesmo.

TABELA 4.4.3 - Valores críticos das estatísticas C₀, X₀ e V₀ para partição d₂, $\alpha = 0,05$.

Critério	Estatística C ₀ , X ₀ e V ₀
Desigualdade de Bonferroni	$X(2, \alpha/3) = 8,175$
Comparações Múltiplas	$X(4, \alpha) = 9,488$

Caso seja utilizada a estatística θ como critério de seleção da partição ótima, a avaliação da significância será feita através do Intervalo de Confiança de λ , construído a nível α .

Critérios	Valor crítico da distr.assoc.	IC p/λ a nível α
Desigualdade de Bonferroni	^a $Z(\alpha/3)$	0; 0.144
Comparação Múltipla	^b $Z(\alpha)$	0; 0.169

onde $Z^a = \sqrt{N} (\hat{\lambda}(d) - \lambda(d)) \sim N(0, V^a)$

$$V^a = \frac{\theta D(d) * (\theta(d) + \theta - 2\rho)}{(\theta T)^3} \quad \rho = \sum_j W_j c_j \quad j = A, B$$

$$Z^b = \sqrt{N} (\hat{\lambda}(K) - \lambda(K)) \sim N(0, V^b)$$

$$V^b = \frac{\theta D(K) * (\theta(K) + \theta - 2\rho)}{(\theta T)^3} \quad \rho = \sum_{k=1}^K W_k c_k$$

Desta forma, dentro das possíveis partições, a melhor é a d_2 , obtendo os seguintes resultados.

TABELA 4.4.5 - Resultado obtido após aplicação da Técnica AID.

Grupos \ Marca				Total
	A	B	C	
SP + Interior	190	20	190	400
GSP	80	100	20	200
Total	270	120	210	600

5 SIMULAÇÕES

5.0 Introdução

Este capítulo tem como objetivo comparar as várias estatísticas utilizadas para selecionar a partição ótima. No item 5.1 serão descritas as comparações teóricas entre C , X e V , realizadas por Light e Margolin (1971), para uma partição fixa. No item 5.2 estão descritos os resultados obtidos através de simulação de Light e Margolin (1971) e Morgan e Messenger (1973), ainda para uma partição fixa. A primeira compara as estatísticas C com X enquanto que a segunda calcula as correlações de posto entre os vários pares das estatísticas C , X , V , θ e δ .

Complementando as comparações, foram realizadas outras simulações visando estudar o comportamento de $C(d)$, $XSE(d)$ e $VSE(d)$ que no caso serão referenciadas por C_0 , X_0 e V_0 , respectivamente. A seleção destas estatísticas se prende ao fato de C , X e V apresentarem a mesma distribuição assintótica e algumas propriedades comuns. Nos itens 5.3.1 a 5.3.6 estão descritos os resultados empíricos destas simulações sob diversas condições, bem como as correlações lineares entre os vários pares das estatísticas.

No item 5.4 estão descritos os resultados obtidos por Morgan e Messenger (1973) sobre o comportamento da estatística $\theta E(d)$ empregada iterativamente na procura de grupos ótimos, com várias variáveis preditoras.

As simulações mencionadas requerem a geração de tabelas $K \times I$, onde em cada linha está apresentado o resultado observado

de N_k ensaios multinomiais de parâmetros $\underline{\pi}_k = (\pi_{k1}, \dots, \pi_{kl})$.

A maioria das simulações é realizada sob a hipótese de igualdade dos parâmetros $\underline{\pi}_k = \underline{\pi}$. Estabelecido este parâmetro e os tamanhos $N_1, N_2, \dots, N_K$, é possível gerar tabelas $K \times I$ e calcular as estatísticas pertinentes. Desse modo, a cada simulação é necessário fixar:

K - número de classes da variável preditora

I - número de resposta da variável resposta

$\underline{\pi}$ - vetor de probabilidade da distribuição multinomial

$\underline{N} = (N_1, N_2, \dots, N_K)$

R - número de tabelas geradas (réplicas)

A geração das tabelas $K \times I$ foi efetuada através de sub-rotina GMULT do pacote IMSL, implantado no CCE-USP.

5.1 Alguns resultados teóricos

As estatísticas X e C , expressas em 3.3.1 e 3.2.11, podem ser reescritas como :

$$X = \sum_{i=1}^I \frac{N}{n_i} \sum_{k=1}^K \left(n_{ki} / \sqrt{N_k} - \frac{n_i \sqrt{N_k}}{N} \right)^2 \quad (5.1.1)$$

$$C = \sum_{i=1}^I \frac{(I-1)(N-1)}{2 * SST(K)} \sum_{k=1}^K \left[n_{ki} / \sqrt{N_k} - \frac{n_i \sqrt{N_k}}{N} \right]^2 \quad (5.1.2)$$

Comparando 5.1.1 com 5.1.2, observa-se que ambos são de

forma:

$$\sum_{i=1}^I w_i \sum_{k=1}^K \left[n_{ki} / \sqrt{N_k} - \frac{n_i \sqrt{N_k}}{N} \right]^2 \quad (5.1.3)$$

com $w_i = \frac{N}{n_i}$ em (5.1.1)

$w_i = \frac{(I-1)(N-1)}{2 * SST(K)}$ em (5.1.2)

Light e Margolin (1971) demonstraram que para $I = 2$, as duas estatísticas se relacionam através da expressão:

$$X = \frac{N}{N-1} C \quad (5.1.4)$$

constatando que, para o caso de variáveis respostas dicotômicas, os valores destas 2 estatísticas são bastante próximos.

Outro resultado importante demonstrado por Margolin e Light (1974) relaciona as estatísticas V e X e, que pode ser resumido no seguinte teorema:

Teorema 5.1.1 : Para qualquer tabela $2 \times I$, com $N_1 = N_2 = N/2$ e $n_i \neq 0$ para qualquer i , vale a seguinte relação:

$$(2 \ln 2) X > V > X \quad (5.1.5)$$

A demonstração deste teorema encontra-se no apêndice.

A desigualdade 5.1.5 possibilita achar o limite inferior do nível descritivo de V nesta situação. O valor da estatística V será sempre maior que o de X e, conseqüentemente, terá sempre um nível descritivo menor que o da estatística de X , nestas condições.

5.2 Simulações para partições fixas

A simulação realizada por Light e Margolin (1971) tem como finalidade comparar as estatísticas C e X com suas distribuições assintóticas para $I > 2$. Neste caso, ambos têm a mesma distribuição χ^2 com $(K-1) \times (I-1)$ graus de liberdade, sob H_0 .

Os parâmetros utilizados nesta simulação estão indicados na

Tabela 5.2.1, enquanto que na Tabela 5.2.2 estão apresentadas algumas medidas de interesse, tais como média, variância, quantis a 90%, 95% e 99% e a correlação de postos entre C e X.

TABELA 5.2.1 - Parâmetros utilizados na simulação de Light e Margolin (1971).

I	$\pi_{k1}, \dots, \pi_{kI}$	K	N _k
3	1/3 1/3 1/3	2	20, 40
		2	30, 30
		2	30, 100
		2	100, 100
		4	30, 30, 30, 30
	.70 .15 .15	2	20, 40
		2	30, 30
		2	60, 60
		2	100, 100
	1/2 1/4 1/4	2	15, 15
	.90 .05 .05	2	20, 40
	.94 .03 .03	2	80, 80
	.48 .48 .04	2	80, 80
	.30 .30 .40	2	50, 50
		3	50, 50, 50
4	1/5 1/5 1/5 2/5	5	50, 50, 50, 50
		2	50, 50
		3	80, 80, 80
5	1/5 1/5 1/5 1/5 1/5	4	80, 80, 80, 80
		2	50, 50
		3	100, 100, 100
6	.16 .16 .16 .16 .16 .20	4	100, 100, 100, 100
		2	100, 100
7	.10 .10 .10 .10 .10	2	100, 100, 100,
	.10 .30	6	100, 100, 100

TABELA 5.2.2 - Estatísticas descritivas obtidas através de simulação de Light e Margolin (1971).

Simulation number	π	K	n_k	Mean	Variance	Percentiles			Correlação de postos X e C
						Reference percentile	Empirical critical value		
1	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	2	100 all	2.000 ^a 1.973 ^b 1.982 ^c	4.000 ^a 3.766 ^b 3.758 ^c	.90 .95 .99	.890 .944 .990	4.440 5.763 9.238	.972
2	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	2	30 all	2.000 ^a 1.994 ^b 2.027 ^c	4.000 ^a 3.889 ^b 3.918 ^c	.90 .95 .99	.890 .951 .989	4.439 6.034 9.092	.975
3	(.70, .15 .15)	2	30 all	2.000 ^a 2.027 ^b 2.077 ^c	4.000 ^a 4.160 ^b 3.826 ^c	.90 .95 .99	.905 .955 .991	4.720 6.211 9.463	.954
4	(.70, .15 .15)	2	60 all	2.000 ^a 1.951 ^b 1.963 ^c	4.000 ^a 3.940 ^b 3.380 ^c	.90 .95 .99	.890 .947 .992	4.433 5.870 9.854	.963
5	(.70, .15 .15)	2	100 all	2.000 ^a 1.951 ^b 1.963 ^c	4.000 ^a 3.940 ^b 3.380 ^c	.90 .95 .99	.900 .953 .990	4.606 6.149 9.310	.961
6	$(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$	2	15 all	2.000 ^a 2.024 ^b 2.078 ^c	4.000 ^a 4.224 ^b 3.911 ^c	.90 .95 .99	.895 .956 .992	4.528 6.265 9.770	.952
7	(.94, .03, .03)	2	80 all	2.000 ^a 1.986 ^b 2.013 ^c	4.000 ^a 3.989 ^b 3.033 ^c	.90 .95 .99	.904 .953 .990	4.710 6.166 9.126	.961
8	(.48, .48 .04)	2	80 all	2.000 ^a 2.020 ^b 2.101 ^c	4.000 ^a 5.498 ^b 3.399 ^c	.90 .95 .99	.916 .968 .995	4.973 6.970 10.619	.948
9	(.30, .30 .40)	2	50 all	2.000 ^a 1.889 ^b 1.929 ^c	4.000 ^a 3.562 ^b 3.752 ^c	.90 .95 .99	.895 .942 .982	4.496 5.671 8.085	.961
10	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	2	20, 40	2.000 ^a 1.954 ^b 1.989 ^c	4.000 ^a 3.822 ^b 3.973 ^c	.90 .95 .99	.896 .952 .988	4.569 6.117 9.019	.922
11	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	2	30, 100	2.000 ^a 1.940 ^b 1.957 ^c	4.000 ^a 3.602 ^b 3.587 ^c	.90 .95 .99	.890 .938 .986	4.422 5.447 8.688	.912
12	(.70, .15 .15)	2	20, 40	2.000 ^a 2.029 ^b 2.051 ^c	4.000 ^a 4.061 ^b 3.446 ^c	.90 .95 .99	.900 .945 .990	4.618 5.805 9.306	.935
13	(.90, .05 .05)	2	20, 40	2.000 ^a 2.014 ^b 2.092 ^c	4.000 ^a 4.024 ^b 3.867 ^c	.90 .95 .99	.902 .954 .970	4.642 6.182 9.344	.926
14	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	2	50 all	3.000 ^a 2.919 ^b 2.965 ^c	6.000 ^a 6.095 ^b 5.725 ^c	.90 .95 .99	.901 .947 .989	6.298 7.684 11.180	.942
15	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	2	50 all	4.000 ^a 3.999 ^b 4.062 ^c	8.000 ^a 7.959 ^b 7.926 ^c	.90 .95 .99	.895 .947 .987	7.619 9.260 12.813	.927
16	(.30, .30 .40)	3	50 all	4.000 ^a 4.012 ^b 4.038 ^c	8.000 ^a 8.005 ^b 7.954 ^c	.90 .95 .99	.897 .950 .988	7.678 9.460 13.095	.956
17	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	3	80 all	6.000 ^a 6.116 ^b 6.155 ^c	12.000 ^a 12.190 ^b 12.113 ^c	.90 .95 .99	.903 .949 .984	10.868 12.544 15.949	.940
18	(.16, .16, .16, .16, .16, .20)	2	100 all	5.000 ^a 5.035 ^b 5.060 ^c	10.000 ^a 9.967 ^b 9.854 ^c	.90 .95 .99	.900 .954 .988	9.266 11.391 14.688	.905
19	(.20, .20 .20, .40)	4	80 all	9.000 ^a 8.908 ^b 8.971 ^c	18.000 ^a 17.054 ^b 16.336 ^c	.90 .95 .99	.895 .945 .986	14.283 16.621 21.064	.941
20	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	4	30 all	6.000 ^a 6.280 ^b 6.347 ^c	12.000 ^a 12.860 ^b 13.018 ^c	.90 .95 .99	.926 .961 .986	11.770 13.357 16.128	.968
21	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	3	100 all	8.000 ^a 7.967 ^b 7.992 ^c	16.000 ^a 15.887 ^b 16.028 ^c	.90 .95 .99	.890 .940 .987	12.986 15.047 19.580	.926
22	(.30, .30 .40)	5	50 all	8.000 ^a 7.794 ^b 7.815 ^c	16.000 ^a 15.389 ^b 15.438 ^c	.90 .95 .99	.890 .940 .981	12.993 15.027 18.554	.970
23	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	4	100 all	12.000 ^a 12.200 ^b 12.198 ^c	24.000 ^a 25.296 ^b 24.856 ^c	.90 .95 .99	.908 .956 .993	18.873 21.569 27.498	.932
24	(.10, .10, .10 .10, .10, .10 .10, .30)	6	100 all	35.000 ^a 35.035 ^b 34.940 ^c	70.000 ^a 80.488 ^b 69.204 ^c	.90 .95 .99	.918 .958 .991	47.617 51.247 59.326	.912

^a Approximated.

^b Observed C.

^c Observed χ^2 .

Constatações:

1) Os valores das estatísticas X e C independem das estruturas de probabilidade, da distribuição multinomial dependem apenas do número grau de liberdade, $V = (K-1)*(I-1)$.

2) Mesmo para valores pequenos de N , a estatística C é bastante próxima de sua distribuição assintótica sob H_0 . Semelhante comportamento se observa em X .

3) As médias e variâncias das estatísticas X e C se aproximam, respectivamente, da média e variância, de suas distribuições assintóticas sob H_0 .

4) O quantil do valor empírico da estatística C é bastante próximo do nível de confiança fixo, de sua distribuição assintótica sob H_0 .

5) É alta a correlação de postos entre as duas estatísticas (todos acima de 0,90).

Das constatações, pode-se concluir que as distribuições de C e X se aproximam de uma distribuição qui-quadrado, mesmo para valores pequenos de N . Os valores das duas estatísticas C e X são bastante próximos, podendo serem usadas indiferentemente, no caso de uma partição fixa. E a alta correlação de pontos indica que as duas estatísticas têm a mesma tendência.

Além desta simulação, Light e Margolin (1971) realizaram uma outra para observar o comportamento das estatísticas C e X sob diferentes hipóteses alternativas específicas, com os parâmetros

indicados na Tabela 5.2.3.

TABELA 5.2.3 - Parâmetros utilizados na Simulação de Light e Margolin (1971).

		π		
K = 2	sob H0	1/3	1/3	1/3
I = 3	sob H1	.40	.30	.30
$N_k = (100, 100)$		.44	.28	.28
R = 10.000		.52	.24	.24
		.60	.20	.20
		.72	.14	.14
		.36	.36	.28
		.39	.39	.22
		.42	.42	.16
		.45	.45	.10
		.48	.48	.04

De cada réplica, calcularam-se os poderes empíricos das estatística X e C a níveis de 1%, 5% e 10%. Os resultados obtidos estão apresentados na Tabela 5.2.4.

TABELA 5.2.4 - Estudo empírico do poder comparativo entre as estatística C e X, obtido através de Simulação de Light e Margolin (1971).

	π sob H_0	π sob H_1	Nível	Poder empírico		Correl. de posto entre X e C
				X	C	
1	.333	.400	.01	.040	.040	.942
	.333	.300	.05	.120	.117	
	.333	.300	.10	.207	.207	
2	.333	.440	.01	.093	.120	.928
	.333	.280	.05	.278	.287	
	.333	.280	.10	.398	.406	
3	.333	.520	.01	.435	.493	.932
	.333	.240	.05	.673	.701	
	.333	.240	.10	.780	.800	
4	.333	.600	.01	.857	.892	.901
	.333	.200	.05	.947	.961	
	.333	.200	.10	.972	.974	
5	.333	.720	.01	.992	.996	.898
	.333	.140	.05	.997	.999	
	.333	.140	.10	.998	.999	
6	.333	.360	.01	.032	.032	.958
	.333	.360	.05	.103	.099	
	.333	.280	.10	.185	.176	
7	.333	.390	.01	.167	.138	.932
	.333	.390	.05	.368	.333	
	.333	.220	.10	.494	.464	
8	.333	.420	.01	.511	.406	.907
	.333	.420	.05	.756	.701	
	.333	.160	.10	.858	.809	
9	.333	.450	.01	.926	.827	.911
	.333	.450	.05	.976	.955	
	.333	.100	.10	.989	.980	
10	.333	.480	.01	.997	.978	.892
	.333	.480	.05	.999	.998	
	.333	.040	.10	.999	.999	

Constatações:

1) Para estruturas com alta probabilidade num único nível de resposta (unimodal) (casos 1 a 5), observou-se que o poder da estatística C supera o de X.

2) Para estruturas alternativas com alta probabilidade em 2 dos níveis (distribuições bimodais) de resposta (casos 6 a 10), observou-se que o poder da estatística X supera o de C. Além disto, os poderes empíricos de X e C alcançam valores "razoáveis" nos casos 4, 5, 9 e 10. Os casos 4 e 5 são os que apresentam estruturas de probabilidade "extremas" de distribuições unimodais e os casos 9 e 10 são os que apresentam estruturas de probabilidade "extremas" de distribuições bimodais. Pode-se concluir que as estatísticas X e C são boas para casos extremos de unimodalidade e bimodalidade.

Com o objetivo de obter as correlações de postos entre os vários pares de estatística C, X, V, θ e δ , Morgan e Messenger (1973) realizaram simulação com os seguintes parâmetros:

$$I = 10 \quad K = 2$$

$$\underline{\pi} = (1/10, 1/10, \dots, 1/10)$$

$$N_1 + N_2 = 10.000$$

$$R = 10$$

Em cada réplica, os 10.000 ensaios multinomiais de parâmetros $\underline{\pi}$ são segmentados em 2 grupos, sendo que cada ensaio é designado em um dos 2 grupos de modo equiprovável. Efetuada a partição, calculam-se as estatísticas de interesse e a correlação de postos

dos diversos pares de estatística, obtendo os resultados apresentados a seguir:

TABELA 5.2.5 - Matriz de correlação de postos obtidos através de simulação de Morgan e Messenger (1973).

	δ	V	C	X
V	.82			
C	.87	.96		
X	.82	1.00	.96	
θ	.24	.24	.29	.24

Constatações

1) Uma forte correlação entre X, C e V. δ também se correlaciona com o trio mencionado, porém de uma maneira mais fraca.

2) θ não se correlaciona com nenhuma das 4 estatísticas anteriores.

Através das correlações de postos, pode-se concluir que os valores de C, X e V devem ser bastante próximos, sendo indiferente usar qualquer uma das 3 estatísticas. O valor da estatística δ deve seguir a mesma tendência de C, X e V. A estatística θ deve ter um comportamento diferenciado, visto não se correlacionar fortemente com nenhuma das estatísticas anteriores.

5.3 Simulações para partições ótimas

5.3.1 Introdução

Neste item são feitas algumas simulações para estudar o comportamento das estatísticas C , X e V , obedecendo às regras Livre e Monotônica. Os parâmetros empregados nesta etapa estão indicados na Tabela 5.3.1.1.

TABELA 5.3.1.1 - Parâmetros de simulação de C , X e V .

K	I	π	Entropia
3	3	.94 .03 .03	0.83
		.69 .23 .08	0.30
		.48 .48 .03	0.27
		1/3 1/3 1/3	0.00
	4	1/4 1/4 1/4 1/4	0.00
	5	1/5 1/5 1/5 1/5 1/5	0.00
		2/5 3/20 3/20 3/20 3/20	0.10
		3/5 1/10 1/10 1/10 1/10	0.38
		2/5 2/5 1/15 1/15 1/15	0.33
		9/20 9/20 2/60 2/60 2/60	0.55
	8	1/8 1/8 ... 1/8	0.00
	10	1/10 1/10 ... 1/10	0.00
5	3	1/3 1/3 1/3	0.00
	4	1/4 1/4 1/4 1/4	0.00
	5	1/5 1/5 1/5 1/5 1/5	0.00
	8	1/8 1/8 1/8 ... 1/8	0.00
	10	1/10 1/10 1/10...1/10	0.00

Estas simulações visam obter informações sobre:

1) Comportamento das estatísticas para vários parâmetros Π , mantendo fixos K e I. As informações foram obtidas para (K = 3 e I = 3) e (K = 3 e I = 5). Os resultados estão relatados no item 5.3.2.

2) Comportamento das estatísticas para vários valores de K e I. As simulações foram feitas com estruturas de equiprobabilidade para valores (I = 3, 4, 5, 8, 10 e K = 3) e (I = 3, 4, 5, 8, 10 e K = 5). Os resultados observados estão comentados no item 5.3.3.

3) Distribuição empírica dos valores gerados nos 2 itens anteriores. Utilizando-se a aproximação de Wilson-Hilferty (Kendall e Stuart, vol.1), tenta-se identificar se a distribuição empírica de C , X e V tem uma distribuição qui-quadrado. Os resultados estão descritos no item 5.3.4.

4) A coincidência das partições ótimas obtidas pelas 3 estatísticas. No item 5.3.5 estão os resultados observados.

5.3.2 Comportamento das estatísticas para valores fixos de K e I

A realização desta simulação deve-se ao interesse em se constatar a uniformidade das distribuições empíricas das estatísticas, obtidas com diferentes parâmetros, fixando K e I. Caso fosse observada a semelhança, "bastaria" simular apenas para a estrutura de equiprobabilidade pois ela "representaria" o rol dos parâmetros Π . Para que não fossem estudadas populações com mesma estrutura de probabilidade, foi utilizado como critério de escolha

dos parâmetros, o coeficiente de entropia ($p_i \ln p_i$). Valores altos da entropia indicam que o parâmetro se afasta da estrutura de equiprobabilidade. Foram escolhidos parâmetros com valores de coeficiente de entropia diferenciados, conforme mostra a Tabela 5.3.1.1.

Os resultados observados através destas simulações estão dispostos nas Tabelas A-1, A-2, que se encontram no Apêndice.

Constatações:

K = 3 e I = 3	K = 3 e I = 5
1) As médias das 3 estatísticas são bastante próximas, independentemente do parâmetro $\mathbb{I}$	1) Idem
2) As variâncias de X_i e V_i são bastantes semelhantes, independentes do parâmetro $\mathbb{I}$. As variâncias de C_i são semelhantes em apenas 2 casos: (.94, .03, .03) e (1/3, 1/3, 1/3). Nos demais casos, observam-se valores muito maiores do que os 2 casos mencionados.	2) Idem Não se observou semelhança nas variâncias de C_i nos casos simulados.
3) Os quantis a diversos níveis de X_i e V_i são muito semelhantes, independentemente do parâmetro $\mathbb{I}$. Para os obtidos através de C_i , a semelhança ocorreu em 2 casos: (.94, .03, .03) e (1/3, 1/3, 1/3); nos demais casos, os valores são muito maiores que os 2 casos mencionados.	3) Idem Não se observou semelhança dos quantis de C_i nos casos estudados.

Estas constatações ocorreram com as 2 regras de partição em estudo: (Livre e Monotônica).

5.3.3 - Comportamento das estatísticas para valores variados de l e K .

A finalidade desta simulação é observar o comportamento das estatísticas geradas segundo estruturas de equiprobabilidade com as seguintes características:

- $K = 3$ e variando valores de l ($l = 3, 4, 5, 8, 10$)
- $K = 5$ e variando valores de l ($l = 3, 4, 5, 8, 10$)

Os resultados observados desta fase estão apresentados nas Tabelas A-3, A-4, que se encontram no Apêndice.

Constatações:

$I = 3, 4, 5, 8, 10$

$K = 3$

$K = 5$

1) As médias e variâncias das 3 estatísticas C , X e V são muito semelhantes.

1) Idem

2) A diferença entre as médias observadas e média assintótica tende a crescer, conforme cresce o valor de I ,

2) Idem

Regra	$\neq$ quando $I=3$	$\neq$ quando $I=10$
Livre	1,5	3,0
Monotônica	1,0	1,5

Regra	$\neq$ quando $I=3$	$\neq$ quando $I=10$
Livre	2,5	7,0
Monotônica	1,0	1,8

3) A diferença entre a variância observada e variância assintótica decresce conforme cresce I .

3) A diferença entre a variância observada e variância assintótica se comportou conforme abaixo, segundo as regras de partição.

Regra	$\neq I = 3$	$\neq I = 10$
Livre	+2,0	-1,0
Monotônica	+2,0	-1,0

Regra	$\neq I = 3$	$\neq I = 10$
Livre	+4,5	-1,0
Monotônica	+2,0	+5,0

4) As diferenças entre a média obtida com regra Livre e regra Monotônica são $\pm$ "constante" para qualquer valor de I .

4) As diferenças entre as regras livres e monotônicas crescem conforme aumenta o valor de I .

5) As variâncias obtidas segundo as 2 regras apresentaram comportamentos semelhantes (vide gráfico)

5) As variâncias obtidas segundo regra Livre apresentaram um comportamento bastante diferenciado do obtido segundo regra Monotônica.

Notação: μ_M Média segundo regra Monotônica
 μ_L Média segundo regra Livre
 Var_M Variância segundo regra Monotônica
 Var_L Variância segundo regra Livre

Os gráficos 5.3.3.1 e 5.3.3.2 ilustram estas constatações, salientando que a média e a variância das 3 estatísticas foram substituídas pela "média" geral devido à primeira constatação, apenas para efeito de confecção dos gráficos.

Além destas, outras constatações, referentes ao efeito do aumento de K , foram anotadas:

- 1) $\mu_{M,K=3} \approx \mu_{M,K=5}$, para qualquer I
- 2) $\mu_{L,K=5} > \mu_{M,K=3}$, para qualquer I
- 3) $\text{Var}_{M,K=5} > \text{Var}_{L,K=3}$, para qualquer I

Na Tabela 5.3.3.1 estão as correlações lineares entre os vários pares das estatísticas C , X e V , segundo várias condições, podendo-se observar, também, altas correlações entre os vários pares de estatísticas (Todos superiores a .94).

Gráfico 5.3.3.1 - Médias obtidas segundo Regras de Partição.

Gráfico 5.3.3.2 - Variâncias obtidas segundo Regras de Partição.

TABELA 5.3.3.1 - Correlações lineares de X_0 , C_0 , V_0 .

K	I	Regra Livre			Regra Monotônica		
		Correlações entre			Correlações entre		
		C_0 e X_0	C_0 e V_0	V_0 e X_0	C_0 e X_0	C_0 e V_0	V_0 e X_0
3	3	.9990	.9986	.9997	.9990	.9987	.9997
	4	.9971	.9969	.9995	.9977	.9973	.9996
	5	.9957	.9947	.9992	.9993	.9956	.9963
	8	.9911	.9894	.9982	.9920	.9902	.9983
	10	.9875	.9849	.9978	.9897	.9871	.9981
5	3	.9989	.9984	.9994	.9933	.9928	.9985
	4	.9978	.9968	.9987	.9900	.9728	.9816
	5	.9966	.9942	.9981	.9688	.9898	.9751
	8	.9914	.9854	.9952	.9751	.9621	.9678
	10	.9880	.9816	.9951	.9703	.9413	.9628

Nas constatações pode-se concluir que, mesmo aumentando o número de populações e o número de respostas, as médias e variâncias das 3 estatísticas C_0 , X_0 e V_0 são muito próximas em situação de equiprobabilidade.

É alta a correlação linear entre pares das estatísticas C_0 , X_0 e V_0 , indicando que o valor das 3 estatísticas devam ser próximos.

5.3.4 Distribuição empírica das estatísticas

Neste item serão verificadas se as distribuições empíricas das estatísticas C_0 , X_0 e V_0 se aproximam a uma distribuição do tipo aX_0^2 . Para isto, será utilizada a aproximação de Wilson-Hilferty (Kendall e Stuart, Vol.1), constando que:

$$(X_0^2 / u)^{1/3} \sim N \left(1 - \frac{2}{9u}; \frac{2}{9u} \right)$$

Considere uma variável genérica Y , cuja distribuição não é conhecida. Se o gráfico de $\sqrt[3]{y} \times \Phi^{-1}(F(y))$ apresentar uma reta, então:

$$\sqrt[3]{Y} \sim N(\mu, \sigma^2) \text{ ou } Y \sim aX_0^2 \text{ com}$$

$$\mu = (au)^{1/3} \left(1 - \frac{2}{9u}\right)$$

$$\sigma = (au) \left(\frac{2}{9u}\right)$$

(Bussab, 1976).

Na Figura 5.3.4.1 está o gráfico de $\sqrt[3]{X_0}$ contra $\Phi^{-1}[F(X_0)]$, observando uma tendência linear. Os gráficos das outras estatísticas apresentaram comportamento semelhante. Como os dados apresentaram uma tendência linear, pode-se ajustar uma regressão linear segundo Métodos de Mínimos Quadrados de $\Phi^{-1}[F(Y)]$ em $\sqrt[3]{Y}$. Isto foi aplicado às estatísticas C_0 , X_0 e V_0 , obtendo as retas constantes na Tabela 5.3.4.1.

NORMAL INVERSA DA DISTR. EMPIRICA

FIGURA 5.3.4.1 - Gráfico de $\sqrt[3]{x_0}$ contra $F(x_0)$.

TABELA 5.3.4.1 - Coeficientes angulares e lineares das retas ajustadas.

π	Critério de seleção	K = 3						K = 5					
		Regra Livre			Regra Monotônica			Regra Livre			Regra Monotônica		
		Coef.		R^2	Coef.		R^2	Coef.		R^2	Coef.		R^2
		Linear	Ang.		Linear	Ang.		Linear	Ang.		Linear	Ang.	
1/3...1/3	X	-3.73	2.67	0.997	-3.53	2.66	0.994	-5.53	3.22	0.997	-3.06	2.36	0.998
	$\emptyset$												
	C	-3.72	2.66	0.997	-3.52	2.65	0.994	-5.56	3.24	0.997	-3.07	2.37	0.998
	$\emptyset$												
1/4...1/4	V	-3.73	2.66	0.997	-3.52	2.65	0.994	-5.52	3.22	0.997	-3.04	2.35	0.998
	$\emptyset$												
	X	-5.05	3.15	0.994	-4.65	3.03	0.992	-6.03	3.62	0.997	-3.71	2.51	0.997
	$\emptyset$												
1/5...1/5	C	-5.06	3.15	0.994	-4.67	3.04	0.992	-6.05	3.63	0.996	-3.69	2.50	0.998
	$\emptyset$												
	V	-5.02	3.13	0.993	-4.64	3.02	0.991	-6.01	3.60	0.997	-3.70	2.50	0.997
	$\emptyset$												
1/6...1/6	X	-5.70	3.33	0.998	-5.29	3.17	0.998	-7.64	3.73	0.992	-4.07	2.47	0.997
	$\emptyset$												
	C	-5.76	3.32	0.998	-5.29	3.18	0.998	-7.63	3.72	0.991	-4.02	2.44	0.997
	$\emptyset$												
1/8...1/8	V	-5.77	3.32	0.998	-5.26	3.15	0.998	-7.57	3.69	0.990	-4.05	2.45	0.997
	$\emptyset$												
	X	-7.36	3.51	0.997	-6.96	3.41	0.995	-10.35	4.42	0.997	-5.54	2.82	0.998
	$\emptyset$												
1/10...1/10	C	-7.32	3.49	0.997	-6.96	3.42	0.995	-10.33	4.41	0.997	-5.56	2.83	0.998
	$\emptyset$												
	V	-7.28	3.47	0.997	-6.96	3.41	0.995	-10.30	4.38	0.997	-5.52	2.80	0.998
	$\emptyset$												
1/10...1/10	X	-8.43	3.76	0.998	-7.57	3.50	0.997	-11.39	4.56	0.998	-6.36	2.94	0.995
	$\emptyset$												
	C	-8.31	3.71	0.998	-7.47	3.45	0.997	-11.26	4.51	0.998	-6.28	2.92	0.996
	$\emptyset$												
1/10...1/10	V	-8.39	3.73	0.998	-7.51	3.46	0.996	-11.26	4.50	0.998	-6.26	2.90	0.996
	$\emptyset$												

Da Tabela 5.3.4.1 pode-se constatar que:

1) As retas ajustadas às 3 estatísticas são muito semelhantes dentro de uma estrutura de equiprobabilidade (fixado K e I).

2) As retas apresentam alto valor do coeficiente de explicação (todos superiores a 98%).

3) Os valores dos Resíduos estão variando entre -0.3 a 0.15 que, padronizados, correspondem a ± 2.0 . Não apresentam aleatoriedade, visto que a construção de $\hat{F}^{-1}(y)$ não foi de forma aleatória.

Apesar da não aleatoriedade dos resíduos, pode-se considerar que a distribuição C , V e X se aproxima de uma distribuição do tipo aX_{ν}^2 , com a e ν estimados pelos valores apresentados na Tabela 5.3.4.2.

TABELA 5.3.4.2 - Valores estimados de a e $\checkmark$ das retas de regressão.

π	Estatíst	K = 3				K = 5			
		Regra Livre		Regra Monotônica		Regra Livre		Regra Monotônica	
		a	$\checkmark$	a	$\checkmark$	a	$\checkmark$	a	$\checkmark$
1/3...1/3	X	0.835	3.916	0.866	3.395	0.773	7.140	1.086	2.634
	$\emptyset$								
	C	0.982	3.371	0.874	3.364	0.760	7.260	1.079	2.641
	$\emptyset$								
1/4...1/4	V	0.915	3.749	0.883	3.341	0.777	7.130	1.100	2.580
	$\emptyset$								
	X	0.800	5.775	0.845	4.864	0.668	10.719	1.126	3.499
	$\emptyset$								
1/4...1/4	C	0.800	5.775	0.840	4.881	0.665	10.722	1.118	3.497
	$\emptyset$								
	V	0.817	5.679	0.856	4.813	0.678	10.605	1.130	3.478
	$\emptyset$								
1/5...1/5	X	0.703	8.094	0.770	6.623	0.703	12.873	1.189	4.390
	$\emptyset$								
	C	0.707	8.020	0.767	6.623	0.704	12.855	1.239	4.237
	$\emptyset$								
1/5...1/5	V	0.711	8.045	0.781	6.569	0.726	12.548	1.233	4.298
	$\emptyset$								
1/8...1/8	X	0.775	12.477	0.809	11.081	0.601	22.164	1.139	7.341
	$\emptyset$								
	C	0.792	12.251	0.820	10.930	0.619	21.500	1.150	7.236
	$\emptyset$								
1/8...1/8	V	0.798	12.263	0.833	10.853	0.628	21.402	1.194	7.067
	$\emptyset$								
1/10...1/10	X	0.712	16.432	0.779	13.684	0.522	30.479	1.053	10.266
	$\emptyset$								
	C	0.744	15.766	0.814	13.120	0.540	29.537	1.082	9.926
	$\emptyset$								
1/10...1/10	V	0.726	16.295	0.798	13.509	0.542	29.649	1.090	9.934
	$\emptyset$								

5.3.5 Valores críticos obtidos de vários critérios.

Nas Tabelas 5.3.5.1 e 5.3.5.2 estão os quantis a 95%, obtidos de várias maneiras:

- Desigualdade de Bonferroni
- Comparações Múltiplas

- Regressão
- Observado Empírico

TABELA 5.3.5.1 - Valores críticos das estatísticas C , X e V a 95% obtidos de maneiras diferentes ($K = 3$).

π	Estatística	Regra Livre				Regra Monotônica			
		Empírico	Bonferroni	Regressão	C.Mult.	Empírico	Bonferroni	Regressão	C.Mult.
1/3...1/3	X	7.71		8.14		7.70		7.34	
	θ								
	C	8.68	8.175	8.13	9.48	7.72	7.38	7.38	9.48
	θ								
1/4...1/4	V	8.40		8.23		7.72		7.38	
	θ								
	X	9.38		9.58		8.84		8.95	
	θ								
1/5...1/5	C	9.42	10.24	9.62	12.59	9.06	9.35	8.94	12.59
	θ								
	V	9.52		9.63		9.12		8.99	
	θ								
1/5...1/5	X	11.21		11.06		10.77		10.45	
	θ								
	C	11.18	12.11	11.07	15.51	10.55	11.14	10.35	15.51
	θ								
1/8...1/8	V	11.49		11.12		11.06		10.51	
	θ								
	X	16.62		16.86		15.73		16.04	
	θ								
1/8...1/8	C	16.40	17.07	16.92	23.68	15.91	16.01	15.90	23.68
	θ								
	V	16.62		16.99		15.73		16.04	
	θ								
1/10...1/10	X	19.01		19.21		18.31		18.22	
	θ								
	C	19.19	20.23	19.30	28.87	18.28	19.02	18.41	28.87
	θ								
1/10...1/10	V	19.27		19.44		18.73		18.49	
	θ								

TABELA 5.3.5.2 - Valores críticos das estatísticas C_0 , X_0 e V_0 obtidos através de vários critérios para $K = 5$ a 95%.

Estrutura Estatística		Regra Livre				Regra Monotônica			
		Empírico	Bonferroni	Regres.	C.Mult.	Empírico	Bonferroni	Regres.	C.Mult.
1/3...1/3	X	11.24		11.04		7.34		7.90	
	θ								
	C	11.13	11.51	10.97	15.51	7.20	8.75	7.85	15.51
	θ								
	V	11.32		10.99		7.33		7.90	
1/4...1/4	X	12.78		12.01		10.22		9.68	
	θ								
	C	12.43	13.81	12.79	21.03	10.05	10.86	9.69	21.03
	θ								
	V	12.74		12.93		10.19		9.75	
1/5...1/5	X	16.05		15.40		11.82		12.35	
	θ								
	C	16.01	15.87	15.47	26.03	11.77	12.77	12.48	26.30
	θ								
	V	16.21		15.55		11.70		12.53	
1/8...1/8	X	20.70		19.96		16.96		16.50	
	θ								
	C	20.76	22.76	20.00	41.34	16.56	16.60	16.47	41.34
	θ								
	V	20.90		20.26		17.28		16.72	
1/10...1/10	X	22.79		23.33		19.82		20.15	
	θ								
	C	22.93	24.76	23.40	43.77	19.42	21.04	19.95	43.77
	θ								
	V	22.93		23.56		19.10		20.22	

Pode-se observar que os valores críticos obtidos através da Desigualdade de Bonferroni se aproximam dos valores simulados, ao passo que os obtidos através da Comparação Múltipla se afastam muito dos valores simulados. Isto implica que a desigualdade Bonferroni é um critério de avaliação recomendado.

5.3.6 - Coincidência da partição ótima escolhida através de critérios diferentes

Para verificar se os três critérios X , C e V selecionam a mesma partição, foi construída a Tabela 5.3.6.1 onde, para cada estrutura de probabilidade, apresenta-se a porcentagem de partições ótimas coincidentes entre (X e C), (C e V) e (X e V).

TABELA 5.3.6.1 - Porcentagem de coincidência na seleção da partição ótima através de critérios diferentes.

π	K = 3						K = 5					
	Regra Livre			Regra Monotônica			Regra Livre			Regra Monotônica		
	X, C	C, V	X, V	X, C	C, V	X, V	X, C	C, V	X, V	X, C	C, V	X, V
	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
1/3...1/3	96.0	96.5	99.5	99.5	99.5	99.5	98.6	97.8	98.0	98.6	97.8	98.0
1/4...1/4	98.1	97.8	99.7	98.9	98.6	99.5	96.0	92.6	95.0	98.2	97.0	98.6
1/5...1/5	97.7	97.3	99.1	99.1	99.0	99.7	95.4	94.0	97.0	98.2	97.6	99.0
1/8...1/8	95.6	95.8	97.4	98.6	98.0	98.6	93.0	91.8	96.2	96.8	96.2	97.4
1/10...1/10	95.3	94.6	97.9	96.0	95.8	98.3	90.0	88.4	93.0	97.4	94.4	96.0

Da Tabela 5.3.6.1 pode-se constatar que:

1) É alta a porcentagem de coincidência na seleção da partição ótima por diferentes critérios (todos acima 88%), significando que os 3 critérios diferentes podem conduzir ao mesmo resultado.

3) Aparentemente, conforme cresce o valor de l , a porcentagem

de coincidência tende a decrescer (bem pouco).

4) A porcentagem de coincidência é maior na regra monotônica. Quanto menor o valor de I e K, maior a porcentagem de coincidência.

5.4 Propriedades amostrais de θE_0 .

Neste item será descrita a simulação realizada por Morgan e Messenger (1973) para estudar o comportamento da estatística θE_0 . Ressalta-se que a diferença desta simulação com as anteriores se deve ao fato de ter sido feita simulação de diversas variáveis preditoras e aplicações iterativas do processo de procura de melhor partição.

Os parâmetros utilizados neste estudo são:

$$\begin{aligned} i &= 1, 2 \\ \prod_k (X_j) &= [1/2, 1/2] \quad k = 1, 2, \dots, 5, \text{ qualquer } X_j \\ j &= 1, 2, \dots, 8 \\ \sum_{k=1}^5 \sum_{i=1}^2 n_{ki} (X_j) &= 500 \\ R &= 10 \end{aligned}$$

A simulação consistia nas seguintes operações:

1) Seleciona-se a melhor partição dicotômica segundo o critério θE_0 para cada variável preditora X_j , sendo associada a cada uma a estatística $\theta E_0(X_j)$, $j = 1, \dots, 8$.

2) Seleciona-se a preditora X_0 para partição, aquela que satisfizer

$$\theta E_0(X_j) = \max (\theta E_0(X_j), \quad j = 1, \dots, 8), \quad \text{obtendo-se os}$$

grupos A e B.

3) Os passos 1) e 2) são repetidos para os grupos A e B, obtendo-se $\theta E_{O/A} (X_0)$ e $\theta E_{O/B} (X_0)$. É selecionado como grupo para partição aquele que apresentar maior $\theta E(X_0)$, obtendo-se 3 grupos.

4) O procedimento é aplicado novamente nestes 3 grupos, obtendo-se 4 grupos e assim, sucessivamente, até 14 grupos. Em cada estágio anota-se o valor de $\theta E_0(p)$, $p = 0, \dots, 13$ e o seu incremento

$$INC(p) = \theta E_0(p+1) - \theta E_0(p)$$

dos incrementos de acordo com o tamanho da amostra.

FIGURA 5.4.1 - Estatísticas descritivas de $INC(p)$.

5) Com 10 réplicas, calcularam-se algumas estatísticas descritivas que estão representadas na Figura 5.4.1.

Constatações:

Os incrementos começam a se estabilizar para número de grupos superior a 8, com valor médio girando em torno de 0.08. Para um número de grupos inferior a 8, mostra-se uma relação crescente com número de grupos.

Outras simulações realizadas por Morgan e Messenger (1973) mostraram que as curvas da Figura 5.4.1 são aproximadamente proporcionais à raiz quadrada do recíproco do tamanho da amostra. Assim, podem-se reescalonar as curvas da Figura por $\sqrt{500/N}$ para se ter valores apropriados das estatísticas de acordo com o tamanho da amostra.

Além disso, as curvas servem como indicador para avaliar como "significativa" a partição selecionada pelo critério θE_0 , caso o valor da estatística associada fosse superior a curva dos valores máximos. Caso os valores das estatísticas estiverem entre as curvas de média e do máximo, não se deve tomar nenhuma decisão. Como a simulação foi feita sob H_0 , os valores dos incrementos não devem ser grandes. Pode-se concluir, também, que aumentando a amostra deve diminuir o incremento.

6 APLICAÇÕES PRÁTICAS

Neste capítulo serão apresentadas aplicações da técnica descrita em 3 exemplos. O primeiro exemplo, ilustra o uso da técnica descrita iterativamente numa situação com apenas uma variável preditora. Os resultados obtidos e procedimentos empregados estão descritos no item 6.1. O item 6.2 ilustra através de 2 exemplos o uso da técnica numa situação com diversas variáveis preditoras. O primeiro exemplo é uma outra maneira de analisar os dados do item 6.1, considerando o caso de 2 variáveis preditoras. O segundo exemplo é uma aplicação com mais de 2 variáveis preditoras.

6.1 Exemplo com uma variável preditora

Suponha uma situação onde são controladas P variáveis preditoras, cada uma com $k(p)$ classes, $p = 1, \dots, P$ e o pesquisador tem interesse em verificar quais das $\prod_{p=1}^P k(p)$ combinações possíveis possuem comportamentos semelhantes em relação à variável resposta. Uma das maneiras de resolver este problema é aplicar a técnica AID para dados categorizados iterativamente, utilizando-se as $\prod_{p=1}^P k(p)$ combinações não vazias como se fossem classes de uma única variável preditora.

Neste caso, para aplicar a técnica foram utilizados os seguintes procedimentos

- 1) a estatística X_0 para seleccionar a partição "ótima";

- 2) desigualdade de Bonferroni para avaliar a significância da partição "ótima" individualmente;

3) a partir da 2^a iteração, usualmente haverá 2 ou mais caselas disponíveis para partição, tendo, cada uma, sua partição ótima correspondente. Escolhe-se a partição que apresentar o menor nível descritivo, porém não é conhecida a distribuição da estatística X_0 (embora tenha simulada para alguns valores de I e K , conforme capítulo 5). Em vista disso, necessita-se adotar um critério para selecionar a melhor partição. Os critérios mais usuais são:

3.1) Através do critério empírico: maior coeficiente de explicação R^2 , pois acredita-se que, quanto maior o R^2 , mais significativa é a partição. Este é o critério mais utilizado na prática devido à sua facilidade e também devido à indisponibilidade da distribuição de X_0 . Como regra prática, a partição que irá produzir o maior coeficiente de explicação (R^2) corresponde à que está associada ao maior valor de X_0 . Apesar do coeficiente de explicação ser calculado através da estatística C_0 , a regra prática pode ser aplicada visto que C_0 e X_0 apresentam alta correlação linear (superior 0.95) e também alta porcentagem de coincidências na seleção da partição ótima.

3.2) Através da "significância de Bonferroni" (Kass, 1980); escolhe-se a partição que apresentar o menor nível de "significância de Bonferroni". Kass (1980) o definiu como sendo o resultado do produto do nível descritivo de X_0 obtido segundo a distribuição de X_0 sob H_0 com o multiplicador de Bonferroni (Tabela 4.3.1). Este é o critério adotado no programa CHAID de Kass (1980) na seleção da "melhor" casela.

Neste exemplo, será adotado o critério empírico na seleção da melhor.

4) desigualdade de Bonferroni para avaliar a significância global da partição selecionada em 3). Segundo recomendação de Kass (1980), a avaliação global da casela selecionada em 3 é feita novamente pela desigualdade de Bonferroni, levando em consideração o número de partições acumuladas até esta etapa na determinação do valor crítico. Caso seja avaliada como significativa, a partição da casela selecionada acontece.

Este processo é repetido até não satisfazer o item 4.

Os dados utilizados nesta aplicação fazem parte do relatório RAE-SEA-8505, que estão reproduzidos na Tabela 6.1.1. São dados referentes a uma pesquisa sobre aderência de células onde elas são classificadas em aderentes ou não, controlados o estado de ativação (A) e o tipo de revestimento do eritrócito (B). Foram pesquisados 3 estados de ativação: A1 - Residentes Normais

A2 - Estimulados

A3 - Tripsinizados,

e 4 tipos de revestimento:

B1 - IgM

B2 - α M

B3 - Salina

B4 - Soro anti-E

que combinados resultam em 12 classes.

TABELA 6.1.1 - Dados sobre aderência de células.

Classes	Ocorreu ader.	Não ocorreu ader.	Total
A1B1	29	771	800 ✓
A1B2	786	214	1000
A1B3	0	800	800 ✓
A1B4	543	257	800
A2B1	33	767	800 ✓
A2B2	437	363	800
A2B3	0	800	800 ✓
A2B4	416	384	800
A3B1	507	293	800
A3B2	666	134	800
A3B3	269	531	800 ✓
A3B4	535	265	800
Total	4221	5579	9800

Neste caso, a variável preditora é a combinação do estado de ativação e tipo de revestimento, resultando em 12 classes e a variável resposta é dicotômica (a célula adere ou não).

Os resultados obtidos através do emprego da técnica interativamente estão apresentados nas Tabelas 6.1.2 e 6.1.3 e na Figura 6.1.1.

FIGURA 6.1.1 - Resultados obtidos através do emprego da técnica no item 6.1.

269
800

TABELA 6.1.2 - Composição dos grupos obtidos segundo a técnica AID.

Grupos	Classe	Ocorreu ader.	nao ocorreu ader.	Total
1	A1B3	0	800	800
	A2B3	0	800	800
	SOMA	0	1600	1600
2	A1B1	29	771	800
	A2B1	33	767	800
	SOMA	62	1538	1600
3	A3B3	269	531	800
	SOMA	269	531	800
4	A2B2	437	363	800
	A2B4	416	384	800
	SOMA	853	747	1600
5	A1B4	543	257	800
	A3B1	507	293	800
	A3B4	535	265	800
	SOMA	1585	815	2400
6	A1B2	786	214	1000
	A3B2	666	134	800
	SOMA	1452	348	1800

TABELA 6.1.3 - Grupos Finais

Grupo	Composição	S	N	Casela
1	A1B3, A2B3,	0.0	1.0	10
2	A1B1, A2B1,	0.04	0.96	11
3	A3B3,	0.34	0.66	5
4	A2B2, A2B4,	0.53	0.47	8
5	A1B4, A3B1, A3B4	0.66	0.34	9
6	A1B2, A3B2,	0.81	0.19	7

Para obter estes resultados, primeiro procura-se a partição ótima entre todas as partições livres dicotômicas possíveis segundo a estatística X^2 , sendo a mesma julgada significativa através da desigualdade de Bonferroni. A partição dicotômica selecionada como ótima tem a seguinte composição: um grupo contendo as classes A1B3, A2B3, A1B1, A2B1, A3B3 (casela 2) e o outro grupo com as classes A2B2, A2B4, A1B4, A3B1, A3B4, A1B2 e A3B2 (casela 3).

Pode-se observar na casela 1 da figura 6.1.1 que a proporção de células aderentes é de 0.43. Com a partição, nota-se um comportamento diferenciado da proporção de células aderentes nas caselas resultantes; na casela 2, a proporção decresceu de 0.43 para 0.08, ao passo que, na casela 3, cresceu de 0.43 para 0.67.

Antes de prosseguir a divisão, procura-se a "melhor" partição livre-dicotômica segundo a estatística X^2 , em cada um dos grupos resultantes, obtendo as partições relacionadas na Tabela 6.1.4,

sendo avaliadas como significativas individualmente, segundo o critério da Desigualdade de Bonferroni. O próximo passo é escolher a partição mais significativa ($< \text{nível descritivo}$) entre as relacionadas, para que seja provavelmente dicotomizada na etapa seguinte. Para se determinar o nível descritivo é necessário conhecer a distribuição de X_0 , que foi simulada para alguns valores de K , I e gama, conforme capítulo 5. Neste caso, especificamente quando $I = 2$, foi possível estimar os níveis descritivos através dos resultados das simulações efetuadas por Bussab (1976); os valores correspondentes estão na Tabela 6.1.4.

TABELA 6.1.4 - Melhores partições das caselas 2 e 3.

Casela	Melhor partição, segundo X		X_0	Valor crítico segundo Bonferroni	R^2	Nível des- critivo estimado	B = n. partições dicotômicas
	A	B					
2	A3B3	A1B3, A2B3, A1B1, A2B1	846.64	8.802	0.21	$\frac{1}{3}(22.56)$	15
3	A1B2, A2B2, A2B4, A1B4, A3B2	A3B1, A3B4	369.91	11.657	0.06	$\frac{1}{3}(17.50)$	63

Apesar da disponibilidade, foi mantido o critério empírico para selecionar a melhor partição; no caso, a partição correspondente à casela 2 foi selecionada, visto atender o critério adotado.

Foi avaliada como significativa a casela selecionada através da Desigualdade de Bonferroni, levando em consideração o número

acumulado de partições julgadas até esta etapa na determinação do valor crítico, formando as caselas 4 e 5. Este processo é repetido até chegar aos resultados da Figura 6.1.1.

Na Tabela 6.1.5 estão os valores críticos utilizados na avaliação das partições selecionadas, bem como a seleção das caselas nas várias etapas. O termo casela-pai corresponde à casela selecionada como a melhor entre as disponíveis à partição e caselas filho são as resultantes da partição da casela-pai.

Comparando os resultados obtidos através da técnica AID com as constantes do relatório RAE 8505, constatou-se que os grupos finais são os mesmos.

TABELA 6.1.5 - Valores críticos utilizados e seleção das caselas nas várias etapas.

Etapa	Casela selecionada (pai)	Caselas geradas (filhos)	Caselas disponíveis p/próx.etapa	B=Multip. Bonferroni	B acumulado	X_0	Valor crítico Bonferroni	Concl
1	1	2, 3	2, 3	2047	2047	3.337.40	18.465	sig
2	2	4, 5	3, 4, 5	15	2062	846.64	18.966	sig
3	3	6, 7	4, 6, 7	63	2125	369.91	19.168	sig
4	6	8, 9	4, 7, 8, 9	15	2140	65.36	19.371	sig
5	4	10, 11	7, 8, 9, 10, 11	7	2147	63.22	19.576	sig
6	7	12, 13	8, 9, 10, 11	1	2148	6.36	19.886	não s

6.2 - Exemplos com mais de uma variável preditora.

6.2.1 - Primeiro exemplo

Este primeiro exemplo é um caso com 2 variáveis preditoras. Os procedimentos adotados neste exemplo foram:

1) Relaciona-se a melhor partição de cada variável preditora segundo a estatística X_0 .

2) Utiliza-se a desigualdade de Bonferroni para avaliar a significância das partições relacionadas no item 1).

3) Escolhe-se a partição mais significativa entre as relacionadas (α nível descritivo). Como isto não é possível, devido ao não conhecimento da distribuição de X_0 , então adotou-se o critério empírico maior coeficiente de explicação (R^2) para seleção da melhor partição.

4) Avalia-se a partição selecionada no item 3) através da desigualdade de Bonferroni, levando em consideração o número acumulado de partições até então avaliadas na determinação do valor crítico. Caso seja avaliada como significativa, a partição acontece.

Estes procedimentos foram aplicados ao conjunto de dados do item 6.1, considerando agora, no caso, 2 variáveis preditoras:

A - estado de ativação com 2 classes e

B - tipo de revestimento com 4 classes, obtendo os resultados apresentados na Figura 6.2.1 e na Tabela 6.2.1.1.

FIGURA 6.2.1 - Resultados obtidos através do emprego da técnica no item 6.2.1.

TABELA 6.2.1.1 - Grupos finais.

Grupo final	% de células aderentes	% de células não aderentes	Composição	Tamanho do grupo	N ^o da casela na Fig. 6.2.1
1	0	100	A1B3, A2B3	1600	12
2	4	96	A1B1, A2B1	1600	13
3	34	66	A3B3	800	8
4	53	47	A2B2, A2B4	1600	6
5	63	37	A3B1	800	9
6	67	33	A1B4, A3B4	1600	10
7	81	19	A1B2, A3B2	1800	11

Para chegar a estes resultados:

1) procurou-se a melhor partição livre-dicotômica entre as possíveis partições dicotômicas de cada variável preditora, obtendo-se as partições relacionadas na Tabela 6.2.1.2; as partições foram avaliadas individualmente como significativas, segundo o critério da Desigualdade de Bonferroni.

TABELA 6.2.1.2 - Melhor partição de variável A e B - 1^a etapa.

Variável preditora	"melhor" partição segundo X_0		X_0	Valor crítico X^2 ($\alpha, 1$)	R^2	B = n. partições dicot. possíveis
	A	B				
A	A1 A2	A3	679.00	5.63	0.07	3 - 1
B	B1 B3	B2, B4	2516.99	7.22	0.26	7 - 1

2) Escolheu-se a partição mais significativa entre as relacionadas. De acordo com o critério empírico, a partição correspondente à variável preditora B é selecionada como a mais significativa.

3) A partição selecionada no item anterior foi avaliada como significativa através da Desigualdade de Bonferroni, produzindo as caselas 2 e 3 (Figura 6.2.1). Seguindo orientação de Kass (1980), quando há mais de uma variável preditora é necessário considerar na determinação do valor crítico, o número total de partições possíveis. No caso, há 3 partições possíveis da variável preditora A e 7 partições possíveis correspondentes à variável preditora B, totalizando assim 10 partições possíveis.

4) Repete-se o item 1) para as caselas 2 e 3, obtendo as partições mencionadas nas Tabelas 6.2.1.3. Foram consideradas significativas estas partições (individual) segundo o critério da Desigualdade de Bonferroni. A partição da variável preditora A, na casela 2, é selecionada para provável divisão na próxima etapa, visto satisfazer o critério empírico adotado. Constata-se a significância global da partição selecionada, através da Desigualdade de Bonferroni, levando em consideração o número acumulado de partições julgadas na determinação do valor crítico. Em vista disto, a casela 2 foi dividida, produzindo as caselas 4 e 5.

TABELA 6.2.1.3 - Melhor partição das caselas 2 e 3 - 2^a etapa.

Casela	Variável preditora	melhor partição segundo X_0		X_0	Valor crítico $\chi^2 (\alpha/B, 1)$	R^2	B = n. de partições possíveis
		A	B				
2	A	A1, A2	A3	1607.7	5.63	6.33	3
	B	B1	B3	130.0	5.02	0.03	1
3	A	A2	A1, A3	222.26	5.63	0.04	3
	B	B2	B4	42.67	5.02	0.01	1

5) Repete-se o item 1 nas caselas disponíveis à partição, nos casos 3, 4 e 5 e assim, sucessivamente, até chegar à Figura 6.2.1.

Na Tabela 6.2.1.4 estão apresentados os valores críticos utilizados na avaliação global das partições e a seleção da melhor casela para partição nas várias etapas.

TABELA 6.2.1.4 - Valores críticos e seleção da casela nas várias etapas.

Etapa	Casela selecionada (pai)	Caselas geradas (filho)	Caselas disp. p/prox.estagio	Multip. de Bonferroni B acum.	B acum.	$\bar{X}_O$	valor critico	Conclusão
1	1	2, 3	2, 3	10	10	2516.99	7.96	sig
2	2	4,5	3, 4, 5	4	14	1607.7	8.60	sig
3	3	6, 7	4, 5, 6, 7	4	18	222.3	9.08	sig
4	5	8, 9	4, 6,7	1	19	141.34	9.27	sig
5	7	10, 11	4, 6, 7,10,11	2	21	73.49	9.40	sig
6	4	12, 13	6,7,10,11,12,13	2	23	63.22	9.56	sig
7	11	14, 15	6,7,10,12,13, 14, 15	1	24	6.36	9.71	não sig

Comparando os resultados obtidos através das duas estratégias diferentes aplicados ao mesmo conjunto de dados, (Tabelas 6.1.3 e 6.2.1.1), pode-se constatar que:

- 1) Este último processo produziu 7 grupos finais, ao passo que o primeiro, 6 grupos.
- 2) A composição dos grupos finais é semelhante nos 2 processos.
- 3) No último processo é mais visível, do que no primeiro que a variável preditora B explica melhor a variação da variável resposta do que a variável A.

Por estas constatações, pode-se concluir que estratégias dife-

rentes podem conduzir a resultados diferentes. A escolha da estratégia depende muito da finalidade da análise. Quando se quer estudar o comportamento das caselas, o primeiro procedimento é o mais recomendado enquanto que o efeito das variáveis preditoras é melhor avaliado através do segundo procedimento, onde são eliminadas algumas possibilidades de partição.

6.2.1-Segundo exemplo

Neste segundo exemplo, será aplicada a técnica do AID para um conjunto de dados, envolvendo diversas variáveis preditoras e uma variável resposta.

Os procedimentos adotados neste exemplo foram:

1) relaciona-se a melhor partição dicotômica de cada variável preditora, segundo a estatística X_0 .

2) Utiliza-se a desigualdade de Bonferroni para avaliar a significância das partições relacionadas no item 1).

3) Escolhe-se a partição mais significativa entre as relacionadas (menor nível descritivo). Como isto não é possível devido ao não conhecimento da distribuição de X_0 , utilizou-se um critério empírico (maior coeficiente de explicação) para seleção da melhor partição.

4) Avalia-se a partição selecionada no item 3, através dos seguintes critérios:

4.1) Desigualdade de Bonferroni, levando em consideração o número acumulado de partições até então avaliadas na determinação

$$2 < R \leq 5 \text{ S.M.}$$

$$R > 5 \text{ S.M.}$$

onde S.M. = Salário Mínimo (na época Cr\$ 600.000,00)

A descrição das variáveis preditoras utilizadas está apresentada na Tabela 6.2.2.1.

TABELA 6.2.2.1 - Variáveis preditoras utilizadas no exemplo 2.

Nome da preditora	n. de classes	Tipo de partição	Rótulo das classes
posição na família	2	livre	chefe, não chefe
cor	2	livre	branca, não branca
local de residência	2	livre	SP, NSP
tempo de residência	2	livre	menos de 3 anos, mais de 3 anos
sexo	2	livre	masculino, feminino
idade	5	monotônica	10 a 14, 15 a 17, 18 a 24, 25 a 39, mais de 40 anos
instrução	6	flutuante	s/decl, analfabeto, 1- grau inc e comp, no mínimo 2o grau incomp.
ramo de atividade	4	livre	ind de transf, serviços, comércio, outros
posição na ocupação	5	livre	empr.doméstica, empregador, assalariado, autônomo e outros.

Os resultados obtidos através do emprego iterativo da técnica, estão apresentados na Figura 6.2.2.

Para obter estes resultados, procura-se a melhor partição gama-dicotômica de cada variável preditora segundo o critério adotado (X^2), produzindo as partições relacionadas na Tabela 6.2.2.2, sendo todas avaliadas individualmente como significativas segundo o critério da Desigualdade de Bonferroni. Em seguida, deve-se escolher a partição mais significativa entre as relacionadas; porém, em vista do desconhecimento da distribuição de X^2 , utilizou-se o critério empírico: maior coeficiente de explicação (R^2).

TABELA 6.2.2.2 - Melhor partição das variáveis preditoras - 1ª etapa.

Preditora	Regra de partição	No. de poss. partições=B	Melhor partição		$X^2 (\alpha/B)$	X^2 obs.	Conclusão
			A	B			
cor	livre	2	branco	n branco	341.59	7.815	signif.
local de residência	livre	2	SP	n SP	14.23	7.815	signif.
tempo de residência	livre	2	até 3 anos	+3 anos	90.48	7.815	signif.
sexo	livre	2	homens	mulheres	1.183.88	7.815	signif.
posição na ocupação	livre	15	emp.dom.+ outras	assalariado, emp.auton.	2.238.51	13.810	signif.
idade	Monotônica	4	até 17	+ 17 anos	1.478.92	10.863	signif.
instrução	flutuante	5	até 1 grau	+ 2 grau incom.	1.412.44	11.345	signif.
Ramo	livre	7	ind.transf.	n.ind.transf.	309.14	12.121	signif.
Posição na família	livre	2	chefe	N chefe	1.781.6	7.815	signif.

Obedecendo a este critério, selecionou-se a partição correspondente à variável posição de ocupação. Tal partição satisfaz o conjunto de critérios adotados, produzindo as caselas 2 e 3 com os seguintes perfis:

TABELA 6.2.2.3 - Perfil das caselas 2 e 3.

Casela	Até 1 SM	+ 1 a 2 SM	+ 2 a 5 SM	+ 5 SM
2	0,20	0,26	0,26	0,28
3	0,81	0,10	0,01	0,08

Na casela 2 estão as pessoas alocadas nas posições de ocupação: empregador, assalariado e autônomo e na casela 3 estão as empregadas domésticas e outras posições de ocupação.

Continuando o procedimento; na tabela 6.2.2.4 estão relacionadas as partições das caselas 2 e 3 selecionadas segundo o critério X para cada variável preditora.

0

Entre as partições relacionadas, a correspondente à variável preditora Idade da casela 2 foi selecionada como a "melhor", visto atender ao critério empírico adotado. Sua significância global foi constatada pelo conjunto de critérios adotados, produzindo as caselas 4 e 5.

Agora são 3 caselas disponíveis para partição, as de número 3, 4 e 5. Selecionam-se as partições segundo o critério X^0 para cada preditora em cada casela disponível; entre essas, escolhe-se a mais significativa segundo o critério empírico, avalia-se sua significância global e, assim, o processo continua até não atender ao conjunto de critério.

Na Tabela 6.2.2.5 estão algumas estatísticas utilizadas para avaliação global da partição selecionada pelo critério empírico, nas várias etapas.

TABELA 6.2.2.4 - Relação das partições das caselas 2 e 3 segundo o critério X_0 para cada variável preditora.

Casela	Preditora	Tipo de partição	n. de partições pos.	Melhor partição		X_0 obs.
				A	B	
2	Cor	livre	1	branco	n branco	215.46
	local de resid.	livre	1	S.P.	n S.P.	25.18
	tempo de resid.	livre	1	até 3 anos	+ 3 anos	58.61
	Sexo	livre	1	homens	mulheres	556.53
	Posição de ocupação	livre	3	empregador	assal. + autonom.	378.74
	Idade	Monotônica	4	até 17 anos	+ 17 anos	1521.70
	Instrução	flutuante	5	até 1.grau completo	+ 2.grau incomp.	994.63
	Ramo	livre	7	transf.+serv	comerc. + outros	175.17
	Posição na família	livre	1	chefe	n chefe	1287.21
3	Cor	livre	1	branco	n branco	26.81
	local de resid.	livre	1	SP	n SP	19.86
	tempo de resid.	livre	1	até 3 anos	+ 3 anos	3.54
	Sexo	livre	1	homens	mulheres	118.62
	Posição de ocupação	livre	1	emp. dom.	outras	25.60
	Idade	Monotônica	4	até 24 anos	+ 24 anos	56.52
	Instrução	flutuante	5	até 1.grau completo	+ 2.grau incompleto	336.70
	Ramo	livre	7	transf. + serv.	com. + outros	23.63
	Posição na família	livre	1	chefe	n chefe	89.25

TABELA 6.2.2.5 - Valores e caselas selecionadas nas várias etapas.

Etapa	Pai	Filho	Caselas disponíveis prox. etapa	Mult. Bonferroni	Bonf. acumul.	χ^2_c	$\chi^2_{obs.}$	SSE(d)	R^2 (%)	Conclusão
1	1	2, 3	2, 3	36	36	15.820	2.238.51	468.23	9.84	
2	2	4, 5	3, 4, 5	24	60	17.039	1.521.70	165.87	3.48	
3	4	6, 7	3, 5, 6, 7	22	82	17.751	807.81	92.49	1.94	significa-
4	6	8, 9	3, 5, 7, 8, 9	21	103	18.237	771.93	112.45	2.36	tiva para
5	7	10, 11	3, 5, 8, 9, 10, 11	21	124	18.649	582.37	69.81	1.47	2 R > 0,5%
6	3	12, 13	5, 8, 9, 10, 11	22	146	19.066	336.70	21.79	0.46	
7	9	14, 15	5, 8, 10, 11, 14, 15	19	165	19.320	188.98	18.38	0.39	significa-
8	10	16, 17	5, 8, 11, 14, 15, 16, 17	16	181	19.490	184.30	23.28	0.49	tiva para
9	14	18, 19	5, 8, 11, 15, 16, 17, 18	17	198	19.741	131.15	13.96	0.29	2 R > 0,2%
10	11	20, 21	5, 8, 15, 16, 17, 18, 20, 21	19	217	19.920	110.41	17.98	0.38	
11	18	22, 23	5, 8, 15, 16, 17, 20, 21, 22, 23	16	233	20.182	100.46	13.89	0.29	
12	16	24, 25	5, 8, 15, 17, 18, 21, 22, 23, 24, 25	16	249	20.269	99.22	5.4	0.11	

Apenas a título de curiosidade, calculou-se o valor de $\lambda(K)$ acumulado apenas para confrontar o comportamento do coeficiente de explicação (R^2) acumulado. Na Figura 6.2.3 estão as curvas referentes a estas 2 estatísticas, demonstrando coerência numa possível finalização do processo na 5- etapa.

Na Tabela 6.2.2.6 estão caracterizados os grupos finais, adotando-se o conjunto de critérios. Caso se adotasse incremento superior a 0,2%, chegar-se-ia aos resultados ilustrados na Figura 6.2.2.

TABELA 6.2.2.6 - Caracterização dos grupos fixos para $R^2 > 0,5\%$.

Casela n. na fig. 6.2.2	Características do Grupo	tamanho do grupo	Proporção de pessoas			
			até 1 S.M.	+ 1 a 2 S.M.	+ 2 a 5 S.M.	+ 5 S.M.
3	empregada doméstica + outras posições	1291	81.1	9.8	1.3	7.9
5	empregador, assalariado ou autônomo, com menos de 18 anos	845	70.8	20.8	1.5	6.9
11	empregador, assalariado ou autônomo, com mais de 18 anos, não chefe, com escolaridade até 1. grau completo	2896	32.9	36.7	15.6	14.7
9	empregador, assalariado ou autônomo, com mais de 18 anos, chefe, com escolaridade até 1. grau completo.	4206	11.6	26.4	34.7	27.3
10	empregador, assalariado ou autônomo, com mais de 18 anos, não chefe, no mínimo 2. grau incompleto	1985	11.4	24.8	35.4	28.4
8	empregador, assalariado ou autônomo, com mais de 18 anos, chefe, no mínimo 2. grau incompleto.	1480	2.6	5.9	25.7	65.7

Pela Figura 6.2.2, pode-se perceber que algumas variáveis preditoras coletadas, tais como: cor, local de residência, ramo de atividade, tempo de residência não foram utilizadas nas partições.

FIGURA 6.2.3 - Valores de λ (K) e R^2 acumulado.

7 CONCLUSÃO

Este trabalho pretendeu abordar alguns aspectos da Técnica AID quando a variável resposta é categorizada. O primeiro aspecto discutido foi o tipo de restrição nas partições associado à variável preditora que neste trabalho foi designado genericamente como **Regra Gama**. Foram estudadas as 3 regras mais comuns: livre, monotônica e flutuante. Pode-se constatar através dos exemplos do Capítulo 2 que, partindo da definição da regra Gama, constrói-se o conjunto de partição $\Phi(J, \text{Gama})$ a partir do qual é feita a seleção da melhor partição segundo algum critério. Através de simulações constatou-se que a regra Gama influi na determinação de valores críticos obtidos para uma dada estatística. Destas constatações, conclui-se que a escolha da regra é importante na técnica AID, tanto na determinação dos grupos, como na distribuição da estatística utilizada.

Outro assunto estudado neste trabalho foram os critérios utilizados para avaliar a homogeneidade de uma partição fixa. Foram definidos os critérios C , X , V , θ e δ , estudadas suas propriedades e sua utilização na seleção da melhor partição. Através das simulações geradas sob diferentes escolhas do parâmetro π feitas por Light e Margolin, concluiu-se que a distribuição de C sob H_0 independe do parâmetro π , o mesmo acontecendo com a estatística X , visto que as médias, variâncias e quantis a vários níveis são bastante próximos para qualquer parâmetro π e para K e I fixos. Além disso, verificou-se também, através de simulações, que a estatística C é mais poderosa para casos unimodais, enquanto que a

estatística X é mais poderosa para os casos bimodais, pelo menos em situações com respostas tricotômicas. Assim, pode-se concluir que nenhum dos dois critérios é uniformemente mais poderoso que o outro; o poder dos dois critérios depende do valor do parâmetro considerado na hipótese alternativa.

Para casos de respostas dicotômicas ($l=2$) os valores das estatísticas X e C são bem próximos, visto que se relacionam

através da expressão
$$X = \frac{N}{N-1} \cdot C$$

Nas situações onde $K = 2$, e $N_1 = N_2$, o valor da estatística V é sempre maior que o da estatística X , visto que vale a desigualdade do Teorema de Light e Margolin.

$$2 \ln 2 \cdot X > V > X$$

Os valores das 3 estatísticas X , C e V parecem ter a mesma tendência, já que as correlações de postos calculadas para os vários pares foram todas superiores a 0.90.

Por outro lado, θ deve apresentar um comportamento diferenciado pois não se correlaciona com nenhuma das 4 estatísticas apresentadas. Já a estatística δ deve seguir levemente a tendência de X , C e V , por causa de suas correlações com esse trio que, é em torno de 0,80.

Com relação às dificuldades em identificar a distribuição das estatísticas utilizadas para selecionar a partição ótima, através de simulações, concluiu-se que as distribuições de X e V depen-

dem apenas dos parâmetros K , I e γ e devem ter comportamento muito semelhantes visto apresentarem estatísticas descritivas (tais como: média, variância, correlação de Pearson superior a 0.90) bastante próximas. Já a distribuição de C_0 , além de depender dos parâmetros mencionados, depende também de π , já que em apenas alguns casos apresentam comportamento semelhante ao de X_0 e V_0 . De modo geral, sua média é bastante próxima de X_0 e V_0 , porém sua variância é maior que as variâncias de X_0 e V_0 . Além disto, as 3 estatísticas apresentam fortes indícios de terem uma distribuição do tipo AX_{γ}^2 , fato sugerido pelas regressões lineares ajustadas: $\Phi^{-1}(F(y)) = a + b \sqrt[3]{y}$ (onde y é uma das 3 estatísticas) apresentando um alto coeficiente de explicação (R^2). Estas retas podem ser utilizadas para estimar os níveis descritivos em situações semelhantes às simuladas.

Com relação aos possíveis agrupamentos obtidos, os critérios C_0 , X_0 e V_0 devem conduzir a grupos finais muito semelhantes, já que foi alta a porcentagem de coincidência (superior a 85% de coincidência) nos casos simulados. Sobre os critérios utilizados para julgar a "vantagem" da partição ótima, concluiu-se que a Desigualdade de Bonferroni é um critério recomendado, pois apresentou valores críticos bastante próximos aos valores simulados, ao passo que o critério de Comparações Múltiplas apresentou valores críticos muito distantes (sempre maior) dos valores gerados.

Nas aplicações práticas, notou-se entretanto, a necessidade de adotar um conjunto de critérios para controlar a qualidade dos "resultados" principalmente quando se utiliza critério de seleção onde o tamanho do grupo é considerado no seu cálculo, por exem-

plo, X_0 onde valores altos desta estatística vão gerar divisões desnecessárias. Além disso, quando várias variáveis preditoras são utilizadas na análise, necessita-se adotar estratégias para selecionar a "melhor variável preditora", a "melhor casela", pois na maioria dos casos não é possível calcular o nível descritivo de cada variável preditora, dado que não se dispõe da distribuição da estatística. A solução mais usual é adotar um critério empírico ou selecionar através da significância de Bonferroni. Esta última disponível somente no programa CHAID.

APÊNDICE

A1 - Demonstração do teorema (Light e Margolin):

Teorema. Para qualquer tabela 2×1 , com $N_1 = N_2 = N/2$, vale a desigualdade

$$(2 \ln 2) * X_i > V_i > X_i$$

Demonstração:

Para uma tabela 2×1 , com $N_1 = N_2 = N/2$ e $n_i > 0$, para todo i , tem-se

$$X_i = \sum_{i=1}^I X_i$$

$$\text{onde } X_i = \sum_{k=1}^2 [n_{ki} - (n_i/2)]^2 / (n_i/2)$$

$$= (n_{1i} - n_{2i})^2 / (n_{1i} + n_{2i})$$

X_i é a contribuição do nível de resposta i no cálculo de X .

De modo semelhante, V pode ser escrito como:

$$V = \sum_i V_i$$

$$\text{onde } V_i = 2 \sum_{k=1}^2 n_{ki} \ln [2n_{ki} / (n_{1i} + n_{2i})] \text{ e } 0 \ln 0 \text{ é definido como zero}$$

V_i é a contribuição do i -ésimo nível de resposta no cálculo da estatística V .

Aqui será provado um resultado mais forte que o teorema, que para todo i , vale a desigualdade:

$$(2 \ln 2) * X_i > V_i > X_i$$

A igualdade da direita só ocorrerá se e somente se $n_{k1} = n_{k2}$

$= n_i / 2$, daí $X_i = 0 = V_i$; a igualdade da esquerda ocorrerá se $n_i = 0$ ou $n_{k2} = 0$, mas nunca ambos, resultando $V_i = n_i (2 \ln 2)$ e $X_i = n_{k1}$

A fim de facilitar o entendimento da demonstração, será suprimido o índice i , conforme se segue:

$$X = (n_1 - n_2)^2 / (n_1 + n_2) \quad (1)$$

$$V = 2 \{ n_1 \ln n_1 + n_2 \ln n_2 - (n_1 + n_2) \ln [(n_1 + n_2)/2] \} \quad (2)$$

Considere a função $x \ln x$, $x > 0$

$$\frac{\partial x \ln x}{\partial x} = 1 + \ln x$$

$$\frac{\partial^{(j)} x \ln x}{\partial x} = (-1)^j (j-2)! x^{-(j-1)}$$

$x \ln x$ pode ser expandido numa série de Taylor em torno de $x_0 > 0$ arbitrário.

Então,

$$x \ln x = x_0 \ln x_0 + (x - x_0) (1 + \ln x_0) + \sum_{j=2}^{\infty} \frac{(x-x_0)^j (-1)^j}{x_0^{j-1} (j-1)!}$$

com o lado direito convergente para $2x_0 > x > 0$.

$$\text{Seja } x_0 = (n_1 + n_2)/2 = \bar{n}$$

$$\begin{aligned} & \text{Expandindo (2) numa série de Taylor, obtém-se } n_1 \ln n_1 + \quad (3) \\ n_2 \ln n_2 &= 2 \bar{n} \ln \bar{n} + \sum_{k=1}^2 (n_k - \bar{n}) (1 + \ln \bar{n}) + \sum_{j=2}^{\infty} \sum_{k=1}^2 \frac{(n_k - \bar{n})^j (-1)^j}{(\bar{n})^{j-1} (j-1)!} \end{aligned}$$

Como $(n_1 - \bar{n}) + (n_2 - \bar{n}) = 0$ e $(n_1 - \bar{n})^j + (n_2 - \bar{n})^j = 0$ para j ímpar

$$\text{e } (n_1 - \bar{n})^{2r} = (n_2 - \bar{n})^{2r} = \frac{(n_1 - n_2)^{2r}}{2}$$

Daí, substituindo em (3),

$$\text{resulta em } n_1 \ln n_1 + n_2 \ln n_2 = (n_1 + n_2) \ln \left(\frac{n_1 + n_2}{2} \right) +$$

$$2 \sum_{r=1}^{\infty} \frac{(n_1 - n_2)^{2r}}{(n_1 + n_2)^{2r-1} (2r) (2r-1)}$$

Daí,

$$V_i = 2 \left(n_1 \ln n_1 + n_2 \ln n_2 - (n_1 + n_2) \ln \left(\frac{n_1 + n_2}{2} \right) \right) =$$

$$\frac{(n_1 - n_2)^2}{n_1 + n_2} + 2 \sum_{r=2}^{\infty} \frac{(n_1 - n_2)^{2r}}{(n_1 + n_2)^{2r-1} (2r) (2r-1)} \quad (4)$$

O primeiro termo de (4) é X_i e o segundo termo é não negativo. Só será igual a zero se $n_1 = n_2$

Portanto, $V_i > X_i$.

Além disso,

$$V_i \leq 2 \frac{(n_1 - n_2)}{(n_1 + n_2)} \left[\frac{1}{2 \times 1} + \sum_{r=2}^{\infty} \frac{1}{2r(2r-1)} \right]$$

Só atinge a igualdade se e somente se $|n_1 - n_2| = n_1 + n_2$, ou seja, ou $n_1 = 0$ ou $n_2 = 0$. Segue-se portanto, que:

$$V_i \leq 2 \frac{(n_1 - n_2)^2}{(n_1 + n_2)} \left[\sum_{r=1}^{\infty} \frac{1}{2r(2r-1)} \right] =$$

$$= 2 \sum_{r=1}^{\infty} \frac{(-1)^{r+1}}{r} = 2 \sum_i \ln 2$$

Daf, provou-se que:

$$2 \sum_i \ln 2 > V > \sum_i X$$

Somando em i ; obtém-se que

$$2 \sum_i \ln 2 > V > \sum_i X$$

TABELA A.1 - Resultados observados de X_0 , C_0 e V_0 - $K = 3$ e $I = 3$.

π	Estatística simulada	Regra Livre					Regra Monotônica				
		Estatísticas		Quantis a			Estatísticas		Quantis a		
		Média	Variância	90%	95%	99%	Média	Variância	90%	95%	99%
.94,.03,.03	X_0	3.27	5.46	6.13	7.71	11.32	2.84	4.97	5.61	7.06	10.72
	C_0	3.31	6.50	6.73	8.68	12.51	2.86	5.60	5.76	7.06	12.03
	V_0	3.43	6.28	6.61	8.40	12.98	2.97	5.74	6.12	7.44	11.69
1/3;1/3;1/3	X_0	3.38	5.63	6.60	7.91	10.59	2.94	5.09	6.06	7.70	10.36
	C_0	3.38	5.68	6.63	7.97	10.52	2.94	5.14	6.08	7.72	10.52
	V_0	3.39	5.73	6.72	8.00	10.72	2.95	5.21	6.12	7.77	10.37
.48,.48;.03	X_0	3.20	4.52	6.25	7.18	9.53	1.84	4.10	5.68	6.94	9.12
	C_0	3.52	9.66	7.44	10.39	13.74	3.07	8.15	6.62	9.40	12.68
	V_0	3.23	4.59	6.47	7.34	9.48	2.88	4.22	5.83	6.98	9.12
.69,.23;.08	X_0	3.14	5.31	6.15	7.54	10.04	2.75	4.77	5.62	6.92	9.42
	C_0	3.15	7.36	6.51	8.95	11.96	2.75	6.28	5.88	7.70	11.76
	V_0	3.18	5.59	6.33	7.40	9.95	2.77	4.96	5.45	7.01	9.73

TABELA A2 - Resultados observados de C, X e V para K = 3 e
I = 5.

π	Estatística simulada	Regra Livre					Regra Monotônica				
		Estat. descritivas		Quantis			Estat. descritivas		Quantis		
		Média	Variância	90%	95%	99%	Média	Variância	90%	95%	99%
0,40	X	5,86	8,89	9,79	11,58	14,49	5,29	8,64	9,05	10,90	14,30
0,15	C	5,97	11,80	10,29	13,06	17,26	5,38	11,57	9,43	12,48	17,26
0,15	V	5,90	9,08	9,66	11,63	14,97	5,32	8,80	9,29	11,24	14,69
0,6	X	6,10	9,76	10,28	11,77	16,48	5,55	9,72	9,91	11,61	15,14
0,1	C	6,21	14,06	11,53	13,93	18,35	5,58	13,40	10,37	12,71	18,13
0,1	V	6,18	10,25	10,36	11,97	15,89	5,62	10,28	9,83	11,70	15,53
0,8	X	6,10	10,60	10,52	12,12	15,08	5,49	9,82	9,77	11,17	14,06
0,05	C	6,27	19,14	11,54	13,97	21,25	5,56	17,43	10,14	12,81	19,81
0,05	V	6,26	11,63	10,76	12,35	16,45	5,60	10,46	9,97	11,74	14,71
0,40	X	6,06	10,23	10,24	12,02	15,82	5,39	9,13	9,32	10,78	14,99
0,40	C	6,51	22,34	12,06	15,27	21,39	5,66	18,23	10,71	14,05	21,35
0,066...0,066	V	6,17	11,21	10,50	11,99	16,38	5,49	10,06	9,70	11,42	15,78
0,45	X	6,20	10,85	10,33	12,83	16,64	5,41	10,17	9,24	11,30	16,47
0,45	C	6,62	32,15	13,71	17,89	28,50	5,66	26,42	12,26	15,91	25,52
0,03...0,03	V	6,42	12,40	10,40	13,23	17,97	5,59	11,31	9,75	11,78	16,52

TABELA A.3 - Resultados observados de X_0 , C_0 , V_0 para $K = 3$ e vários I .

π	Estatística	Regra Livre					Regra Monotônica				
		Estat. descritivas		Quantis			Estat. descritivas		Quantis		
		Média	Variância	90%	95%	99%	Média	Variância	90%	95%	99%
1/3 1/3 1/3	X_0	3,27	5,46	6,13	7,71	11,32	2,94	5,09	6,06	7,70	10,36
	C_0	3,31	6,50	6,73	8,68	12,51	2,94	5,14	6,08	7,72	10,52
	V_0	3,43	6,28	6,61	8,40	12,98	2,95	5,21	6,12	7,72	10,37
1/4 1/4 1/4 1/4	X_0	4,62	7,39	8,20	9,38	13,93	4,11	6,95	7,51	8,94	12,97
	C_0	4,62	7,39	7,92	9,42	13,91	4,10	6,89	7,49	9,06	13,10
	V_0	4,64	7,58	8,28	9,52	13,92	4,12	7,05	7,59	9,12	13,08
1/5 1/5 1/5 1/5 1/5	X_0	5,69	8,00	9,68	11,21	14,36	5,10	7,86	8,71	10,77	14,36
	C_0	5,67	8,02	9,80	11,18	13,53	5,08	7,79	8,84	10,55	13,53
	V_0	5,72	8,14	9,84	11,49	14,17	5,13	8,01	8,79	11,06	13,94
1/8 ... 1/8	X_0	9,67	14,99	14,66	16,31	20,84	8,96	14,49	13,99	15,73	20,51
	C_0	9,70	15,36	14,87	16,40	21,03	8,96	14,69	14,18	15,91	20,41
	V_0	9,78	15,60	14,79	16,62	21,04	9,04	15,06	14,27	16,03	20,11
1/10 ... 1/10 ($n_R = 200$)	X_0	11,70	16,65	16,89	19,01	23,32	10,66	16,61	15,98	18,31	21,31
	C_0	11,73	17,46	17,09	19,19	24,52	10,68	17,38	16,20	18,28	22,21
	V_0	11,83	17,17	17,09	19,27	23,58	10,78	17,21	16,25	18,73	21,71
1/10 ... 1/10 ($n_R = 400$)	X_0	12,08	17,47	17,79	19,83	24,26	10,97	16,96	16,49	18,69	23,01
	C_0	12,06	17,23	17,67	19,48	23,09	10,96	16,91	16,63	18,83	22,71
	V_0	12,16	17,87	17,93	20,03	24,62	11,03	17,36	16,59	18,57	23,51

TABELA A.4 - Resultados observados de X , C e V para $K = 5$ e
vários l .

π	Estatística	Regra livre					Regra Monotônica				
		Estat.descritivas			Quantis		Estat.descritivas		Quantis		
		Média	Variância	90%	95%	99%	Média	Variância	90%	95%	99%
1/3 ... 1/3	X	5,52	8,53	9,30	11,24	14,58	2,86	6,21	5,93	7,34	12,07
	O										
	C	5,52	8,39	9,37	11,13	14,71	2,85	6,15	5,98	7,20	11,75
	O										
	V	5,54	8,61	9,41	11,32	14,70	2,86	6,34	5,92	7,33	12,16
	O										
1/4 ... 1/4	X	7,16	9,57	11,04	12,78	16,60	3,94	8,87	8,14	10,22	13,92
	O										
	C	7,13	9,48	10,92	12,43	16,62	3,91	8,74	8,05	10,05	13,46
	O										
	V	7,19	9,75	11,24	12,74	16,60	3,93	8,88	8,01	10,19	13,45
	O										
1/5 ... 1/5	X	9,05	12,73	13,96	16,05	21,00	5,22	12,41	9,98	11,82	15,48
	O										
	C	9,05	12,75	13,65	16,01	21,22	5,25	13,01	10,02	11,77	17,69
	O										
	V	9,11	13,23	13,73	16,21	21,42	5,30	13,70	10,10	11,70	16,88
	O										
1/8...1/8	X	13,32	16,01	18,87	20,70	25,12	8,36	19,04	14,32	16,96	20,04
	O										
	C	13,31	16,48	18,76	20,76	25,33	8,32	19,13	14,29	16,56	20,86
	O										
	V	13,44	16,88	18,73	20,90	25,50	8,44	20,16	14,47	17,28	20,06
	O										
1/10 1/10	X	15,91	16,60	21,09	22,79	26,26	10,81	22,76	17,04	19,82	23,58
	O										
	C	15,95	17,22	21,15	22,99	26,37	10,74	23,24	16,76	19,42	24,31
	O										
...1/10	V	16,07	17,43	21,31	22,93	26,72	10,83	23,62	17,13	19,10	24,37
	O										

BIBLIOGRAFIA

- ANDERSON R.J.; LANDIS, J.R. (1980) - Catanova for multidimensional contingency tables: Nominal scale response / Communication Statistics - Theory and methods, 1191-1206.
- ANDERSON, R.J. & LANDIS, J.R. (1982) - Catanova for Multidimensional Contingency tables: ordinal scale response - Communication Statistics - Theory and Methods -, 257-270.
- ANDREWS, F.M.; MORGAN, J.N.; SONQUIST, J.A., KLEIN, L. (1973) - Multiple classification analysis Institute for Social Research,^a The University of Michigan, 2- edição.
- ANDREWS, F.M.; MESSENGER, R.C. (1973) - Multivariate nominal scale analysis - Survey Research Center - The University of Michigan. Michigan.
- BUSSAB, W. O. (1976) - Hierarchical dichotomous partitions in cluster analysis, Ph.d. Thesis. - University of London.
- CARMER, S.G.; LIN, W.T. (1983) - Type I error rates for divisive clustering methods for grouping means in analysis of variance - Communication Statistics - Simulation and Computation. 456-466.
- ELPHINSTONE C.D.; LONERGAN A.T.; FATTI, L.P.; HAWKINS, D.M. (1985) - An empirical investigation into the application of some statistical techniques to the classification of remotely sensed data - Internal Report - CSIR - n 40 - Pretoria.

- ENGELMAN, L.; HARTGAN, J.A. (1969) - Percentage points on a test for cluster - J.A.S.A. - 1647-1648.
- FIELDING, A.; O'MUIRCHEARTAIGH, C.A. (1976) - Binary segmentation in survey analysis with particular reference to AID - The Statistician, vol. 26, n.1, 17-28.
- FIELDING, A. (1977) - Binary segmentation: The Automatic Interaction Detector and Related Techniques for Exploring Data Structure. in The Analysis of Survey Data - vol. I - Exploring data structures. Edited by Colm A. O'Muircheartaigh and Clive Payne.
- FLEISS, J.L. (1973) - Statistical Methods for Rates and Proportions. New York - Wiley.
- GABRIEL, K.R. (1966) - Simultaneous test procedures for multiple comparisons on categorical data - J.A.S.A. - 1081-1096.
- GILULA, Z. (1985) - On the analysis of heterogeneity among populations - J.R. Statistic Society, 47, n.1, 76-83.
- GINI, C.W. (1938) - Variabilita e mutabilita contributo allo studio delle distribuzioni e relazioni statistiche, Studi Economico-Giuridice della R. Universita di Cagliari.
- GOODMAN, L.A. & KRUSKAL, W.H. (1954) - "Measures of Association for Cross Classification" - J.A.S.A., 49, 732-64.
- GROSS, W.F. (1984) - A note on Chi-Squared tests with survey data - J.R. Statistical Society - B 46, n.2, 270-272.

- HALPERN, J. (1982) - Maximally selected chi-square statistics for small samples - *Biometrics*, 38-1017-1023.
- HAWKINS, D.M.; KASS, G.V. (1982) - Automatic Interaction Detection in Topics in Applied Multivariate Analysis - (Chapter IV) - 269-302. C.U.P.
- HEYMANN, C. (1981) - XAID - An extended automatic interaction detector - Internal Report - CSIR - n 28 - Pretoria.
- HOLT, D.; SCOTT, A.J.; EWINGS, P.D. (1980) - Chi-squared tests with survey data - *J.R. Statistical Society - A*, 143, part 3, 303-320.
- ISRAELS, A.Z.; VAN DRIEL, J. (1983) - Use of the chi-square statistic for selecting explanatory variables in multiway tables - *Quality and Quantity*, 17, 103-110.
- KASS, G.V. (1980) - An exploratory technique for investigating large quantities for categorical data - *Applied Statistics*, 29, n.2, 119-127.
- KASS, G.V. (1980) - Material didático que acompanha o programa CHAID - CSIR - Pretoria.
- KENDALL, M.G. & STUART, A. (1973) - The advanced theory of statistics, vol. 1, 2, 3 - London, Charles Griffin and Co. Ltd.

- KULLBACK, S. (1959) - Information Theory and Statistics - N. York, John Wiley.
- LIGHT, R.J. & MARGOLIN, B.H. (1971) - An analysis of variance for categorical data - J.A.S.A. - vol. 66, n.335, 534-544.
- MARGOLIN, B.H. & LIGHT, R.J. (1974) - An analysis of variance for categorical data, II: Small sample comparisons with chi-square and other competitor - J.A.S.A. - vol. 69. N 347 - 755-764.
- MESSENGER, R.; MENDELL, L. (1972) - A modal search technique for predictive nominal scale multivariate analysis. J.A.S.A. - vol. 67 - 768-772.
- MILLER R.; SIEGMUND, D. (1982) - Maximally selected chi-squared statistics - Biometrics 38, 1011-1016.
- MORGAN, J.M. & MESSENGER, R.C. (1973) - THAID - A sequential analysis program for the analysis of nominal scale dependent variables - Survey Research Center - Institute for Social Research - The University of Michigan - Ann. Arbor, Michigan,
- MORGAN, J.N.; SONQUIST, J.A. (1963) - Problems in the analysis of survey data - J.A.S.A.-58, 415-434.
- SHAFFER, J.P.; GILLO, M.W. (1974) - A multivariate extension of the correlation ration (M-AID) - Educational and Psychological

Measurement - 34, 521-524.

SONQUIST, J.A.; BAKER, E.L.; MORGAN, J.N. (1973) - Searching for structure - Survey Research Center - The University of Michigan - Ann. Arbor, Michigan.

THEIL, H. (1970) - On the estimation of relationships involving qualitative variables - American Journal of Sociology - vol. 76, n.1 - 103-154.

THOMPSON, V.R. (1972) - Sequential dichotomisation two techniques - The Statistician - vol. 21 n.3 - 181-194.