UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Análise de Variância para Dados Categóricos: Uma Aplicação em Genética

Autora: Talita Armborst

Orientadora: Profª Sidia Maria Callegari Jacques

Monografia apresentada para
a obtenção do Título de
Bacharel em Estatística

Porto Alegre, março de 2003.

Para minha mãe,
Vera, uma mulher
maravilhosa.

AGRADECIMENTOS

A minha mãe Vera Rejane Armborst, pela ajuda nos momentos difíceis, compreensão pelos muitos momentos que quis, mas não pude, estar presente. E por ser uma grande mulher, amiga e mãe.

Ao meu namorado Guilherme Ehrenbrink, pelo carinho e atenção quando mais precisava. Além da contribuição na monografia na parte da genética.

A minha mestra, Sidia Maria Callegari Jacques, por sua energia e disposição e pelos conhecimentos na área de Genética e Estatística, demonstrados no meu tempo de Bolsista e na criação desta Monografia.

À professora Maria Cátira Bortolini, por ceder o material dos Índios, utilizado como um dos exemplos deste trabalho e por seu auxílio na parte de Genética de Populações e de DNA.

Aos professores do Departamento de Estatística que me transmitiram o conhecimento necessário para exercer esta profissão e saber um pouco mais da vida.

Aos colegas, pela ajuda nos estudos e pelas festas compartilhadas.

E a todas as pessoas que contribuíram nesta fase da minha vida.

RESUMO

A Análise de Variância para Dados Categóricos (CATANOVA) é similar à Análise de Variância para Dados Contínuos com apenas uma variável resposta (ANOVA), com a diferença que naquela, a variável resposta é qualitativa. Light e Margolin (1971) criaram a CATANOVA para um fator (tabelas bidimensionais) com o objetivo de testar a igualdade de duas ou mais proporções populacionais. Um dos subprodutos da ANOVA é a estatística R^2, que mede a proporção da variação total na variável resposta explicada por um fator. Na CATANOVA, foi criada a estatística C a partir de R^2, que mede a associação entre um fator e uma variável resposta categórica. A estatística **C** é testada através da distribuição Qui-Quadrado. Anderson e Landis (1980) expandiram a CATANOVA para mais de um fator (tabelas multidimensionais), onde um fator seria o de maior interesse e os outros entrariam como fatores de controle. Neste caso, é possível medir a associação (R^2) para cada um dos fatores sem controlar pelos outros, para o fator de maior interesse controlando pelos demais fatores e para a combinação de todos os fatores. Estas medidas de associação são mais adequadas do que o χ^2 calculado pois freqüentemente os dados não satisfazem as suposições para este teste. A área de genética pode vir a utilizar a CATANOVA em grande escala, pois nela é comum se ter uma variável resposta nominal, como nos exemplos utilizados neste trabalho: tipo de gônadas (gônadas normais e gônadas disgênicas) e tipo de nucleotídeos (A, C, T, G). No último caso, a CATANOVA, como uma nova forma de comparar populações com base em seqüências genômicas, mostrou-se extremamente útil ao aumentar o poder discriminatório do Fator 1 (população indígena). Em ambos os exemplos, a CATANOVA extraiu mais informações dos dados do que as análises tradicionais. Dessa forma, a CATANOVA é uma ferramenta importante e deve ser utilizada na análise de dados genéticos.

SUMÁRIO

ÍNDICE DE TABELAS

Capítulo 3: CATANOVA PARA UM FATOR: ASPECTOS TEÓRICOS

Capítulo 4: CATANOVA PARA MAIS DE UM FATOR: ASPECTOS TEÓRICOS

Capítulo 5: CATANOVA APLICADA À GENÉTICA DE POPULAÇÕES

1. INTRODUÇÃO

A Análise de Variância tem como objetivo testar a igualdade de três ou mais tratamentos para uma variável resposta.

A mais difundida é a ANOVA (*Analysis of Variance*), que é a análise de variância para dados contínuos. Ela testa a igualdade de três ou mais médias populacionais com base na partição da variabilidade total nos componentes associados a fontes identificáveis de variação. Na sua forma mais simples, a soma de quadrados total (SQTotal) é dividida em soma de quadrados entre tratamentos (SQT) e a soma de quadrados dentro dos tratamentos (SQE).

Uma outra análise de variância pouco conhecida é a Análise de Variância para Dados Categóricos, também conhecida como CATANOVA (de *Categorical ANOVA*), que é a análise de variância para variáveis nominais (variáveis qualitativas não-ordenadas). Ela é similar à ANOVA, contudo a CATANOVA testa a igualdade de duas ou mais proporções populacionais.

Outra diferença entre a ANOVA e a CATANOVA é a de que na primeira os componentes SQT e SQE são independentes, e conseqüentemente SQT e o SQTotal não o são. Somente SQT e SQTotal serão independentes assintoticamente quando houver um grande tamanho de amostra em cada tratamento (Light e Margolin,1971).

Na CATANOVA, os dados categóricos são organizados em tabelas de contingência. No caso de tabelas bidimensionais, uma marginal é fixa e a outra é variável. A marginal fixa pode ser chamada de Fator 1 e seus níveis identificam diferentes grupos, categorias de um possível fator ou amostras de diferentes populações. A marginal variável representa as categorias de resposta. Este cenário é típico de ensaios comparativos ou testes de homogeneidade entre populações.

Em 1971, Light e Margolin desenvolveram a CATANOVA para dados agrupados em tabelas bidimensionais com apenas um fator, a partir do Índice de Biodiversidade, criado por Gini (1912) e Simpson (1949) para medir variabilidade em variáveis categóricas, e de idéias oriundas da análise de

variância para dados contínuos. Eles partiram do conceito de componentes de variação sob um modelo multinomial para criar a estatística **C** e uma medida de associação chamada R^2.

O objetivo de Light e Margolin foi o de criar uma medida de associação de fácil interpretação para contrabalançar a excessiva ênfase dada, na literatura, aos testes de significância em tabelas de contingência. Foi a primeira medida de associação para dados categóricos que mede a proporção da variação total na variável resposta explicada pelo fator.

Estes autores ainda compararam o poder da estatística **C** com a χ^2 em grandes amostras e mostraram a relação entre estas duas estatísticas quando a variável resposta é dicotômica.

Posteriormente, para o caso específico de pequenas amostras, Margolin e Light (1974) compararam a estatística **C** com a χ^2 e a estatística de informação (2Î) de Kullback (1962), que é idêntica a duas vezes o logaritmo natural da razão de verossimilhança. Também compararam a medida de associação R^2 com o Lambda de Goodman e Kruskal (t_b) (1954).

A CATANOVA para dados organizados em tabelas de contingência multidimensionais envolvendo vários fatores foi desenvolvida para o caso da variável resposta ser em escala nominal por Anderson e Landis (1980), e em escala ordinal por Anderson e Landis (1982). Estes autores reforçaram a importância da medida de associação como forma de se obter mais informações dos dados, através de suas formas como medida de associação simples, múltipla e parcial para dados categóricos.

A CATANOVA tem grande possibilidade de utilização em estudos genéticos, porém aparentemente poucos pesquisadores estão se dedicando a produzir soluções usando esta técnica, como Andrade e Pinheiro (2002) que enfoca seqüências genômicas no vírus HIV. Por esta razão, no presente trabalho a CATANOVA é abordada com o auxílio de dois trabalhos com assuntos bem distintos.

2. OBJETIVOS

A CATANOVA foi criada em 1971 visando a encontrar uma alternativa para uma análise de variância quando a variável resposta não é quantitativa, mas categórica nominal.

Apesar de ter sido criada há mais de três décadas atrás, ela é pouco difundida e se utiliza erroneamente o Teste Qui-Quadrado ignorando algumas suposições necessárias para seu uso em tabelas de contingência.

Uma das vantagens da CATANOVA é utilizar a medida de associação R^2 para dados categóricos, a qual exprime a proporção da variação explicada pela variável de interesse.

Uma das áreas em que a CATANOVA pode ser difundida é a Genética. Assim, as aplicações no presente trabalho são voltadas para esta área, que está tendo grande expansão nas últimas décadas e necessita cada vez mais de análises adequadas para cada caso.

Dessa forma, o objetivo do presente trabalho é apresentar, de forma didática, as idéias que nortearam a construção da técnica e os procedimentos para sua execução, ilustrando-os com dois exemplos de dados reais da pesquisa em Genética. No primeiro, o objetivo é avaliar se existe diferença no tipo de gônadas da descendência da mosca-das-frutas *Drosophila willistoni*, dependendo da linhagem e da temperatura utilizadas. O segundo visa analisar se existe diferença entre três tribos indígenas brasileiras (Gavião, Xavante e Zoró) com base em seqüências genômicas de DNA mitocondrial.

3. CATANOVA PARA UM FATOR: ASPECTOS TEÓRICOS

3.1. Primeiros passos: índice de Biodiversidade de Gini-Simpson

O primeiro passo para a concretização do teste para CATANOVA foi dado em 1938 por Gini que criou um Índice para medir a variabilidade em uma amostra, usando dados qualitativos. Em 1949, Simpson propôs o mesmo índice aparentemente com o desconhecimento do Índice de Gini para um contexto ecológico. Por esta razão, este índice é conhecido atualmente como Índice de Biodiversidade Gini-Simpson:

$$I_S(p) = 1 - \sum_{i=1}^{I} p_i^2 \qquad \textbf{3. 1}$$

no qual $p_i=N_i/N$ é a probabilidade da categoria i, sendo que a espécie varia de 1 a I.

Já que não é possível utilizar a média e a variância para dados categóricos, a idéia destes autores foi criar uma medida de variação para o caso de ausência de concentração, com base apenas na freqüência observada em cada categoria.

Gini observou que a soma dos quadrados dos desvios da média para n medidas quantitativas pode ser expressa somente como uma função dos quadrados da diferença par a par para todos $\binom{n}{2}$ pares.

Se as medidas quantitativas são denotas por X_1, X_2,...,X_n, então a **soma de quadrados dos desvios** será:

$$SQ = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \qquad \textbf{3. 2}$$

onde: $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, e d_{ij} é a Distância Euclidiana: $d_{ij} = X_i - X_j$.

Um segundo passo foi dado por Gini em 1938 com o desenvolvimento de uma medida de variação para dados qualitativos a partir dos resultados de 1912. As medidas quantitativas passaram a ser medidas qualitativas referentes a uma categoria das I possíveis.

Cada uma das respostas qualitativas deve pertencer a somente uma das I categorias possíveis.

A $d_{ij}=f(X_i,X_j)$ passa a ser dicotômica na qual:

$$d_{ij} \begin{cases} 1 & \text{, se } X_i \text{ e } X_j \text{ indicarem categorias diferentes} \\ 0 & \text{, se } X_i \text{ e } X_j \text{ indicarem a mesma categoria} \end{cases} \qquad \textbf{3. 3}$$

Logo, a **variância** para respostas categóricas $X_1,..,X_n$ será:

$$\frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}^2 = \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij} \qquad \textbf{3. 4}$$

Os dados são reunidos em categorias em que n_i é o número de respostas na categoria *ith*, i=1,..,I. Através de um vetor Φ, reúnem-se todos n_i: $\Phi=(n_1,...,n_I)$; de forma que o somatório de todos n_i é o tamanho da amostra: $\sum_{i=1}^{I} n_i = n$.

Assim, a **variação total das respostas** pode ser descrita como:

$$D_n = \frac{1}{2n}\left[\sum_{i\neq j} n_i n_j\right] = \frac{1}{2n}\left[n^2 - \sum_{i=1}^{I} n_i^2\right] = \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{I} n_i^2 \qquad \textbf{3. 5}$$

E o tamanho amostral é:

$$n = \sum_{j=1}^{J} n_{+j} = \sum_{i=1}^{I} n_{i+} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \qquad \textbf{3. 6}$$

no qual n_{ij} representa o número de respostas na categoria i para o tratamento j, j=1,..,J.

3.2. Componentes de variação em tabelas de contingência

Light e Margolin (1971) desenvolveram a Análise de Variância para Dados Categóricos dispostos em tabelas de contingência bidimensional com um fator como mostrado na Tabela 3.1.

Tabela 3. 1. Tabela de Contingência Bidimensional utilizada em CATANOVA.

	Categorias da Variável Resposta				
Tratamento	1	2		I	Total
1	n_{11}	n_{21}	...	N_{I1}	n_{+1}
2	n_{12}	n_{22}	...	n_{I2}	n_{+2}
...	...	...	...	...	...
J	n_{1J}	n_{2J}	...	n_{IJ}	n_{+J}
TOTAL	n_{1+}	n_{2+}	...	n_{I+}	$n_{++}=n$

Para criar a CATANOVA, eles se basearam no Índice de Biodiversidade Gini-Simpson partindo da variação total das respostas dada na equação 3.5.

Na realização do modelo proposto por Light e Margolin deve-se assumir que a Categoria e Tratamento são estocasticamente independentes.

- **Soma Total dos Quadrados ou Variação Total (SQTotal)**

Partindo da variância total das respostas (D_n) criada por Gini, a Soma Total dos Quadrados será.

$$SQTotal = \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{I} n_{i+}^2 \qquad \textbf{3. 7}$$

sendo i: $n_{i+} = \sum_{j=1}^{J} n_{ij}$, i=1,..., I, o total de elementos na coluna.

Outra forma de escrever a fórmula SQTotal é:

$$SQTotal = \frac{n}{2} - \frac{1}{2n}\sum_{i=1}^{I}\left(\sum_{j=1}^{J} n_{ij}\right)^2 \qquad \textbf{3. 8}$$

- **Soma dos Quadrados dentro dos Tratamentos ou a Variação dentro dos Tratamentos (SQE) (Erro ou Resíduo)**

Partindo da equação 3.4, a variação na variável resposta dentro do tratamento *jth* é dada por:

$$\frac{n_{+j}}{2} - \frac{1}{2n_{+j}} \sum_{i=1}^{I} n_{ij}^2 \qquad \textbf{3. 9}$$

sendo: $\mathbf{n_{+j}} = \sum_{i=1}^{I} n_{ij}$, j=1,...,J, o total de elementos na linha j.

Para todos os tratamentos, a soma dos quadrados dentro dos tratamentos será:

$$SQE = \sum_{j=1}^{J} \left(\frac{n_{+j}}{2} - \frac{1}{2n_{+j}} \sum_{i=1}^{J} n_{ij}^2 \right) = \frac{n}{2} - \frac{1}{2} \sum_{j=1}^{J} \frac{1}{n_{+j}} \sum_{i=1}^{I} n_{ij}^2 \qquad \textbf{3. 10}$$

Para modelos balanceados onde $n_{+j}=n_+$ (todos os n_j são iguais), SQE se simplifica para:

$$SQE = \frac{1}{2n_+} \left(Jn_+^2 - \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij}^2 \right) \qquad \textbf{3. 11}$$

- **Soma de Quadrados entre Tratamentos ou Variação entre Tratamentos (SQT)**

A soma de quadrados entre tratamentos é obtida pela diferença entre a soma de quadrados total (SQTotal) e a soma de quadrados dentro de tratamentos (SQE):

$$SQT = SQTotal - SQE = \frac{1}{2} \left(\sum_{j=1}^{J} \frac{1}{n_{+j}} \sum_{i=1}^{I} n_{ij}^2 \right) - \frac{1}{2n} \sum_{i=1}^{I} n_{i+}^2 \qquad \textbf{3. 12}$$

sendo $\mathbf{n_{+i}} = \sum_{j=1}^{J} n_{ij}$, i=1,..., I

$\mathbf{n_{+j}} = \sum_{i=1}^{I} n_{ij}$, j=1,...,J.

Para modelos balanceados onde $n_{+j}=n_{+}$ (todos os n_j são iguais):

$$SQT = \frac{1}{2n_{+}J}\left[J\left(\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij}^2\right) - \sum_{i=1}^{I} n_{i+}^2\right] \qquad \textbf{3. 13}$$

- **Graus de Liberdade**

Os graus de liberdade para a CATANOVA são análogos aos da ANOVA: **(n-1)** para a variação total, **(n-J)** para variação dentro dos tratamentos e **(J-1)** para a variação entre tratamentos.

- **Quadrados Médio:**

São obtidas pela razão entre soma de quadrados (SQ) e constantes de proporcionalidade que podem ser usadas como análogas aos graus de liberdade da ANOVA "One-Way". Assim,

Quadrado Médio Total: $QMTotal = \frac{SQTotal}{(n-1)}$ **3. 14**

Quadrado Médio dentro dos Tratamentos: $QME = \frac{SQE}{(n-J)}$ **3. 15**

Quadrado Médio entre os Tratamentos: $QMT = \frac{SQT}{(J-1)}$ **3. 16**

3.3. Medida de Associação na CATANOVA

Na Análise de Variância para Dados Contínuos, a medida de associação R^2 é a proporção da variação total (SQTotal) explicada pela variação entre os tratamentos (SQT).

Light e Margolin (1971) aplicaram esta definição de R^2 para dados contínuos na Análise de Variância para Dados Categóricos.

$$R^2 = \frac{SQT}{SQTotal} = \frac{\left(\sum_{j=1}^{J} \frac{1}{n_{+j}} \sum_{i=1}^{I} n_{ij}^2\right) - \frac{1}{n} \sum_{i=1}^{I} n_{i+}^2}{n - \frac{1}{n} \sum_{i=1}^{I} n_{i+}^2} \qquad \textbf{3. 17}$$

onde $\mathbf{n_{+i}} \sum_{j=1}^{J} n_{ij}$, i=1,..., I

$\mathbf{n_{+j}} = \sum_{i=1}^{I} n_{ij}$, j=1,...,J.

Logo, R^2 mede a associação entre os tratamentos e as categorias da variável resposta.

As propriedades da medida de associação R^2 são as seguintes:

1) R^2 varia de 0 a 1.

2) R^2 =0 se $\frac{n_{ij}}{n_{+j}} = f_i$, ou seja, não há associação entre os tratamentos e as variáveis respostas. Note que f_i é a freqüência média na categoria i.

2) R^2 =1 se para cada j, j=1,...,J existe um *i*, i=1,..., I no qual $n_{ij}=n_{+j}$, ou seja, há total associação entre os tratamentos e as variáveis respostas.

3) R^2 é indefinido e todos os componentes são zeros se existe uma categoria i que possua todos os elementos estudados, $n_{i+}=n$.

4) R^2 não muda se todos os n_{ij} são multiplicados pelo mesmo valor positivo.

3.4. Modelo Multinomial

O Modelo Multinomial é o modelo adequado para testar a hipótese na CATANOVA.

As respostas de cada tratamento, (n_{1j},..., n_{Ij}), seguem uma distribuição multinomial. Nesta, os dados em cada categoria se comportam de forma dicotômica, isto é, pertencem à categoria ou não.

Assim, em todos os tratamentos temos uma distribuição multinomial, que é, portanto, uma generalização da distribuição binomial:

$$P\{(n_{1j},\ldots,n_{Ij})\} = \binom{n_{+j}}{n_{1j},\ldots,n_{Ij}} \prod_{i=1}^{I} (p_{ij})^{n_{ij}} \qquad \textbf{3. 18}$$

sendo : $\sum_{i=1}^{I} p_{ij} = 1$, $p_{ij}>0$, i=1,..., I e j=1,..., J.

O modelo do produto de multinomiais para todos tratamentos será:

$$\Pr\{(n_{11},\ldots,n_{I1},\ldots,n_{1J},\ldots,n_{IJ})\} = \prod_{j=1}^{J} P\{(n_{1j},\ldots,n_{Ij})\} \qquad \textbf{3. 19}$$

$$= \prod_{j=1}^{J}\left[\binom{n_{+j}}{n_{1j},\ldots,n_{Ij}} \prod_{i=1}^{I} (p_{ij})^{n_{ij}}\right]$$

sendo $n_{+j}>0$, para todo j, p_{ij} a probabilidade de que uma unidade experimental no tratamento j corresponda a resposta i, e $\sum_{i=1}^{I} p_{ij} = 1$, para todo j.

3.5. Estatística C de Light e Margolin

A hipótese nula principal da Análise de Variância para Dados Categóricos é a de que todos os tratamentos tenham a mesma estrutura multinomial, podendo ser resumida por:

H_0: $\mathbf{p_{ij} = p_{i+}}$, para todo i e j.

Para testar esta hipótese, Light e Margolin (1971) desenvolveram a estatística **C**:

$$C = (n-1)(I-1)\left[\frac{\left(\sum_{j=1}^{J}\frac{1}{n_{+j}}\sum_{i=1}^{I} n_{ij}^2\right) - \frac{1}{n}\sum_{i=1}^{I} n_{i+}^2}{n - \frac{1}{n}\sum_{i=1}^{I} n_{i+}^2}\right] \qquad \textbf{3. 20}$$

A estatística **C** pode ser reescrito usando as componentes SQT e SQTotal:

$$C = \frac{(n-1)(I-1)SQT}{SQTotal} \qquad 3.21$$

Esta estatística se aproximada assintoticamente de uma distribuição Qui-Quadrado com (I -1)(J-1) graus de liberdade.

A relação entre SQT e SQTotal representa a proporção da variabilidade na variável resposta que pode ser atribuída a divisão em tratamentos, como na ANOVA:

$$R^2 = \frac{SQT}{SQTotal} \qquad 3.22$$

Considerando a equação 3.22, pode-se escrever a estatística **C** como função de R^2:

$$C = (n-1)(I-1)R^2 \qquad 3.23$$

Então, para testar a significância da medida de associação R^2, multiplica-se a estatística **C** por (n-1)(I-1) e aproxima-se esta distribuição sob H_0 pela distribuição Qui-Quadrado com (I -1)(J-1) graus de liberdade.

3.6. Tabela de Análise de Variância

As componentes de variação podem se resumir numa tabela tradicional de ANOVA, modificando-se apenas as fórmulas da Soma de Quadrados em cada componente. A estatística para a hipótese nula será a estatística **C** de Light e Margolin (1971). Esta tabela é apresentada na Tabela 3.2.

Tabela 3. 2. Tabela de Análise de Variância para Dados Categóricos com Um Fator.

Causas de Variação	GL	SQ	QM	C
Tratamentos	J-1	SQT	$QMT = \frac{\text{SQT}}{\text{J - 1}}$	$C = \frac{(n-1)(I-1)SQT}{SQTotal}$
Erro Experimental	n-J	SQE	$QME = \frac{\text{SQE}}{\text{n - J}}$	
Total	n-1	SQTotal	$QMTotal = \frac{\text{SQTotal}}{\text{n - 1}}$	

3.7. Comparação da estatística C com a estatística χ^2

3.7.1. Quando o número de categorias na variável resposta é dois

Light e Margolin (1971) demonstraram que a estatística χ^2 de Pearson pode ser expressa em função da estatística **C** se o número de categorias for igual a dois.

A estatística χ^2 tem como função:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\frac{\left(n_{ij} - \frac{n_{+j}n_{i+}}{n}\right)^2}{\frac{n_{+j}n_{i+}}{n}}\right] \qquad \textbf{3. 24}$$

Reescrevendo:

$$\chi^2 = \sum_i \frac{n}{n_{i+}} \sum_j \left(\frac{n_{ij}}{\sqrt{n_{+j}}} - \frac{n_{i+}}{n}\sqrt{n_{+j}}\right)^2 \qquad \textbf{3. 25}$$

Na CATANOVA, a Variação entre Grupos, SQT, pode ser reescrita:

$$SQT = \frac{1}{2}\sum_i \sum_j \left[\frac{n_{ij}}{\sqrt{n_{+j}}} - \frac{n_{i+}}{n}\sqrt{n_{+j}}\right]^2 \qquad \textbf{3. 26}$$

A partir da equação **3.21**, a estatística **C** pode ser apresentada como:

$$C = \frac{(I-1)(n-1)}{2SQTotal}\sum_i \sum_j \left[\frac{n_{ij}}{\sqrt{n_{+j}}} - \frac{n_{i+}}{n}\sqrt{n_{+j}}\right]^2 \qquad \textbf{3. 27}$$

Logo, com a introdução de uma ponderação, χ^2 e **C** podem ser representadas na forma:

$$\sum_i w_i \sum_j \left(\frac{n_{ij}}{\sqrt{n_{+j}}} - \frac{n_{i+}}{n}\sqrt{n_{+j}}\right)^2 \qquad \textbf{3. 28}$$

onde $w_i = \frac{n}{n_{i+}}$ para χ^2 e $w_i = \frac{(I-1)(n-1)}{2SQTotal}$ para **C**.

Como I=2, a estatística χ^2 pode ser expressa:

$$\chi^2 = \frac{n}{2}\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)SQT = \frac{1}{2}\left(\frac{n^2}{n_{1+}n_{2+}}\right)SQT \qquad \textbf{3. 29}$$

e a SQTotal pode ser simplificada para:

$$SQTotal = \frac{2n_{1+}n_{2+}}{n} \qquad \textbf{3. 30}$$

Portanto, a estatística χ^2 pode ser expressa como uma função da estatística **C**:

$$\chi^2 = \left(\frac{n}{n-1}\right)\frac{(n-1)SQT}{SQTotal} = \left(\frac{n}{n-1}\right)C \qquad \textbf{3. 31}$$

Observa-se, a partir desta equação, que para os casos de variável resposta ser dicotômica a estatística χ^2 será sempre maior que **C**, podendo ser praticamente igual quando o tamanho amostral for grande.

Como a estatística **C** é igual a $(n-1)(I-1)R^2$, a estatística χ^2 pode ser expressa também em função da medida de associação R^2 quando há 2 categorias na variável resposta:

$$\chi^2 = nR^2(2-1) = nR^2 \qquad \textbf{3. 32}$$

3.7.2. Quando o número de categorias na variável resposta é três

Light e Margolin (1971) realizaram simulações em amostras grandes para comparar a estatística **C** com a χ^2 de Pearson, quando o número de categorias na variável resposta é três e o número de tratamentos (níveis do fator) é dois.

Eles observaram que:

- Se uma probabilidade de resposta for alta e as outras duas forem baixas, o poder de C excede o poder de χ^2.
- Se uma probabilidade de resposta é baixa e as outras duas são altas o poder de χ^2 será maior que **C**.

• O ordenamento das categorias afeta a comparação entre a estatística C com χ^2.

A partir destas simulações observa-se que a estatística χ^2 somente pode ser expressa em função da estatística C ou da medida de associação R^2 somente se o número de categorias na variável resposta é dois.

3.8. Pequenas Amostras

Margolin e Light (1974) observaram que as medidas de associação R^2 e τ_b, uma versão amostral da medida de associação τ de Goodman e Kruskal (1954) são idênticas computacionalmente, porém diferentes no método e na interpretação. A medida τ_b é um coeficiente populacional assimétrico que mede a diminuição relativa na probabilidade de predição incorreta na variável resposta, quando se usa o conhecimento disponível sobre o fator. A fórmula desta medida é:

$$\tau_{ij} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\left(\frac{n_{+j}}{n}\right)p_{ij}^2 - \sum_{i=1}^{I}p_{i+}^2}{1-\sum_{i=1}^{I}p_{i+}^2} \qquad 3.33$$

onde: p_{ij} é a probabilidade de resposta i para um indivíduo, dado que não existe informações sobre o grupo a que pertence.

O estimador para τ_{ij} será:

$$t_{ij} = \frac{\left(\sum_{j=1}^{J}\sum_{i=1}^{I}\frac{n_{ij}^2}{nn_{+j}}\right) - \sum_{i=1}^{I}\frac{n_{i+}^2}{n^2}}{1-\sum_{i=1}^{I}\frac{n_{i+}^2}{n^2}} = \frac{SQT}{SQTotal} = R^2 \qquad 3.34$$

Logo, t_{ij} terá a mesma fórmula e as mesmas propriedades da medida de associação R^2 de Light e Margolin (1971).

Margolin e Light (1974) observaram que as estatísticas C, χ^2 calculado e 2Î, a estatística de informação de Kullback, possuem distribuição aproximada a

uma distribuição Qui-Quadrado com graus de liberdade (I-1)(J-1). A fórmula da estatística de informação de Kullback (Kullback et al., 1962) é igual ao logaritmo da razão de verossimilhança para tabelas de contingência:

$$2\hat{I} = 2\left[n\ln n + \sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij}\ln n_{ij} - \sum_{i=1}^{I} n_{i+}\ln n_{i+} - \sum_{j=1}^{J} n_{+j}\ln n_{+j} \right] \qquad 3.35$$

A partir de simulações, estes autores observaram também que, em pequenas amostras, a distribuição da estatística **C** se aproxima melhor de uma $\chi^2_{(I-1)(J-1)}$ do que o faz a distribuição do χ^2 calculado, sendo que a distribuição nula de 2Î é a que menos se aproxima dessa distribuição teórica. Além disso, o valor da estatística 2Î será sempre maior ou igual do que o do χ^2 calculado, quando se comparam dois grupos quanto a uma variável resposta com qualquer número de categorias, logo este é mais conservador do que 2Î. Assim, os autores aconselham não usar 2Î para testar independência em tabelas com amostras pequenas, preferindo a estatística **C**.

3.9. CATANOVA e MANOVA

Light e Margolin (1971) observaram que a CATANOVA se assemelha a Análise de Variância Multivariada (MANOVA), que é usada quando há duas ou mais variáveis respostas (dependentes) intervalares e variáveis independentes categóricas. Essa semelhança ocorre porque a variável resposta da CATANOVA pode ser dividida em múltiplas variáveis, considerando cada categoria uma variável dicotômica.

Segundo Light e Margolin (1971), a estatística de teste para a MANOVA é uma função monotonicamente crescente da estatística de teste para CATANOVA.

No caso em que o número de categorias é dois, a MANOVA torna-se equivalente a uma ANOVA com uma variável resposta do tipo 0,1. Neste caso, Cochran (1950) sugere o seguinte teste para a ANOVA:

$$\frac{SQT/(J-1)}{SQE/(n-J)} \qquad \textbf{3. 36}$$

Sob H_0, esta razão é comparada com uma $F_{J-1,\,n-J}$.

Já no caso do teste de CATANOVA, a estatística seria:

$$C = \frac{(n-1)SQT}{SQTotal} \qquad \textbf{3. 37}$$

Sob H_0, este teste é realizado usando-se uma χ^2_{J-1}.

Logo, estas estatísticas não são equivalentes quando o número de categorias é igual a dois. Embora as estatísticas de teste se relacionem de forma monótona, a distribuição de referência não obedece à mesma relação. Isto ocorre porque a distribuição de referência na CATANOVA trata a Soma Total dos Quadrados (SQTotal) como uma constante.

4. CATANOVA PARA MAIS DE UM FATOR: ASPECTOS TEÓRICOS

Partindo dos resultados de Light e Margolin (1971) sobre CATANOVA para tabelas bidimensionais contendo um fator, Anderson e Landis (1980, 1982) expandiram a CATANOVA para tabelas de contingência multidimensionais contendo dois ou mais fatores.

Como a estatística **C** pode ser escrita como uma função da medida de associação R^2 para dados categóricos, Anderson e Landis (1980) fizeram uma analogia à regressão linear múltipla para dados contínuos e desenvolveram medidas de associação parcial e múltipla para o caso da variável resposta ser em escala nominal.

Anderson e Landis (1982) também estenderam a metodologia da CATANOVA para os casos da variável resposta em escala ordinal. Esta metodologia não será abordada nesta monografia.

O objetivo destes autores em criar a CATANOVA para vários fatores, tanto para variável resposta nominal como ordinal, não é somente fazer uma expansão para mais de um fator e sim obter mais informações do que simplesmente a proporção da variação total explicada pela variação entre os tratamentos.

Anderson e Landis (1980) desenvolveram a Análise de Variância para Dados Categóricos em escala nominal para o caso de dois fatores. Eles recomendam utilizar o **Fator 1** como a variável de maior interesse e o **Fator 2** como a variável de controle.

Também comentam que esta análise pode ser estendida para mais do que dois fatores. Nestes casos, eles recomendam reunir todas as variáveis de controle cruzando as classificações de tal modo a constituírem as categorias de um único fator de controle, o Fator 2. Por exemplo, a variável de controle 1 possui os níveis A e B, e a variável de controle 2, os níveis 1, 2 e 3. Assim, o Fator 2 tem seis níveis, resultantes das combinações A1, A2, A3, B1, B2 e B3.

Já o Fator 1 continuaria sendo a variável de maior interesse. E as medidas de associação parciais múltiplas podem ser vistas como medidas de associação parcial.

Em uma tabela de contingência tridimensional, há dois fatores independentes que são o Fator 1, j=1,...,J, e o Fator 2, k=1,...,K, e uma variável dependente que é a resposta com as categorias em escala nominal, i=1,...,I.

Essas medidas de associações foram criadas para medir a variação explicada da variável dependente:

a) pelo Fator 1 ou pelo Fator 2 como "fatores principais";

b) pelo efeito da interação entre o Fator 1 e o Fator 2;

c) pelo efeito parcial do Fator 2 dado o Fator 1 ou o efeito parcial do Fator 1 dado o Fator 2. (Ou, em outras palavras, a variação explicada pelo Fator 1 controlando pelo Fator 2).

O número de indivíduos em cada categoria da variável resposta, conforme Fator 1 e Fator 2, são apresentadas na forma de tabela tridimensional na Tabela 4.1.

Tabela 4. 1. Tabela de Contingência Tridimensional utilizada em CATANOVA.

		Categorias da Variável Resposta				
Fator 1	**Fator 2**	1	2		I	Total
1	1	n_{111}	n_{211}	⋯	n_{I11}	n_{+11}
1	2	n_{112}	n_{212}	⋯	n_{I12}	n_{+12}
⋮	⋮	⋮	⋮	⋯	⋮	⋮
1	K	n_{11K}	n_{21K}	⋯	n_{I1K}	n_{+1K}
Total		n_{11+}	n_{21+}		n_{I1+}	n_{+1+}
2	1	n_{121}	n_{221}	⋯	n_{I21}	n_{+21}
2	2	n_{122}	n_{222}	⋯	n_{I22}	n_{+22}
⋮	⋮	⋮	⋮	⋯	⋮	⋮
2	K	n_{12K}	n_{22K}	⋯	n_{I2K}	n_{+2K}
Total		n_{12+}	n_{22+}	⋯	n_{I2+}	n_{+2+}
⋯	⋯	⋯	⋯	⋯	⋯	⋯
J	1	n_{1J1}	n_{2J1}	⋯	n_{IJ1}	n_{+J1}
J	2	n_{1J2}	n_{2J2}	⋯	n_{IJ2}	n_{+J2}
⋮	⋮	⋮	⋮	⋯	⋮	⋮
J	K	n_{1JK}	n_{2JK}	⋯	n_{IJK}	n_{+JK}
Total		n_{1J+}	n_{2J+}	⋯	n_{IJ+}	n_{+J+}
TOTAL		n_{1++}	n_{2++}	⋯	n_{I++}	$n_{+++}=n$

4.1. Componentes de variação

Anderson e Landis (1980) expandiram a proposta de Light e Margolin (1971), partindo de conceitos da regressão linear múltipla e propuseram formas para calcular as principais medidas de associação. Na tabela 4.2 estão os componentes de variação necessários para obter estas medidas.

Tabela 4. 2. Medidas de Variação (Soma de Quadrados, SQ) necessárias para o cálculo das medidas de associação em tabelas de contingência.

Componentes	Fórmulas
SQ Total	$SQTotal = \frac{n}{2} - \frac{\sum_{i=1}^{I} n_{i++}^2}{2n}$
SQ do Erro Residual dentro do Fator 1(j) e o Fator 2(k)	$SQE(j,k) = \frac{n}{2} - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}^2}{2n_{+jk}}$
SQ do Erro Residual dentro de Fator 2(k)	$SQE(k) = \frac{n}{2} - \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} n_{i+k}^2}{2n_{++k}}$
SQ do Erro Residual dentro do 1° nível de Fator 2(k)	$SQE(k=1) = \frac{n_{++1}}{2} - \frac{\sum_{i=1}^{I} n_{i+1}^2}{2n_{++1}}$
SQ do Erro Residual dentro de Fator 1(j)	$SQE(j) = \frac{n}{2} - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij+}^2}{2n_{+j+}}$
SQ da Regressão da Variável resposta no Fator 1(j) e no Fator 2(k)	$SQR(j,k) = SQTotal - SQE(j,k) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}^2}{2n_{+jk}} - \frac{\sum_{i=1}^{I} n_{i++}^2}{2n}$
SQ da Regressão de i em j e k, dado j=1 e k=1	$SQR(j=1,k=1) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{i11}^2}{2n_{+11}} - \frac{\sum_{i=1}^{I} n_{i11}^2}{2n_{+11}}$
SQ da Regressão de i em j, ajustando por k	$SQR(j \mid k) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}^2}{2n_{+jk}} - \frac{\sum_{i=1}^{I}\sum_{k=1}^{K} n_{i+k}^2}{2n_{++k}}$
SQ da Regressão de i em j	$SQR(j) = SQTotal - SQE(j)$
SQ da Regressão de i em k	$SQR(k) = SQTotal - SQE(k)$

As componentes SQTotal, SQR(j,k) e SQE(j,k) são expansões para dois fatores dos componentes SQTotal, SQT e SQE respectivamente, criados por Light e Margolin (1971) para o caso de um fator e uma variável resposta.

4.2. Medidas de associação para CATANOVA

As medidas de associação para dados categóricos são calculadas da mesma forma usada para dados contínuos: SQT/SQTotal.

Como no caso de medida de associação para um fator, a medida de associação para dois ou mais fatores varia de 0 a 1.

As principais medidas de associação obtidas pela CATANOVA são as seguintes:

- **Medidas de Associação Simples**

As medidas de associação simples como R_{ij}^2 e R_{ik}^2 são semelhantes as criadas por Light e Margolin (1971).

R_{ij}^2 indica o efeito individual do Fator 1 que explicaria a variação na variável resposta sem considerar o Fator 2.

$$R_{ij}^2 = \frac{SQR(j)}{SQTotal} \qquad 4.1$$

R_{ik}^2 indica a proporção da variação na variável resposta explicada apenas pelo Fator 2.

$$R_{ik}^2 = \frac{SQR(k)}{SQTotal} \qquad 4.2$$

- **Medida de Associação Múltipla**

A medida de associação múltipla, que é a proporção da variação da variável resposta explicada pela combinação do Fator 1 e do Fator 2, mostra a reunião dos dois efeitos:

$$R^2_{ijk} = \frac{SQR(j,k)}{SQTotal} = \frac{\dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n^2_{ijk}}{n_{+jk}} - \dfrac{\sum_{i=1}^{I} n^2_{i++}}{n}}{n - \dfrac{\sum_{i=1}^{I} n^2_{i++}}{n}} \qquad 4.3$$

- **Medidas Parciais de Associação**

As medidas parciais de associação, na maioria das vezes, são o principal interesse na Análise de Variância para Dados Categóricos, pois medem a associação de um dos fatores, em geral o mais importante, controlando pelos níveis do segundo fator. Há dois tipos de medida de associação parcial em tabelas tridimensionais:

1) Uma medida de associação parcial do Fator 1 controlando pelo segundo fator; reunindo a informação de todas as categorias do Fator 2:

$$R^2_{ij|k} = SQR(j|k) / SQE(k) \qquad 4.4$$

2) Uma medida de associação parcial do Fator 1 em cada nível do segundo fator:

$$R^2_{ij|(k=1)} = SQR(j|k=1) / SQE(k=1) \qquad 4.5$$

4.3. Modelo Multinomial para tabelas tridimensionais

Segundo Andrade e Pinheiro (2002), em uma tabela tridimensional as respostas de cada grupo seguem o mesmo princípio de distribuição multinomial dado em uma tabela bidimensional.

$$P\{(n_{1jk}, n_{2jk}, ..., n_{Ijk})\} = \binom{N}{n_{1jk}, ..., n_{Ijk}} \prod_{i=1}^{I} (p_{ij})^{n_{ijk}} \qquad 4.6$$

sendo : $\sum_{i=1}^{I} p_{ijk} = 1$ e $p_{ij}>0$, j=1,...,J e k=1,...,K.

O modelo do produto de multinomiais considerando todos os tratamentos, supondo que as categorias do Fator 2 são independentes, será:

$$\prod_{j=1}^{G}\prod_{k=1}^{K} \Pr\{n_{1jk}, n_{2jk}, \ldots, n_{ijk}\} = \prod_{j=1}^{J}\prod_{k=1}^{K}\binom{N}{n_{1jk}, \ldots, n_{Ijk}}\prod_{i=1}^{I}(p_{ijk})^{n_{ijk}} \qquad 4.7$$

$$= \prod_{j=1}^{J}\left[\binom{n_{+j}}{n_{1j}, \ldots, n_{Ij}}\prod_{i=1}^{I}(p_{ij})^{n_{ij}}\right]$$

sendo que: $n_{+j+}>0$, para todo j ; p_{ijk} é a probabilidade de que uma unidade experimental no fator 1 (j) corresponda a uma resposta i ; e $\sum_{i=1}^{I} p_{ij}=1$, para todo j.

4.4. Os testes de associação e a estatística C

Para o teste das diversas medidas de associação, a hipótese nula da Análise de Variância para Dados Categóricos (CATANOVA) para dois fatores deve ser subdividida.

Por exemplo, para a medida de associação parcial de maior interesse, que mede o efeito do Fator 1 controlando pelo Fator 2, a hipótese nula é a de que todos os níveis do Fator 1 tenham a mesma estrutura multinomial, categoria a categoria do Fator 2. Ou seja, o vetor de proporções na categoria 1 do Fator 2 é o mesmo em todos os níveis do Fator 1. Resumidamente,

H_0: $p_{ijk} = p_{i+k}$ para todo j=1,2,...,J

Esta hipótese é testada pela estatística C de Anderson e Landis (1980), que é calculada como uma função da medida de associação R^2, criada por Light e Margolin (1971). Para testar a independência em tabelas tridimensionais, pela estatística C, utiliza-se a distribuição Qui-Quadrado. As fórmulas desta estatística e os graus de liberdade para a distribuição nula de

Qui-Quadrado que testa cada medida de associação citadas acima (4.1-4.5) são mostrados na Tabela 4.3.

Tabela 4. 3. Cálculo para a estatística C a partir das medidas de associação.

Medida de Associação		C	gl	Efeito
Simples	R^2_{ij}	$C=(n-1)(I-1)R^2_{ij}$	(I-1)(J-1)	Efeito Principal (Fator 1)
	R^2_{ik}	$C=(n-1)(I-1)R^2_{ik}$	(I-1)(K-1)	Efeito Principal (Fator 2)
Múltipla	R^2_{ijk}	$C=(n-1)(I-1)R^2_{ijk}$	(I-1)(JK-1)	Efeito dos 2 fatores
Parcial	$R^2_{ij\|(k=1)}$	$C=(n_{+1+}-1)(I-1)R^2_{ij\|(k=1)}$	(I-1)(J-1)	Efeito do Fator 1 no nível k=1 do Fator 2
	$R^2_{ij\|k}$	$C=(n-1)(I-1)R^2_{ij\|k}$	(I-1)(J-1)K	Efeito do Fator 1 controlando pelo Fator 2

5. CATANOVA APLICADA À GENÉTICA DE POPULAÇÕES

5.1. Um exemplo aplicado à genética de moscas-das-frutas

Regner et al. (1999) utilizaram a espécie de mosca-das-frutas *Drosophila willistoni*, considerada um paradigma para estudos evolutivos, com a finalidade de identificar os agentes que causam a redução da viabilidade de linhagens híbridas produzidas por cruzamento de populações naturais e de laboratório.

Duas linhagens de *Drosophila willistoni* foram usadas:

- **WIP-11A**: coletada em Eldorado do Sul, Rio Grande do Sul, desde 1991 está sendo criada em laboratório.
- **17A2**: originária da caatinga, próximo a Salvador - Bahia, e está a 30 anos em laboratório.

As linhagens foram cruzadas com indivíduos da mesma linhagem e com indivíduos da outra linhagem, sendo que cada tipo de cruzamento foi feito por 30 fêmeas virgens com 30 machos virgens. Os quatro tipos de cruzamentos possíveis foram realizados em três temperaturas diferentes (18°C, 25°C e 29°C), para observar se os descendentes (F1), divididos em fêmeas e machos, tinham gônadas normais (GN) ou disgênicas (gônadas mal formadas; GD).

A gônada é uma glândula sexual que produz os gametas e segrega hormônios. A gônada feminina é o ovário, e a masculina é o testículo. A disgenesia gonadal reduz a viabilidade e a fecundidade, levando a uma redução na adaptação biológica da espécie.

O objetivo principal do trabalho é saber se existe efeito da temperatura na proporção de gônadas normais e disgênicas na descendência (F1) dos quatro cruzamentos estudados.

Os resultados deste estudo estão apresentados na Tabela 5.1.

Tabela 5. 1. Quantidade de gônadas digênicas e normais na descendência, fêmeas e machos de *Drosophila willistoni*, em diferentes temperaturas.

			F1 Fêmea		F1 Macho	
	Cruzamentos	**Temperatura (°C)**	GD	GN	GD	GN
1	WIP-11A fêmeas	18	4	180	3	171
	X	25	2	172	2	158
	WIP-11A machos	29	6	172	3	123
2	WIP-11A fêmeas	18	2	214	6	210
	X	25	0	180	4	164
	17A2 machos	29	47	133	21	81
3	17A2 fêmeas	18	0	190	4	192
	X	25	2	166	0	156
	WIP-11A machos	29	32	138	14	86
4	17A2 fêmeas	18	4	176	4	158
	X	25	10	172	3	169
	17A2 machos	29	28	144	16	106

Analisou-se separadamente a F1 de fêmeas e machos utilizando a Análise de Variância para Dados Categóricos com tabelas tridimensionais. Os bancos de dados para análise estão em Anexo 1. O **Fator 1 (j)**, o de maior interesse, foi a temperatura, e o **Fator 2 (k)**, o de controle, foi o tipo de cruzamento em *Drosophila willistoni*. A **variável resposta (i)** foi o tipo de gônadas, disgênicas (GD) ou normais (GN). A estatística C foi comparada com a estatística Qui-Quadrado de Pearson nas diversas medidas de associação.

Os cálculos para a realização do teste para a CATANOVA e o Qui-Quadrado foram feitos no Excel®.

A soma de quadrados, separados por descendentes, necessários para calcular as medidas de associação estão apresentados em Anexo 2.

A Tabela 5.2 apresenta os resultados para os descendentes serem fêmeas, e a Tabela 5.3, para os descendentes serem machos.

Tabela 5. 2. Medidas de Associação para o caso dos descendentes serem fêmeas.

Medidas de Associação	R^2	Estatística Baseada no R^2 C	gl	p	Estatística Qui-Quadrado de Pearson χ^2	gl	p	FME
R^2_{ij}	0,0846	183,907	2	0,000	169,635	2	0,000	44,11
R^2_{ik}	0,0101	21,946	3	0,000	21,956	3	0,000	33,27
R^2_{ijk}	0,1185	257,533	11	0,000	256,571	11	0,000	10,71
$R^2_{ij/k=1}$	0,0037	1,985	2	**0,371**	1,989	2	**0,370**	3,90 [a]
$R^2_{ij/k=2}$	0,1812	104,175	2	0,000	104,356	2	0,000	15,31
$R^2_{ij/k=3}$	0,1213	63,914	2	0,000	64,035	2	0,000	10,82
$R^2_{ij/k=4}$	0,0489	26,075	2	0,000	26,124	2	0,000	13,53
$R^2_{ij/k}$	0,1095	237,990	8	0,000	196,505	8	0,000	3,90 [a]

[a] 3 células (50%) possuem valores esperados menores que 5.
FME: freqüência mínima esperada.

Tabela 5. 3. Medidas de Associação para o caso dos descendentes serem machos.

Medidas de Associação	R^2	Estatística Baseada no R^2 C	gl	p	Estatística Qui-Quadrado de Pearson χ^2	gl	p	FME
R^2_{ij}	0,0485	89,791	2	0,000	85,687	2	0,000	19,42
R^2_{ik}	0,0071	13,106	3	0,004	13,113	3	0,004	19,50
R^2_{ijk}	0,0728	134,898	11	0,000	139,832	11	0,000	4,734
$R^2_{ij/k=1}$	0,0011	0,527	2	**0,768**	0,528	2	**0,768**	2,19 [a]
$R^2_{ij/k=2}$	0,0899	43,584	2	0,000	43,674	2	0,000	6,51
$R^2_{ij/k=3}$	0,0767	34,571	2	0,000	34,647	2	0,000	3,98 [b]
$R^2_{ij/k=4}$	0,0499	22,695	2	0,000	22,745	2	0,000	6,15
$R^2_{ij/k}$	0,0662	122,660	8	0,000	101,594	8	0,000	2,19 [a]

[a] 3 células (50%) possuem valores esperados menores que 5.
[b] 1 células (16,7%) possuem valores esperados menores que 5.
FME: freqüência mínima esperada.

Observando as medidas de associação das tabelas 5.2. e 5.3, observa-se que existe diferença entre fêmeas e machos. Sem considerar o tipo de cruzamento, a temperatura explica 8,46% (R^2_{ij}) da variação no tipo de gônadas nas fêmeas, e 4,85%, nos machos. Já controlando pelo tipo de cruzamento, a temperatura explica 10,95% ($R^2_{ij/k}$) da variação no tipo de gônadas nas fêmeas

e 6,62% nos machos, um aumento proporcional de 2,49% para fêmeas e 1,77% para machos. R^2_{ijk} representa o efeito da combinação entre temperatura e tipo de cruzamento: este efeito explica aproximadamente 11,85% da variação em fêmeas e 7,67% em machos. Logo, há indícios de que, na descendência, as fêmeas tiveram maior influência da temperatura sobre a disgenesia gonadal do que os machos.

A partir da Estatística **C** baseada na medida de associação, observa-se que há diferença significativa (α=0,05) em todos os casos, com apenas uma exceção. Logo, a temperatura influi significativamente (α=0,05) no tipo de gônadas, quer analisando separadamente, quer controlando pelo tipo de cruzamento, ou na combinação de temperatura e tipo de cruzamento. Tanto no caso de descendentes machos quanto fêmeas, não houve diferença significativa (α=0,05) apenas quando o efeito da temperatura foi estudada no cruzamento 1 (machos e fêmeas da linhagem WIP-11A). Isto provavelmente se deve ao fato desta linhagem ser originária do Rio Grande do Sul, um clima que contém estações do ano bem definidas, com altas e baixas temperaturas.

Comparando a Estatística **C** com a χ^2 calculado, chega-se as mesmas conclusões. Também é importante ressaltar que a estatística **C** tem maior confiabilidade nos resultados deste exemplo do que a estatística χ^2 calculado, pois neste foram ignoradas suposições como número grande de freqüências esperadas menores que 5 (Tabela 5.2 e 5.3).

Quanto à comparação destas duas estatísticas, Light e Margolin (1971) demonstraram que, para o caso de tabelas bidimensionais com a variável resposta dicotômica, a estatística χ^2 pode ser escrita em função de **C** ou em função da medida de associação R^2:

$$\chi^2 = \left(\frac{n}{n-1}\right)C = nR^2 \qquad \textbf{5.1}$$

Observa-se que no exemplo mostrado, onde as tabelas são tridimensionais com a variável resposta dicotômica, esta função é válida para todas as medidas de associação parcial da temperatura em cada cruzamento

($R^2_{ij/k=1}$, $R^2_{ij/k=2}$, $R^2_{ij/k=3}$, $R^2_{ij/k=4}$) e para a medida de associação entre tipo de cruzamento e disgenesia gonadal sem considerar temperatura (R^2_{ik}), não sendo válida para as demais medidas de associação.

Isto provavelmente ocorre porque as medidas válidas são calculadas a partir de tabelas bidimensionais. Por exemplo, a medida de associação parcial $R^2_{ij/k=1}$ terá como variáveis apenas a temperatura e o tipo de gônadas, restrito ao tipo de cruzamento 1.

Dessa forma, deve-se ter cuidado redobrado em utilizar a função 5.1, que terá que ser reformulada para o caso de dois ou mais fatores. É o que se observa no exemplo de Anderson e Landis (1980) no qual esta correspondência não foi observada (dados não mostrados) em nenhuma medida de associação. Este exemplo envolve dois fatores e uma variável resposta com mais do que duas categorias.

Outro modo de analisar os dados de Regner et al. (1999) que pode ter interesse é controlar ao mesmo tempo o tipo de cruzamento e sexo da descendência, criando-se 8 níveis na variável de controle. Assim, a análise continua tendo dois fatores: o Fator 1 é a variável temperatura; e o Fator 2 é o cruzamentos das categorias das variáveis sexo da descendência e tipo de cruzamento. O banco de dados para esta análise está apresentado na Tabela 5.4.

A soma de quadrados para calcular as medidas de associação está apresentada em Anexo 3.

Nesta análise as medidas parciais em cada cruzamento e em cada sexo são as mesmas já apresentadas na Tabela 5.2 e 5.3, assim serão apresentadas na Tabela 5.5. apenas as demais medidas de associação.

Tabela 5. 4. Banco de dados para análise quando o Fator 2 é a combinação de cruzamento e sexo dos descendentes.

Temperatura (Fator 1)	Fator 2				Tipo de Gônadas	
	Nível	Mãe	Pai	Descendentes	GD	NGD
18	1	WIP-11A	WIP-11A	Fêmea	4	180
	2	WIP-11A	WIP-11A	Macho	3	171
	3	WIP-11A	17A2	Fêmea	2	214
	4	WIP-11A	17A2	Macho	6	210
	5	17A2	WIP-11A	Fêmea	0	190
	6	17A2	WIP-11A	Macho	4	192
	7	17A2	17A2	Fêmea	4	176
	8	17A2	17A2	Macho	4	158
25	1	WIP-11A	WIP-11A	Fêmea	2	172
	2	WIP-11A	WIP-11A	Macho	2	158
	3	WIP-11A	17A2	Fêmea	0	180
	4	WIP-11A	17A2	Macho	4	164
	5	17A2	WIP-11A	Fêmea	2	166
	6	17A2	WIP-11A	Macho	0	156
	7	17A2	17A2	Fêmea	10	172
	8	17A2	17A2	Macho	3	169
29	1	WIP-11A	WIP-11A	Fêmea	6	172
	2	WIP-11A	WIP-11A	Macho	3	123
	3	WIP-11A	17A2	Fêmea	47	133
	4	WIP-11A	17A2	Macho	21	81
	5	17A2	WIP-11A	Fêmea	32	138
	6	17A2	WIP-11A	Macho	14	86
	7	17A2	17A2	Fêmea	28	144
	8	17A2	17A2	Macho	16	106

Tabela 5. 5. Medidas de Associação para o caso do Fator 2 ser a combinação de tipo de cruzamento e sexo dos descendentes.

Efeito	Medidas de Associação	Estatística Baseada no R^2				Estatística Qui-Quadrado de Pearson			
		R^2	C	gl	p	χ^2	gl	p	FME
Temperatura	R^2_{ij}	0,0654	263,422	2	0,000	263,487	2	0,000	61,95
Sexo e Cruzamento	R^2_{ik}	0,0109	43,795	7	0,000	43,806	7	0,000	24,35
Temperatura x Sexo e Cruzamento	R^2_{ijk}	0,1032	415,450	23	0,000	472,321	23	0,000	7,01
Temperatura/ Sexo e Cruzamento	$R^2_{ij/k}$	0,0933	375,742	16	0,000	298,243	16	0,000	2,19[a]

[a] 7 células (14,6%) possuem valores esperados menores que 5.
FME: freqüência mínima esperada.

Observa-se que as medidas de associação mostradas na Tabela 5.5 são médias ponderadas das medidas apresentadas na Tabela 5.2 e 5.3, com exceção da medida de associação da combinação sexo da descendência e tipo de cruzamento, que explica apenas 1,09% da variação no tipo de gônadas. Já a temperatura sem considerar o tipo de cruzamento e o sexo na descendência explica 6,54% da variação no tipo de gônadas (R^2_{ij}). Controlando pelo tipo de cruzamento e o sexo na descendência ($R^2_{ij/k}$) a explicação devida a esta variável é de 9,33%. Por último, a combinação de temperatura e do tipo de cruzamento com o sexo na descendência (R^2_{ijk}) explica 10,32% da variação no tipo de gônadas.

Como nos casos anteriores, existe significância estatística (α=0,05) tanto na estatística **C**, baseada na medida de associação R^2, como na estatística χ^2. No caso do Fator 2 ser a combinação do tipo de cruzamento com o sexo da descendência, a equação 5.1 é válida nos mesmos casos, quais sejam as medidas de associação parcial da temperatura em cada combinação de cruzamento com o sexo na descendência e a medida de associação da combinação do tipo de cruzamento com sexo na descendência sem considerar temperatura.

5.2. Um exemplo aplicado a seqüências genômicas

5.2.1.Genética de Populações e DNA

A Genética é o ramo da Biologia que trata da hereditariedade de características nos organismos e de sua evolução. Após a descoberta de que os genes são compostos de ácido desorribonucléico (**DNA**), na década de 1940, e da descrição da estrutura molecular de DNA por J. Watson e F. Crick em 1953, os avanços científicos nesta área têm sido vertiginosos.

Com o advento das técnicas de manipulação e análises moleculares desenvolvidas a partir da década de 1970 e principalmente na década de 1980, quando ocorreu a descoberta da Reação em Cadeia de Polimerase (**PCR**), a pesquisa científica nesta área cresce em volume e velocidade inimagináveis, gerando enormes quantidades de informação. Dessa forma, torna-se cada vez mais necessário o desenvolvimento de técnicas de análise de dados adaptadas a esta área.

A Análise de Variância para Dados Categóricos (CATANOVA) representa uma ferramenta que pode ser muito útil na análise de dados oriundos de pesquisa em Genética de Populações. A seguir, serão apresentados alguns conceitos básicos usados nesta área.

Gene

Gene é a unidade básica da transmissão da informação genética. Trata-se de pequenos trechos de DNA (ácido desoxirribonucléico) que guardam, em sua seqüência, as informações para a construção de proteínas.

Nucleotídeo

O DNA é compreendido por uma seqüência de nucleotídeos, que são as unidades básicas das seqüências genômicas. Eles são formados por uma base nitrogenada, uma pentose (desoxirribose, no caso do DNA) e um grupo fosfato (PO_4).

As bases nitrogenadas presentes no DNA são: Adenina (**A**), Timina (**T**), Citosina (**C**) e Guanina (**G**). Considerando sua estrutura química, as bases podem ser agrupadas em duas classes: **Purinas** (A e G) e as **Pirimidinas** (C e T).

A molécula de DNA é composta por duas cadeias de nucleotídeos que se dispõem em uma espiral em torno de um eixo imaginário. As bases nitrogenadas de cada cadeia ficam face a face e pontes de hidrogênio entre elas dão estabilidade à molécula. Devido a esta estrutura, o tamanho de um segmento qualquer do DNA é muitas vezes expresso em "número de pares de bases".

A estrutura do DNA destacando a forma do nucleotídeo é mostrada na Figura 5.1.

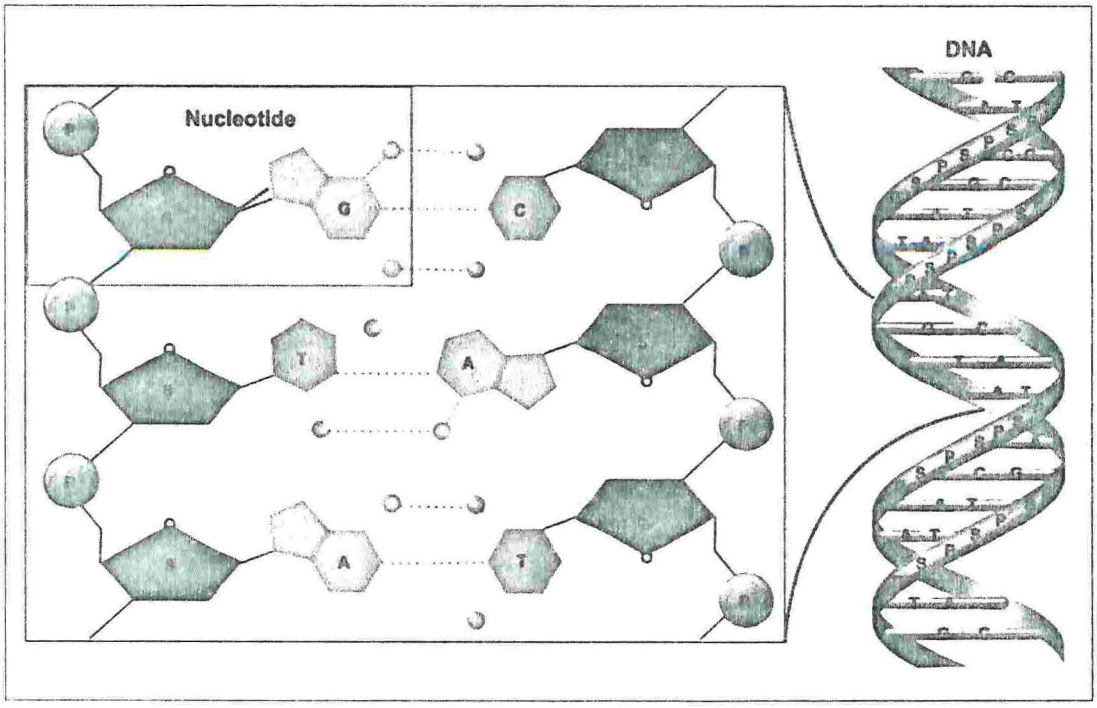

Figura 5. 1. Esquema de um segmento desenrolado da dupla hélice de DNA, mostrando a relação entre as bases complementares que representam o conteúdo informativo de DNA e o esqueleto de açúcar-fosfato (S-P) que é idêntica em todo o DNA.

As diferenças de ordenamento dos nucleotídeos nas seqüências genômicas são características de cada espécie.

Mutação

Mutação é a designação dada a uma alteração na seqüência de DNA. Ela pode ser pontual (em um nucleotídeo) ou envolver trechos maiores. Nas mutações pontuais pode ocorrer a retirada, a adição ou a troca de uma base. Quando uma mutação envolve a troca de uma base da classe das purinas para outra purina ou de uma pirimidina por outra pirimidina, é chamada de **transição**. Quando a troca envolve uma purina para uma pirimidina ou vice-versa, chama-se **transversão**.

Aminoácido

Os aminoácidos são unidades básicas que compõem as proteínas. Existem 20 aminoácidos na natureza. De acordo com as regras do Código Genético, cada aminoácido é determinado por uma seqüência de três nucleotídeos. Estas trincas são chamadas de *códons*. Com três bases diferentes é possível formar 64 arranjos, que são em número maior do que o necessário para codificar os 20 aminoácidos. Assim, verificou-se que códons diferentes podem indicar o mesmo aminoácido. Por exemplo, as trincas GAT e GAC codificam o ácido aspártico.

Um gene, que é uma seqüência de nucleotídeos, codifica a construção de uma proteína, que é uma seqüência de aminoácidos.

DNA Mitocondrial (DNAmit)

DNA mitocondrial é uma molécula de DNA circular, existente em uma organela do citoplasma das células chamada mitocôndria.

Estas mitocôndrias são produtoras de energia. O DNAmit é pequeno (aproximadamente 16000 pares de bases) quanto comparado à forma nuclear (3 bilhões de pares de base). O DNAmit é de herança exclusivamente materna, sendo herdado através das mitocôndrias presentes no citoplasma do óvulo, o qual juntamente com o espermatozóide, gerará o indivíduo.

Região Codificante e Não-Codificante

A seqüência completa do DNA se subdivide em duas partes: a codificante e a não codificante. Em humanos, as partes codificantes representam 3% a 5% do DNA e são elas que caracterizam o indivíduo. Na parte não-codificante, ocorre maior variabilidade, já que nesta região se fixam mais mutação que não são eliminadas pela seleção natural, como o seriam se estivessem na porção codificante e fossem deletérios ao organismo. Além disso, esta região é proporcionalmente maior do que a parte codificante do DNA.

Genética de Populações

A constituição genética de um indivíduo é denominada genótipo e nos casos de organismos diplóides (2n cromossomos), cada indivíduo possui dois conjuntos gênicos, um herdado do pai e outro da mãe.

A Genética de Populações estuda as diferenças, quanto a características herdáveis, entre indivíduos de uma mesma população e entre populações. Para se verificar corretamente se existem diferenças tanto entre indivíduos como entre populações, utilizam-se métodos de análise estatística.

5.2.2. CATANOVA aplicada a seqüências genômicas

Uma das áreas que podem utilizar em larga escala a CATANOVA é a área de genética dirigida para seqüências genômicas. Uma seqüência genômica é um conjunto de informações, em nível de DNA, obtida para um indivíduo.

Isto ocorre porque a CATANOVA é usada para dados não ordenados, assim podendo ser utilizado perfeitamente para comparações de conjuntos de seqüências como os nucleotídeos e os aminoácidos.

O objetivo da análise para seqüências genômicas (cada seqüência representa um indivíduo) é comparar a variabilidade observada em várias posições (sítios do DNA) entre e dentro de populações (grupos). Isto é citado em Andrade e Pinheiro (2002): "Seqüências genômicas não são consideradas em suas características individuais, mas como contribuidoras na variabilidade total de uma distribuição categórica com alta dimensão."

A hipótese principal é verificar se existe diferença entre as populações. No caso de seqüências genômicas, pelos critérios de Anderson e Landis (1980), o Fator 1 seria "grupo" que é o fator mais importante, e o Fator 2 seriam as posições (ou sítios) funcionando como um fator de controle.

Na maioria dos casos de seqüências genômicas não há interesse em trabalhar com tabelas bidimensionais com apenas uma posição que oferecem pouca informação.

Apesar de as posições não serem completamente independentes, uma suposição necessária é a de que exista independência entre as posições ao longo do DNA. Um dos motivos de elas não serem independentes é que, nas mutações, uma posição é compensada por substituições em posições vizinhas. Outro motivo é os antepassados possivelmente serem comuns. Andrade e Pinheiro (2002) comentam que uma maneira de obter resultados parecidos ao da situação de verdadeira independência é utilizar um grande número de posições.

Outra suposição é feita na parte da modelagem da distribuição multinomial. As freqüências assumem uma distribuição hipergeométrica nos casos de seqüências genômicas, por se utilizar uma amostragem sem reposição. Dessa forma, deve-se supor que a população estudada será grande para que sejam minimizados os efeitos da não reposição e assim se poder utilizar a distribuição multinomial.

As categorias da variável resposta podem ser nucleotídeos ou aminoácidos. Logo estas categorias podem variar de 4 (para os nucleotídeos) a 20 (para aminoácidos). A variável resposta é categórica para cada posição. No caso de nucleotídeos, as categorias são: Adenina (A), Timina (T), Citosina (C), Guanina (G).

Quando a variável resposta é o nucleotídeo, os dados tendem a ter mais mutações do tipo transversão do que transição por causa da sua estrutura molecular. Um exemplo é uma posição que originalmente deveria ter o nucleotídeo A: as mutações, se acontecerem, serão com maior freqüência para o nucleotídeo T do que para C e G, que possivelmente terão freqüências zero ou muito próximo disto. Nestes casos, não haverá problema em utilizar a CATANOVA no caso de ter um número grande de caselas zero se todas as categorias tiverem alguma freqüência não-zero.

Os dados de seqüências genômicas podem ser resumidos em forma de tabela tridimensional, como sugerido por Andrade e Pinheiro (2002), mostrado na Tabela 5.6.

É importante ressaltar que aqui está sendo usada a seguinte ordem nos índices de n: Categoria de nucleotídeos (i), Grupo (j) e Posição (k). Assim, o

número de indivíduos em uma dada categoria de nucleotídeo, grupo e posição é n_{ijk} como é visto nas caselas da tabela abaixo. Além disso, o número de indivíduos estudados para cada sítio de DNA ou posição no mesmo grupo (amostra) é dado por N_J. $\overline{N}$ é o tamanho médio nos J grupos.

Tabela 5. 6. Tabela de Contingência utilizada em CATANOVA para um estudo de seqüências de DNA

		Categorias de Nucleotídeos (i)				
Grupo (j)	Sítio do DNA (k)	1	2		$n_{I\,11}$	Total
1	1	n_{111}	n_{211}	⋯	n_{I12}	$n_{+11}=N_1$
1	2	n_{112}	n_{212}	⋯	⋮	$n_{+12}=N_1$
⋮	⋮	⋮	⋮	⋯	n_{I1K}	⋮
1	K	n_{11K}	n_{21K}	⋯	n_{I1+}	$n_{+1K}=N_1$
Total		n_{11+}	n_{21+}		n_{I21}	$n_{+1+}=N_1K$
2	1	n_{121}	n_{221}	⋯	n_{I22}	$n_{+21}=N_2$
2	2	n_{122}	n_{222}	⋯	⋮	$n_{+22}=N_2$
⋮	⋮	⋮	⋮	⋯	n_{I2K}	⋮
2	K	n_{12K}	n_{22K}	⋯	n_{I2+}	$n_{+2K}=N_2$
Total		n_{12+}	n_{22+}	⋯	⋯	$n_{+2+}=N_2K$
⋯	⋯	⋯	⋯	⋯	n_{IJ1}	⋯
J	1	n_{1J1}	n_{2G1}	⋯	n_{IJ2}	$n_{+J1}=N_J$
J	2	n_{1J2}	n_{2G2}	⋯	⋮	$n_{+J2}=N_J$
⋮	⋮	⋮	⋮	⋯	n_{IJK}	⋮
J	K	n_{1JK}	n_{2JK}	⋯	n_{IJ+}	$n_{+JK}=N_J$
Total		n_{1J+}	n_{2J+}	⋯	n_{I++}	$n_{+J+}=N_JK$
TOTAL		n_{1++}	n_{2++}	⋯	$n_{I\,11}$	$n_{+++}=\overline{N}\,JK$

A hipótese nula para medida de associação parcial estabelece que a probabilidade de uma unidade amostral (indivíduo ou seqüência) pertencer à categoria i, no grupo j e posição k, p_{ijk}, é igual à probabilidade de uma unidade amostral pertencer à categoria i e posição k independente do grupo, p_{ik}:

H_0: $p_{ijk}=p_{ik}$, $\forall$ grupo, j=1,...,J; categoria, i=1,..., I; posição, k=1,...,k

Os cálculos das medidas de associação e as estatísticas C foram apresentados no Capítulo 4. Neste caso específico, não há interesse em trabalhar com uma medida de associação parcial do grupo em cada posição (Medida de Associação entre resposta e grupo no sítio 1 $R^2_{ij|(k=1)}$, por exemplo), tendo-se apenas interesse na medida parcial controlando todas as

posições($R^2_{ij|k}$). Isto é, deseja-se avaliar o efeito do "grupo" sobre a freqüência das categorias A, C, T e G, usando informações de vários sítios (ou posições) no DNA.

5.2.2.1. BOOTSTRAP

Andrade e Pinheiro (2002) recomendam utilizar técnicas de reamostragem, como o *bootstrap*, para amostras pequenas, quando há dados faltando (*missing*) ou quando as posições dentro de um grupo não são independentes (por exemplo, o grupo é um indivíduo apenas).

A técnica de bootstrap tenta realizar o que seria desejável na prática: **repetir a experiência**. Esta técnica trata a amostra como se ela representasse exatamente toda a população, o que pode ser um problema se a amostra não for representativa.

O *bootstrappping* tem como objetivo selecionar amostras aleatórias do mesmo tamanho, como se fosse a amostra original, ao contrário da Jackknife e outras técnicas, que forma novas amostras omitindo uma observação diferente de cada vez. Na técnica de *bootstrappping*, todas as observações originais possuem a mesma chance de serem escolhidas de forma aleatória em qualquer estágio. Logo, algumas observações podem não aparecer em nenhuma reamostragem, e em outras, aparecer diversas vezes.

Andrade e Pinheiro (2002) resumem os procedimentos de *bootstrappping* utilizados para um caso específico no qual se deseja testar (apenas) a hipótese principal que avalia se existe diferença entre dois grupos tendo-se tamanhos amostrais (N) iguais. Resumidamente, deve-se:

1) Estimar a $\hat{p}_{ik}$, probabilidade (populacional) de uma unidade amostral pertencer à categoria i e posição k.

2) Calcular a estatística de teste.

3) Gerar N (=n_{jk}) seqüências ou indivíduos em cada um dos J grupos, com K posições em cada um, usando $\hat{p}_{ik}$.

4) Recalcular a estatística de teste (F1) da amostra *bootstrap* e chamá-lo de F_1^* .

5) Repetir os passos quantas vezes se desejar.

O valor de p é obtido por uma razão baseada no número de passos usados. Por exemplo, para 1000 repetições, o valor de p estimado é:

$$\hat{p} = (\text{nº de valores de } F_1^* \geq F_1 \text{ observado}) / 1000 \qquad \textbf{5. 2}$$

Como $\hat{p}$ é uma variável de distribuição binomial, seu erro padrão pode ser estimado através de $\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}}$ e calcula-se então um intervalo de confiança para $\hat{p}$ que servirá como critério para se rejeitar ou não rejeitar a hipótese nula (Roff e Bentzen, 1989).

5.2.2.2. CATANOVA e a MANOVA no caso de seqüências genômicas

No caso de seqüências genômicas, utiliza-se a Análise de Variância Multivariada (MANOVA) porque se analisa as posições como coordenadas de respostas multivariadas.

A CATANOVA é utilizada como alternativa para modelos de MANOVA quando os dados possuem dimensão elevada (várias posições), pois aumenta mais a complexidade da modelagem e do esquema de análise.

Nestas situações, a MANOVA tem pouco poder quando existem muito mais posições do que o número de seqüências.

5.2.2.3.Vantagem da estatística **C** em comparações a outras estatísticas

Em Andrade e Pinheiro (2002), relatam uma outra estatística, além da estatística **C**, a estatística de Pearson de ajuste ou estatística χ^2 para testar homogeneidade:

$$\chi_P^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{J\left(n_{ijk} - \frac{n_{i.k}}{J}\right)^2}{N \times n_{i.k}} \qquad \textbf{5. 3}$$

Os graus de liberdade deste teste seria K(J-1)(I-1).

Um dos problemas em utilizar esta estatística em vez da estatística **C** é o de que os graus de liberdade geralmente são muito grandes, e conseqüentemente o poder pode não ser alto.

Outro problema é as suposições necessárias para o teste: cerca de 80% das freqüências esperadas das caselas serem maiores que 5 e não existir qualquer estrutura de dependência estocástica nas K posições. Sabe-se que estas suposições são dificilmente atendidas no caso de seqüências genômicas. Isto porque é difícil não haver alguma posição com freqüências muito baixas que, conseqüentemente, serão compensadas em outras posições. Já o problema da dependência ocorre tanto na estatística **C** como a χ^2 porque as seqüências genômicas não são totalmente independentes.

Para tentar resolver o problema de freqüências de caselas menores que 5, normalmente utiliza-se o Teste de Fisher. Entretanto, nos casos de seqüências genômicas, mesmo este teste terá valor duvidoso já que o tamanho de cada grupo (N_J) pode ser pequeno enquanto o número dos grupos e posições pode ser grande, o que resultará em tabelas com grande número de caselas com freqüência zero.

Também a independência entre grupos e entre seqüências não implica em um modelo produto de multinomiais por causa das interrelações posicionais.

Por último, a estatística **C** oferece muito mais informações que a χ^2, pois utiliza medidas de associações tipo R^2, que informam sobre a proporção de variabilidade atribuída ao(s) fator(es) em estudo.

5.2.2.4. CATANOVA e Modelos Log-Lineares

Assim como a CATANOVA, técnicas estatísticas usando Modelos Log-Lineares analisam dados categorizados dispostos em tabelas de contingência. O maior interesse em utilizar estes modelos é devido ao fato de que é possível criar Modelos Log-Lineares Hierárquicos, semelhante ao método de parametrização encontrado na ANOVA.

Os Modelos Log-Lineares Hierárquicos podem não ser adequados para dados categóricos nominais se não houver informações adicionais sobre a estrutura do mesmo. No caso de dados genéticos, sabe-se que existe uma certa dependência entre as posições, contudo não se sabe qual é a estrutura exata.

Um algoritmo para realizar análises via modelos Modelos Log-Lineares consta no programa ©SAS versão 8, mas não deve ser confundido com a Análise de Variância para Dados Categóricos, que não foi contemplada neste programa.

5.2.3. Exemplo de CATANOVA aplicado em seqüências genômicas

Bortolini et al. (1997) investigaram a variação dentro e entre populações indígenas quanto ao DNAmit. Eles trabalharam com este marcador genético por ele ser muito informativo quanto à variabilidade genética de grupos.

Os autores desejam comparar dados de DNAmit em seis grupos humanos, indígenas e afro-brasileiros. A CATANOVA será utilizada apenas nos dados de DNAmit dos grupos indígenas.

O banco de dados deste artigo trata de seqüências genômicas de DNA mitocondrial não-codificado. Ao todo são 82 indivíduos agrupados em três populações diferentes. A variável resposta é o tipo de nucleotídeo.

As tribos indígenas estudadas e algumas de suas características são dadas abaixo:

- **Xavante** (XAV): localiza-se no estado de Mato Grosso, no planalto central brasileiro, pertencem ao grupo lingüístico do tronco Jê (Línguas Paleo-americanas).
- **Zoró** (ZOR): localiza-se no estado de Mato Grosso, no sudoeste da Amazônia, pertencem ao grupo lingüístico Tupi-Mondé.
- **Gavião** (GAV): localiza-se no estado de Roraima, no sudoeste da Amazônia, pertencem ao grupo lingüístico Tupi-Mondé como a tribo Zoró.

O número de indivíduos estudados foi 25 em Xavante, 30 em Zoró e 27 em Gavião.

A tabela de contingência deste estudo tem a mesma estrutura dada no Tabela 5.7 referente à CATANOVA aplicada em seqüências genômicas. Esta tabela possui três grupos (tribos), 60 posições e quatro categorias na variável resposta, que são os nucleotídeos (A, T, C, G). Esta tabela está em anexo 4.

Os cálculos para a realização do teste para a CATANOVA foram feitos no Excel®.

Apesar de aparentemente parecer complicado obter as fórmulas para chegar à estatística **C**, elas estão obtidas de forma relativamente fácil se os dados forem colocados na forma da Tabela 5.6.

As somas de quadrados necessárias para os cálculos das medidas de associação estão na Tabela 5.7, e os resultados da Análise de Variância para Dados Categóricos são apresentados na Tabela 5.8. Não serão apresentadas as medidas de associação parcial em cada posição separadamente, por não haver interesse nestas medidas no caso de seqüências genômicas.

Tabela 5. 7. Resultados das Soma de Quadrados necessários para o cálculo das medidas de associação

Soma de Quadrados	Resultados
SQT	1440,242
SQE (j,k)	143,744
SQE (k)	211,012
SQE(j)	1438,991
SQR (j,k)	1296,498
SQR (j/k)	67,268

Tabela 5. 8. Medidas de associação e a estatística C

Efeito	Medidas de Associação	R^2	Estatística Baseada no R^2		
			C	gl	p
Efeito do Grupo	R^2_{ij}	0,0009	12,823	6	0,046
Efeito da Posição	R^2_{ik}	0,8535	12594,929	177	0,000
Efeito da Combinação do Grupo com Posição	R^2_{ijk}	0,9002	13284,167	537	0,000
Efeito do Grupo, controlando pela Posição	$R^2_{ij/k}$	0,3188	4704,326	360	0,000

Neste exemplo, todas as medidas de associação foram estatisticamente significativas (α=0,05). A combinação grupo/posição explicou 90,02% da variação observada nos dados. A variável posição (ou sítio no DNA) considerada sozinha, como efeito principal, é responsável por 85,35% da variação total, mostrando que os sítios diferem muito entre si quanto à

freqüência dos vários nucleotídeos. Já a proporção de variação explicada pelo grupo, sem considerar a posição, é de apenas 0,09%.

Observando os dados de outro modo, pode-se considerar cada posição como uma variável diferente. Conclui-se que a diversidade entre posições (ou sítios) quanto à freqüência das diferentes bases (A, C, T, G) é muito maior (85,35%) do que a observada entre as tribos indígenas (<1%). O fato da proporção explicada pelas posições ser maior do que a explicada pelos grupos se deve ao comportamento altamente variável entre as posições nas seqüências genômicas. Também se observou que os grupos indígenas diferem geneticamente entre si (p=0,046).

Quando se mede o efeito do fator "tribo" sobre a freqüência dos nucleotídeos, controlando pelos sítios estudados, obtém-se um notável aumento no valor de R^2 (31,88%; p<0,001), mostrando que o controle do fator posição levou a um aumento no poder discriminatório do fator tribo. Assim, uma diferença possivelmente pequena entre tribos, quanto às freqüências de A, T, C e G, foi ressaltada ao se controlar pelo fator posição, evidenciando-se deste modo a vantagem de se utilizar uma técnica estatística que extraia o máximo de informação possível dos dados.

6. CONCLUSÃO

O teste tradicionalmente utilizado para dados categóricos nominais (e ordinais), dispostos em tabelas bidimensionais ou multidimensionais, é o Qui-Quadrado. Este, porém, muitas vezes não atende de forma completa o objetivo da pesquisa. Além disso, os dados muitas vezes não satisfazem as pressuposições teóricas deste teste, não se podendo também utilizar outras opções, como o Teste Exato de Fisher.

A Análise de Variância para Dados Categóricos dispostos em escala nominal (CATANOVA) propicia um teste estatístico mais completo e adequado, suficientemente geral para suprir as necessidades em inúmeras situações deste tipo. Ela tem como base medidas de associação R^2 para dados categóricos através das quais se pode obter a proporção da variação total explicada pelas diferentes variáveis de interesse.

Uma área que pode vir a utilizar este teste de forma a extrair mais informação dos dados é a área de genética.

Em um estudo sobre o efeito da linhagem e temperatura em mosca-das-frutas (*Drosophila willistoni*), Regner et al. (1999), utilizaram o Qui-Quadrado em suas comparações. Analisando pela CATANOVA, além das mesmas conclusões, obteve-se maiores informações por causa das medidas de associações encontradas.

E no que se refere a um estudo comparativo entre populações indígenas usando DNA mitocondrial (Bortolini et al, 1997), encontrou-se uma nova forma de analisar as seqüências genômicas, obtendo-se mais poder para mostrar a existência de diferença genética entre as tribos. Isso foi obtido usando uma medida de associação entre o Fator tribo e a resposta, controlando por um segundo fator.

Sem dúvida, a CATANOVA deve ser aprimorada, mas para isso é necessário difundir o conhecimento desta análise. Este trabalho espera ter contribuído para este objetivo, estimulando a realização de novos trabalhos que visem a melhorar esta técnica, a qual poderá a vir a ser utilizada no dia-a-dia, como é utilizado o Teste Qui-Quadrado.

7. PROPOSTAS FUTURAS

A Análise de Variância para Dados Categóricos (CATANOVA) tem grande possibilidade de ser utilizada nas mais diversas áreas, mas para isto é necessário mais trabalhos sobre este assunto. Algumas das propostas que devem ser ressaltadas são:

1. Criar um programa para os cálculos da CATANOVA;
2. Criar testes *post hoc* para comparar tratamentos como existe na ANOVA, mas com a diferença de utilizar proporções ao invés de médias;
3. Comparar a estatística C e χ^2 em tabelas multidimensionais através de simulações;
4. Encontrar uma função mais geral que relacione a estatística χ^2 com a medida de associação R^2;
5. Nos casos de seqüências genômicas, seria interessante encontrar uma forma de ponderar a freqüência das diferentes categorias da variável resposta em cada n_{ijk}, já que a probabilidade de ocorrência dos quatro nucleotídeos não é uniforme no mesmo sítio ou posição.

8. REFERÊNCIAS BIBLIOGRÁFICAS

Anderson RJ, Landis JR. CATANOVA for multidimensional contingency tables: nominal-scale response. **Communication in Statistics – Theory and Methods 9**: 1191-1206. 1980.

Anderson RJ, Landis JR. CATANOVA for multidimensional contingency tables: ordinal-scale response. **Communication in Statistics – Theory and Methods 11**: 257-270. 1982.

Andrade M, Pinheiro H. **Métodos Estatísticos Aplicados em Genética Humana**. 15º Simpósio Nacional de Probabilidade e Estatística (SINAPE). Águas de Lindóia, SP. ABE – Associação Brasileira de Estatística. 2002.

Bortolini MC, Salzano FM, Zago MA, Silva WA Jr, Weimer TA. Genetic variability in two Brazilian ethnic groups: a comparison of mitochondrial and protein data. **American Journal of Physical Anthropology 103**: 147-156. 1997.

Cochran WG. The comparison of percentages in matched samples. **Biometrika 37**: 256-266. 1950.

GINI CW. "Variabilita e Mutabilita" contributo all studio delle distribuzioni e relazioni statische. **Studi Econômico-Giuridici della R. Universita di Cagliari 3**: 159. 1912.

Goodman LA, Kruskal EH. Measures of association for cross-classifications. **Journal of the American Statistical Association 49**: 732-764. 1954.

Kullback S, Kupperman M, Ku HH. Tests for contingency tables and Markov Chains. **Technometrics 4**:573-608. 1962.

Light R J, Margolin BH. An analysis of variance for categorical data. **Journal of the American Statistical Association 66**: 534-544. 1971.

Margolin BH, Light RJ. An analysis of variance for categorical data II: small sample comparison with chi square and other competitors. **Journal of the American Statistical Association 69**: 755-764. 1974.

Regner LP, Abdelhay E, Rohde C, Rodrigues JJS, Valente VL. Temperature-dependent gonadal hybrid dysgenesis in *Drosophila willistoni*. **Genetics and Molecular Biology 22**(2): 205-211. 1999.

Roff DA, Bentzen P. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the Problem of Small Samples. **Molecular Biology and Evolution 6**(5): 539-545. 1989.

SIMPSON EH. The Measurement of Diversity. **Nature 163**: 688. 1949.

9. ANEXO

Anexo 1. Banco de dados para análise no exemplo de CATANOVA aplicado na genética de moscas-das-frutas (Regner et al, 1999).

Tabela 1.1. Fêmeas

		Tipos de Gonadas		
Temperatura (°C)	**Cruzamentos**	**Disgênicas**	**Normais**	Total
18	1	4	180	184
18	2	2	214	216
18	3	0	190	190
18	3	4	176	180
Total		10	760	770
25	1	2	172	174
25	2	0	180	180
25	3	2	166	168
25	4	10	172	182
Total		14	690	704
29	1	6	172	178
29	2	47	133	180
29	3	32	138	170
29	4	28	144	172
Total		113	587	700
TOTAL		137	2037	2174

Tabela 1.2. Machos

		Tipos de Gonadas		
Temperatura (°C)	**Cruzamentos**	**Disgênicas**	**Normais**	Total
18	1	3	171	174
18	2	6	210	216
18	3	4	192	196
18	3	4	158	162
Total		17	731	748
25	1	2	158	160
25	2	4	164	168
25	3	0	156	156
25	4	3	169	172
Total		9	647	656
29	1	3	123	126
29	2	21	81	102
29	3	14	86	100
29	4	16	106	122
Total		54	396	450
TOTAL		80	1774	1854

Anexo 2. Soma de Quadrados para descendentes fêmeas e machos, no exemplo de CATANOVA aplicado na genética de moscas-das-frutas (Regner et al, 1999).

Soma de Quadrados	Descendência	
	Fêmeas	Machos
SST	128,367	76,548
SSE (j,k)	113,153	70,975
SSE (j,k=1)	11,688	7,852
SSE (j,k=2)	36,709	26,415
SSE (j,k=3)	27,953	15,958
SSE (j,k=4)	36,804	20,751
SSE (k)	127,070	76,007
SSE (k=1)	11,731	7,861
SSE (k=2)	44,832	29,023
SSE (k=3)	31,811	17,283
SSE (k=4)	38,697	21,840
SSE (j)	118,350	73,010
SSR (j,k)	15,213	5,573
SSR (j/k)	13,917	5,031
SSR (j/k=1)	0,044	0,009
SSR (j/k=2)	8,122	2,608
SSR (j/k=3)	3,858	1,325
SSR (j/k=4)	1,893	1,089

Anexo 3. Soma de Quadrados para descendentes fêmeas e machos, no exemplo de CATANOVA aplicado na genética de moscas-das-frutas para o caso do Fator 2 ser linhagem x sexo dos descendentes (Regner et al, 1999).

Soma de Quadrados	Descendência
	Fêmeas e Machos
SST	205,310
SSE (j,k)	184,129
SSE (j,k=1)	11,688
SSE (j,k=2)	7,852
SSE (j,k=3)	36,709
SSE (j,k=4)	26,415
SSE (j,k=5)	27,953
SSE (j,k=6)	15,958
SSE (j,k=7)	36,804
SSE (j,k=8)	20,751
SSE (k)	203,077
SSE (k=1)	11,731
SSE (k=2)	7,861
SSE (k=3)	44,832
SSE (k=4)	29,023
SSE (k=5)	31,811
SSE (k=6)	17,283
SSE (k=7)	38,697
SSE (k=8)	21,840
SSE (j)	191,879
SSR (j,k)	21,181
SSR (j/k)	18,948
SSR (j/k=1)	0,044
SSR (j/k=2)	0,009
SSR (j/k=3)	8,122
SSR (j/k=4)	2,608
SSR (j/k=5)	3,858
SSR (j/k=6)	1,325
SSR (j/k=7)	1,893
SSR (j/k=8)	1,089

Anexo 4. Tabela de Contingência do exemplo de CATANOVA aplicado em seqüências genômicas estudados em índios: Xavante (XAV), Zoró (ZOR) e Gavião (GAV); Bortolini et al. (1997).

Grupo	Posição	Categorias de Nucleotídeos			
		A	T	C	G
XAV	16051A	25	0	0	0
XAV	16080A	25	0	0	0
XAV	16092T	0	25	0	0
XAV	16093T	0	16	9	0
XAV	16111C	0	4	21	0
XAV	16144C	0	0	25	0
XAV	16124T	0	25	0	0
XAV	16126T	0	25	0	0
XAV	16129G	0	0	0	25
XAV	16148C	0	0	25	0
XAV	16150C	0	0	25	0
XAV	16168C	0	1	24	0
XAV	16172T	0	25	0	0
XAV	16173C	0	0	25	0
XAV	16175A	25	0	0	0
XAV	16187C	0	0	25	0
XAV	16188C	0	0	25	0
XAV	16189T	0	4	21	0
XAV	16192C	0	0	25	0
XAV	16193C	0	0	25	0
XAV	16209T	0	25	0	0
XAV	16213G	0	0	0	25
XAV	16217T	0	4	21	0
XAV	16223C	0	4	21	0
XAV	16230A	25	0	0	0
XAV	16239C	0	0	25	0
XAV	16241A	16	0	0	9
XAV	16248C	0	0	25	0
XAV	16256C	0	0	25	0
XAV	16260C	0	0	25	0
XAV	16261C	0	0	25	0
XAV	16264C	0	0	25	0
XAV	16265A	25	0	0	0
XAV	16266C	0	0	25	0
XAV	16270C	0	0	25	0
XAV	16278C	0	0	25	0
XAV	16284A	0	0	25	0
XAV	16286C	0	0	25	0
XAV	16290C	0	4	21	0
XAV	16291C	0	0	25	0
XAV	16292C	0	0	25	0
XAV	16293A	25	0	0	0

XAV	16294C	0	0	25	0
XAV	16295C	0	0	25	0
XAV	16298T	0	25	0	0
XAV	16301C	0	0	25	0
XAV	16304T	0	25	0	0
XAV	16311T	0	25	0	0
XAV	16316A	25	0	0	0
XAV	16319G	4	0	0	21
XAV	16320C	0	0	25	0
XAV	16325T	0	25	0	0
XAV	16327C	0	0	25	0
XAV	16328C	0	0	25	0
XAV	16330T	0	25	0	0
XAV	16354C	0	0	25	0
XAV	16355C	0	0	25	0
XAV	16358C	0	0	25	0
XAV	16360C	0	0	25	0
XAV	16362T	0	21	4	0
ZOR	16051A	30	0	0	0
ZOR	16080A	30	0	0	0
ZOR	16092T	0	30	0	0
ZOR	16093T	0	26	4	0
ZOR	16111C	0	6	24	0
ZOR	16144C	0	0	30	0
ZOR	16124T	0	30	0	0
ZOR	16126T	0	30	0	0
ZOR	16129G	0	0	0	30
ZOR	16148C	0	0	30	0
ZOR	16150C	0	0	30	0
ZOR	16168C	0	0	30	0
ZOR	16172T	0	30	0	0
ZOR	16173C	0	0	30	0
ZOR	16175A	28	0	0	2
ZOR	16187C	0	0	30	0
ZOR	16188C	0	0	30	0
ZOR	16189T	0	25	5	0
ZOR	16192C	0	0	30	0
ZOR	16193C	0	0	30	0
ZOR	16209T	0	30	0	0
ZOR	16213G	0	0	0	30
ZOR	16217T	0	29	1	0
ZOR	16223C	0	29	1	0
ZOR	16230A	30	0	0	0
ZOR	16239C	0	0	30	0
ZOR	16241A	30	0	0	0
ZOR	16248C	0	0	30	0
ZOR	16256C	0	1	29	0
ZOR	16260C	0	0	30	0
ZOR	16261C	0	0	30	0
ZOR	16264C	0	0	30	0

ZOR	16265A	30	0	0	0
ZOR	16266C	0	1	29	0
ZOR	16270C	0	0	30	0
ZOR	16278C	0	6	24	0
ZOR	16284A	30	0	0	0
ZOR	16286C	0	0	30	0
ZOR	16290C	0	6	24	0
ZOR	16291C	0	6	24	0
ZOR	16292C	0	0	30	0
ZOR	16293A	30	0	0	0
ZOR	16294C	0	0	30	0
ZOR	16295C	0	0	30	0
ZOR	16298T	0	26	4	0
ZOR	16301C	0	0	30	0
ZOR	16304T	0	28	2	0
ZOR	16311T	0	30	0	0
ZOR	16316A	29	0	0	1
ZOR	16319G	6	0	0	24
ZOR	16320C	0	0	30	0
ZOR	16325T	0	7	23	0
ZOR	16327C	0	4	26	0
ZOR	16328C	0	0	30	0
ZOR	16330T	0	30	0	0
ZOR	16354C	0	0	30	0
ZOR	16355C	0	0	30	0
ZOR	16358C	0	0	30	0
ZOR	16360C	0	0	30	0
ZOR	16362T	0	6	24	0
GAV	16051A	27	0	0	0
GAV	16080A	27	0	0	0
GAV	16092T	0	26	1	0
GAV	16093T	0	27	0	0
GAV	16111C	0	4	23	0
GAV	16144C	0	0	27	0
GAV	16124T	0	27	0	0
GAV	16126T	0	27	0	0
GAV	16129G	0	0	0	27
GAV	16148C	0	0	27	0
GAV	16150C	0	0	27	0
GAV	16168C	0	0	27	0
GAV	16172T	0	27	0	0
GAV	16173C	0	0	27	0
GAV	16175A	21	0	0	6
GAV	16187C	0	0	27	0
GAV	16188C	0	0	27	0
GAV	16189T	0	14	13	0
GAV	16192C	0	0	27	0
GAV	16193C	0	0	27	0
GAV	16209T	0	27	0	0
GAV	16213G	0	0	0	27

GAV	16217T	0	23	4	0
GAV	16223C	0	23	4	0
GAV	16230A	27	0	0	0
GAV	16239C	0	0	27	0
GAV	16241A	27	0	0	0
GAV	16248C	0	0	27	0
GAV	16256C	0	0	27	0
GAV	16260C	0	0	27	0
GAV	16261C	0	4	23	0
GAV	16264C	0	0	27	0
GAV	16265A	27	0	0	0
GAV	16266C	0	0	27	0
GAV	16270C	0	0	27	0
GAV	16278C	0	3	24	0
GAV	16284A	22	0	0	5
GAV	16286C	0	0	27	0
GAV	16290C	0	4	23	0
GAV	16291C	0	0	27	0
GAV	16292C	0	0	27	0
GAV	16293A	27	0	0	0
GAV	16294C	0	0	27	0
GAV	16295C	0	0	27	0
GAV	16298T	0	27	0	0
GAV	16301C	0	0	27	0
GAV	16304T	0	25	2	0
GAV	16311T	0	27	0	0
GAV	16316A	27	0	0	0
GAV	16319G	4	0	0	23
GAV	16320C	0	0	27	0
GAV	16325T	0	8	19	0
GAV	16327C	0	0	27	0
GAV	16328C	0	0	27	0
GAV	16330T	0	27	0	0
GAV	16354C	0	0	27	0
GAV	16355C	0	0	27	0
GAV	16358C	0	0	27	0
GAV	16360C	0	0	27	0
GAV	16362T	0	4	23	0