UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E
COMPUTAÇÃO CIENTÍFICA - IMECC

Estudo da Variabilidade Genética Através de Análise de Variância Para Dados Categorizados em Amostras Não Balanceadas

Roberta de Souza

Profa. Dra. Hildete Prisco Pinheiro
Orientador

Profa. Dra. Cibele Queiroz da Silva
Coorientador

Campinas - SP, Julho 2002

Estudo da Variabilidade Genética Através de Análise de Variância para Dados Categorizados em Amostras Não Balanceadas

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Roberta de Souza e aprovada pela comissão julgadora.

Campinas, 03 de agosto de 2002.

Profa. Dra. Hildete Prisco Pinheiro

Banca Examinadora:
1. Profa. Dra. Hildete Prisco Pinheiro (IMECC/UNICAMP)
2. Prof. Dr. Júlio da Motta Singer (IME/USP)
3. Prof. Dr. Sérgio Furtado dos Reis (IB/UNICAMP)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em Estatística.

Resumo

No estudo de diferenças genéticas entre organismos, geralmente a análise é feita diretamente da molécula de DNA. Assim, podemos ter como resposta um valor do tipo binário (fenótipo dominante ou recessivo) ou, quando observamos diretamente cada nucleotídeo ao longo da seqüência, a resposta é categórica podendo apresentar uma entre quatro categorias. Weir (1990) propôs uma medida de diversidade, a *heterozigosidade*, e desenvolveu uma análise de variância para seqüências de dados binários em amostras balanceadas. Light e Margolin (1974), desenvolveram uma análise de variância para dados categóricos (CATANOVA), que pode ser aplicada ao contexto de seqüências genômicas, considerando-se uma única posição na seqüência. Pinheiro et al. (2000), baseados numa medida de diversidade proposta por Simpson (1949), estenderam a CATANOVA para várias posições ao longo da seqüência. Aqui estendemos o trabalho de Weir e Pinheiro et al. desenvolvendo análises de variâncias para seqüências de dados de natureza binária e categórica de um modo geral (mais de duas categorias de resposta), com o intuito de comparar grupos de diferentes tamanhos (amostras não balanceadas). Para testar a hipótese nula de homogeneidade entre grupos foram encontradas as distribuições assintóticas das estatísticas dos testes nos dois casos: variável resposta binária e com mais de duas categorias. Aplicações destes testes à dados reais são ilustradas utilizando técnicas de reamostragem como *bootstrap* para a geração das distribuições empíricas das estatísticas dos testes.

Abstract

In the study of genetic divergence among organisms, generally the analysis is done directly from the DNA molecule. Therefore, the response could be binary (dominant or recessive phenotype) or, if one observe the nucleotides over the whole sequence, the outcome is categorical being one out of for categories. Weir (1990) proposed a genetic diversity measure, the *heterozygosity*, and developed an analysis of variance of binary data in a balanced design. Light and Margolin (1974) developed an analysis of variance for categorical data (CATANOVA), which can be applied to the context of genomic sequences considering only one position. Pinheiro et al. (2000), based on a diversity measure proposed by Simpson (1949), extend the CATANOVA to various positions along the sequence. Here, we extend the work of Weir and Pinheiro et al. developing analysis of variance for binary and categorical data in general (more than two categories of response) with the purpose of comparing groups in unbalanced designs. In order to test the null hypothesis of homogeneity among groups, the asymptotic distribution of the test statistic were found in both cases: binary and response with more than two categories. Application of tests to real data are illustrated using resampling methods such as the bootstrap to generate the empirical distribution of the test statistics.

Ao meu pai José Carlos de Souza (*in memorian*).

Agradecimentos

À Deus.

Aos meus meus queridos pais, José Carlos e Cláudia, por todo amor e incentivo.

À minha amável irmã Gláucia por todo o apoio e amizade.

Ao meu adorável namorado Anderson, por estar sempre presente com tamanha paciência, incentivo, carinho e amor.

À minha querida orientadora Profa. Hildete P. Pinheiro por, além de outras coisas, toda dedicação e paciência.

À minha doce coorientadora, Profa. Cibele Queiroz da Silva, por todas as correções e sugestões.

Aos professores do Departamento de Estatística, em especial aos professores Aluísio Pinheiro e Filidor E. V. Labra por toda ajuda.

Ao Prof. Sergio F. dos Reis e a Franco de Sousa pela boa vontade e todo o material cedido para o meu aprendizado em conceitos biológicos.

Aos meus colegas do Departamento de Estatística, em especial ao Benilton, Clécio, Samara e Tatiana por, além do companheirismo, toda a colaboração.

Às "figuras"Brutus, Camila, Neale e Mariu por compartilharem diversas situações importantes, sendo inesquecíveis as desestressantes e divertidas noites de quartas-feiras.

À banca examinadora pelas enriquecedoras correções e sugestões.

À CAPES pelo suporte financeiro.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução e Revisão Bibliográfica

1.1 Motivação

No estudo de comparação de populações (grupos), uma ferramenta estatística bastante utilizada é a análise de variância clássica. No entanto, este tipo de análise só é adequada quando a variável resposta é contínua. Nos casos em que a variável resposta é categorizada, outros procedimentos estatísticos precisam ser utilizados. Algumas análises de variâncias foram desenvolvidas para o caso em que a variável resposta é categorizada (Light & Margolin, 1971; Weir, 1990 e Pinheiro *et al.*, 2000), mas as amostras, em geral, são balanceadas. O objetivo deste trabalho é obter estatísticas para que se possa testar a hipótese de homogeneidade entre grupos e assim desenvolver análises de variâncias, para dados categorizados, quando as amostras são não balanceadas.

Uma das motivações para o desenvolvimento de análises de variância para dados categorizados é o estudo da variabilidade genética entre organismos (genética populacional). Uma característica deste estudo é o processo de evolução. Quando animais da mesma espécie se encontram isolados geograficamente (por rios, montanhas, etc) tendem a se diferenciar, pois a mudança de ambiente favorece a ação da seleção natural, o que pode levar a uma mudança na composição genética. A ocorrência de mutações casuais do material genético ao longo do tempo leva a um aumento da variabilidade, que permite a continuidade da atuação da seleção natural. A persistência, a longo prazo, de espécies ameaçadas de extinção depende da existência de variabilidade genética em suas populações como espera-se que esteja ocorrendo com cágados da espécie *Hydromedusa maximiliani*. O es-

tudo da variabilidade genética também pode gerar subsídios para o controle de parasitas, como a espécie *Dermatobia hominis* (mosca do berne), que ameaçam a economia em diversas regiões do país. Estes são exemplos de aplicações das análises de variância para dados categorizados que serão vistos com mais detalhes no capítuo 4.

Diferenças genéticas podem ser obtidas através da análise molecular direta do DNA ou através de proporções de heterozigotos e/ou de homozigotos, entre outros. Com o interesse nesta aplicação serão apresentados alguns conceitos básicos de genética e biologia molecular na Seção 1.2. Algumas medidas de diversidade, para dados categorizados, utilizadas no estudo da variação genética, serão vistas na Seção 1.3. Weir (1990) introduziu a *heterozigosidade* observada como uma medida de diversidade em termos das proporções de heterozigotos e desevolveu a análise de variância para amostras balanceadas (Seção 1.3.1). Outra medida apresentada por Weir, a Diversidade Genética, será vista na Seção 1.3.2. Quando a variável resposta é categorizada, a variância amostral (medida utilizada quando a variável resposta é continua), não é uma medida bem apropriada da variação. Gini (1912) encontrou uma maneira alternativa de caracterizar a variação entre dados categorizados em termos da soma de quadrados das proporções de respostas em cada categoria. Anos depois, Simpson (1949), propôs uma medida proporcional à medida de Gini, o Índice de Simpson. Baseados nestas medidas, Pinheiro *et al.* (2000) apresentaram algumas medidas de diversidade para dados categorizados em geral e, estendendo o trabalho de Light & Margolin (1971), desenvolveram uma análise de variância para dados categorizados (CATANOVA) em amostras balanceadas, considerando como resposta um vetor aleatório com distribuição multinomial (Seção 1.3.3).

No Capítulo 2 foram comparados grupos onde temos como resposta um vetor de valores binários, isto é, os elementos do vetor podem assumir os valores 0 ou 1. Na Seção 2.2 estendemos o trabalho de Weir (1990) para grupos de diferentes tamanhos (amostras não balanceadas). Utilizando uma medida similar, definida em termos das proporções de fenótipos dominantes, foi desenvolvida uma análise de variância para amostras não balanceadas (Seção 2.3). Na Seção 2.4, com base no modelo de Bernoulli, propôs-se uma estatística de teste para a hipótese de homogeneidade entre grupos contra a hipótese alternativa de heterogeneidade entre grupos. A distribuição assintótica da estatística do teste foi obtida assumindo independência entre as populações, entre os indivíduos e entre as posições correspondentes aos elementos do vetor de respostas. Um estudo sobre o poder do teste é apresentado na Seção 2.5.

Uma técnica de análise de variância para dados categorizados é desenvolvida no Capítulo 3. Neste caso, a variável resposta é um vetor onde cada elemento pode corresponder a mais do que duas categorias de respostas. Podemos, por exemplo, estar interessados em comparar grupos de seqüências genômicas. No caso de seqüências de nucleotídeos, cada elemento do vetor de respostas corresponde a uma entre quatro cate-

gorias e, para seqüências de aminoácidos, cada elemento corresponde a uma entre vinte categorias. Estendendo o trabalho de Pinheiro *et al.* (2000) para o caso de amostras não balanceadas, foram obtidas medidas de diversidade e seus respectivos momentos baseando-se no modelo probabilístico multinomial (Seções 3.2, 3.3 e 3.4). Para testarmos a hipótese de homogeneidade entre grupos de diferentes tamanhos propôs-se uma estatística de teste utilizando teoria assintótica; obteve-se a distribuição dessa estatística (Seção 3.5); o poder do teste foi investigado na Seção 3.6.

Os resultados obtidos nos Capítulos 2 e 3 foram aplicados a dados reais no Capítulo 4. Em alguns casos, o tamanho da amostra não é grande o suficiente para que possamos aplicar resultados assintóticos. Por esse motivo, recorremos à metodos de reamostragem, como o *bootstrap*, para a geração da distribuição empírica da estatística do teste para a hipótese nula de homogeneidade entre grupos.

1.2 Conceitos Básicos de Genética e Biologia Molecular

O DNA (ácido desoxirribonucléico) é formado por seqüências de nucleotídeos. Cada nucleotídeo é composto de um fosfato, um açucar, a desoxirribose, e uma base nitrogenada, que podem ser de quatro tipos: *citosina* (**C**), *timina* (**T**), *adenina* (**A**) e *guanina* (**G**). Estas são as bases nitrogenadas que guardam a informação importante para a síntese do RNA (ácido ribonucléico), cuja molécula é composta por uma única cadeia de nucleotídeos sendo o açúcar a ribose no lugar da desoxirribose e o tipo de base *uracila* (**U**) substitui a **T** do DNA. Assim, os nucleotídeos podem ser de quatro tipos de acordo com a base presente, pois o açúcar e o fosfato são componentes invariáveis nos nucleotídeos apresentando uma função unicamente estrutural.

A molécula de DNA é constituída de duas cadeias de nucleotídeos em orientação oposta que se enrolam originando a dupla hélice. As duas cadeias são complementares pois **A** sempre se liga a **T** e **C** sempre pareia com **G**. Portanto, podemos dizer que uma molécula de DNA é constituída por milhares de pares de bases (pb).

O DNA de uma célula humana apresenta um comprimento total de quase 2 metros dividido em vários elementos distintos chamados *cromossomos*. Cada *cromossomo* é formado por uma única molécula de dupla hélice de DNA condensada com proteína. O número e a forma dos cromossomos são características de cada espécie e todas as células de um organismo apresentam o mesmo número de cromossomos (células diplóides) (Farah, 1997).

Em organismos com reprodução sexuada o indivíduo é formado a partir da fusão de um espermatozóide com um óvulo, os *gametas*, que dão origem à célula ovo (célula diplóide). A partir daí inúmeras divisões celulares, do tipo *mitose* (da célula mãe inicial originam-se duas células filhas identicas), vão ocorrendo até a formação do indivíduo adulto. Para manter o número cromossômico da espécie, o óvulo e o espermatozóide devem conter a metade do número de cromossomos presentes nas outras células e portanto os gametas são células haplóides. A divisão celular responsável pela formação dos gametas é chamada de *meiose* (da célula mãe inicial originam-se quatro células filhas haplóides). Portanto, em cada célula humana existem diversos pares de cromossomos sendo que em cada par existe um cromossomo de origem materna e outro de origem paterna (*cromossomos homólogos*).

O *genoma* é o conjunto completo de cromossomos haplóides que contém toda a informação genética de um indivíduo (Farah, 1997). Sendo assim, os gametas possuem uma cópia do genoma, enquanto as células diplóides, também chamadas de células somáticas, apresentam duas cópias. *Gene* é a unidade física funcional da hereditariedade, a qual transmite informação de uma geração para outra. Em termos moleculares, é uma seqüência de DNA, que inclui exons, introns e regiões de controle de transcrição não codificadoras, necessária para a produção de uma proteína funcional ou RNA (Lodish et al, 2000). Existe uma distinção entre os genes e as outras seqüências de DNA presentes no genoma, pois é a partir dos genes que ocorre a síntese de RNA. Esta atividade é chamada de *transcrição*. Os genes humanos representam apenas 2 a 3% no genoma.

O *locus* representa a posição que cada gene ocupa no genoma. Os cromossomos homólogos apresentam os mesmos *loci* e na mesma ordem, ou seja, em cada locus os cromossomos exibem genes para a mesma característica genética. A informação genética, seqüências de bases, contida em cada locus dos cromossomos homólogos pode não ser idêntica. Estas formas alternativas de um gene são chamadas de *alelos*. Dizemos que um indivíduo é *homozigoto* para uma certa característica genética quando, para um dado locus, ele possui alelos idênticos em cada cromossomo homólogo. Quando estes alelos são diferentes o indivíduo é chamado de *heterozigoto*.

O conjunto de todos os alelos presentes em determinados *loci* constitui o *genótipo*, enquanto as características que se observam em um indivíduo resultante de sua constituição genética mais a influência de fatores do meio ambiente representam seu *fenótipo*. Se o fenótipo se expressa mesmo quando somente um cromossomo do par de homólogos apresenta o alelo correspondente dizemos que este fenótipo é *dominante*. Se a expressão de um fenótipo exige a presença de dois alelos do mesmo tipo trata-se de um fenótipo *recessivo*. Os fenótipos podem ser quantitativos ou qualitativos. Um fenótipo é considerado quantitativo se ele é medido numa escala contínua, como a altura e o peso; é considerado qualitativo quando é medido de forma categorizada, como o nível de gravidade de uma doença (dicotômico ou politômico).

1.2.1 Marcadores Moleculares

Os métodos e técnicas da biologia molecular permitem o acesso ao genoma de qualquer organismo, sendo possível investigar a variação genética ao nível do DNA. Pontos de referência nos cromossomos são denominados *marcadores moleculares*. Por marcador molecular define-se todo e qualquer fenótipo molecular oriundo de um gene expresso ou um segmento específico de DNA, correspondente a regiões expressas ou não expressas do genoma (Ferreira e Grattapaglia, 1998). Uma das técnicas disponíveis para a detecção de polimorfismos genéticos moleculares é a classe de marcadores moleculares denominada RAPD, *Random Amplified Polymorphic DNA* ou polimorfismo de DNA amplificado ao acaso. Estes marcadores são gerados pela amplificação de segmentos de DNA através da reação em cadeia de polimerase - PCR (*polymerase chain reaction*). Nesta técnica nenhum conhecimento à priori do DNA é necessário. Utiliza-se uma seqüência arbitrária do DNA, com 9 ou 10 pb (pares de bases), denominada *primer*. Para que haja a amplificação de um fragmento RAPD no genoma analisado, duas seqüências de DNA complementares ao *primer* arbitrário devem estar a uma distância menor do que 4000 pb e em orientação oposta, de maneira a permitir a amplificação do segmento de DNA através da enzima DNA - polimerase. Portanto, o *locus* dentro do genoma que contém o segmento a ser ampliado é obtido aleatoriamente. Então, a freqüência de amplificação de um segmento depende da probabilidade de o *primer* utilizado encontrar duas seqüências complementares (sítios de iniciação) em sentidos opostos ao longo da seqüência de DNA e a uma distância amplificável. Em função da grande quantidade de DNA produzido, este segmento pode ser visualizado diretamente num gel de eletroforese, na forma de uma banda, através de corantes específicos para DNA. Devido a síntese de vários segmentos de DNA dirigidos por um único *primer* pode-se observar várias bandas no gel (Figura 1.1). O polimorfismo genético detectado pelos marcadores moleculares RAPD tem natureza binária, isto é, o segmento amplificado (banda no gel) está presente ou ausente. Basta a diferença de um par de base para causar a não complementaridade do *primer* com o sítio de iniciação. Estes marcadores exibem uma característica fundamental que é o fato de se comportarem como marcadores genéticos dominantes, pois ao se observar uma banda no gel, não é possível distinguir se aquele segmento se originou a partir de uma ou de duas cópias da seqüência amplificada. O segmento de DNA é detectado através da presença da banda em ambos os casos em que o indivíduo é diplóide homozigoto para um locus de RAPD, onde ocorre a amplificação desses dois alelos idênticos (AA) e, no caso em que o indivíduo é heterozigoto (Aa) para o mesmo locus, onde o alelo A é amplificado e o alelo a não o é. Portanto, os genótipos homozigotos dominantes (AA) e heterozigotos (Aa) são colocados juntos na mesma classe fenotípica (presença da banda) enquanto que

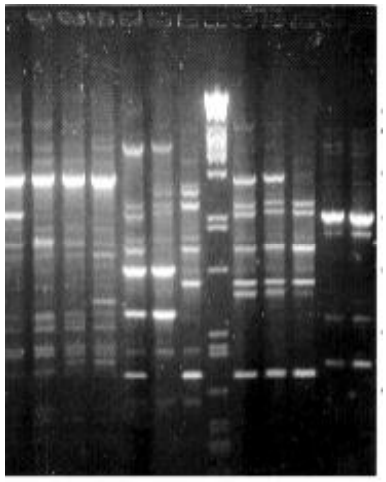

Figura 1.1: Bandas de RAPD

o genótipo homozigoto recessivo (aa), cujo locus não ocorre a amplificação, é identificado pela ausência da banda no gel.

1.3 Medidas de Diversidade e Procedimentos Estatísticos no Estudo da Variação Genética

O estudo da evolução de espécies é caracterizado pelas extensões e causas de variação genética. Existem diferentes maneiras de medir variação genética; entre elas podemos pensar na proporção de heterozigotos numa população, a *heterozigosidade*, pois cada indivíduo carrega diferentes alelos, o que representa a existência de variacão. A presença continuada de diferentes homozigotos também pode resultar em variação; para isso a diversidade genética é uma medida apropriada. Diferenças genéticas também são encontradas por análise molecular direta do DNA. Neste caso a variação pode ser medida comparando-se os nucleotídeos. O Índice de Simpson pode ser utilizado como uma medida de diversidade para a comparação de seqüências genômicas.

1.3.1 Heterozigosidade

Weir (1990) propôs a *heterozigosidade* observada como uma medida de variabilidade genética numa população . Seja n_{luv} o número observado de heterozigotos A_uA_v, $u \neq v$, num locus l numa amostra de tamanho n. Então a proporção amostral de heterozigotos no locus l é

$$\tilde{H}_l = \sum_u \sum_{v>u} \frac{n_{luv}}{n}$$

Se existem m loci, a *heterozigosidade* média é

$$\tilde{H} = \frac{1}{m} \sum_{l=1}^{m} \tilde{H}_l$$

Como $\tilde{H}_l$ é a soma de proporções de heterozigotos que são multinomialmente distribuídas, cada $\tilde{H}_l$ é binomialmente distribuída com

$$\mathrm{E}(\tilde{H}_l) = H_l \quad \text{e} \quad \mathrm{Var}(\tilde{H}_l) = \frac{1}{n} H_l(1 - H_l)$$

onde H_l é a proporção de heterozigotos no locus l na população. Uma outra maneira de representar $\tilde{H}_l$ é

$$\tilde{H}_l = \frac{1}{n} \sum_{j=1}^{n} x_{jl}$$

onde

$$x_{jl} = \begin{cases} 1 \text{ se o indivíduo é heterozigoto no locus } l \\ 0 \text{ caso contrário} \end{cases} .$$

Desta forma,

$$\mathrm{E}(x_{jl}) = H_l \ , \qquad \mathrm{E}(x_{jl}^2) = H_l \qquad \text{e} \qquad \mathrm{E}(x_{jl}\, x_{j'l}) = H_l^2,$$

assumindo que os indivíduos são independentes dentro de uma amostra.

Se $\tilde{H} = \frac{1}{m} \sum_l \tilde{H}_l$ é a estimativa da heterozigosidade média dentro de uma população, então,

$$\mathrm{E}(\tilde{H}) = \frac{1}{m} \sum_l H_l = H$$

Para calcular a variância de $\tilde{H}$, precisamos levar em consideração a covariância entre heterozigosidades em diferentes loci, pois estes não são independentes. Assim,

$$\mathrm{Cov}(\tilde{H}_l,\, \tilde{H}_{l'}) \ = \ \mathrm{E}\left[\left(\frac{1}{n} \sum_j x_{jl} \right) \left(\frac{1}{n} \sum_{j'} x_{j'l'} \right) \right] - H_l H_{l'}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\mathrm{E}\left(\sum_j x_{jl}\,x_{jl'} + \sum_j \sum_{j'\neq j} x_{jl}\,x_{j'l'}\right) - H_l H_{l'} \\
&= \frac{1}{n^2}[nH_{ll'} + n(n-1)H_l H_{l'}] - H_l H_{l'} \\
&= \frac{1}{n}(H_{ll'} - H_l H_{l'})
\end{aligned}
$$

e

$$
\begin{aligned}
\mathrm{Var}(\tilde{H}) &= \frac{1}{m^2}\left[\sum_l \mathrm{Var}(\tilde{H}_l) + \sum_l \sum_{l'\neq l} \mathrm{Cov}(\tilde{H}_l,\, \tilde{H}_{l'})\right] \\
&= \frac{1}{nm^2}\left[\sum_l H_l(1-H_l) + \sum_l \sum_{l'\neq l} (H_{ll'} - H_l H_{l'})\right]
\end{aligned}
$$

onde a heterozigosidade em dois loci $H_{ll'} = \mathrm{E}(x_{jl}\,x_{jl'})$ é a probabilidade de que um indivíduo escolhido aleatoriamente seja heterozigoto nos loci l e l'.

A variância amostral de heterozigosidade por locus é

$$
\begin{aligned}
s_H^2 &= \frac{1}{m-1}\sum_l (\tilde{H}_l - \tilde{H})^2 \\
&= \frac{1}{m}\sum_l \tilde{H}_l^2 - \frac{1}{m(m-1)}\sum_l \sum_{l'\neq l} \tilde{H}_l \tilde{H}_{l'}
\end{aligned}
$$

com

$$
\begin{aligned}
\mathrm{E}(s_H^2) &= \frac{1}{m}\sum_l \left[H_l^2 + \frac{1}{n}H_l(1-H_l)\right] \\
&\quad -\frac{1}{m(m-1)}\sum_l \sum_{l'\neq l}\left[H_l H_{l'} + \frac{1}{n}(H_{ll'} - H_l H_{l'})\right]
\end{aligned}
$$

Note que $\mathrm{E}(\tilde{H}_l^2) = H_l^2 + \mathrm{Var}(\tilde{H}_l)$ e $\mathrm{E}(\tilde{H}_l\,\tilde{H}_{l'}) = H_l H_{l'} + \mathrm{Cov}(\tilde{H}_l,\, \tilde{H}_{l'})$.

Para obtermos a variância entre populações, precisamos levar em consideração a dependência entre membros da mesma população causada pela amostragem genética. Seja

$$
M_l = \mathrm{E}(x_{jl}\,x_{j'l}), \quad j \neq j'
$$

onde M_l é a probabilidade de que dois indivíduos na mesma população sejam heterozigotos. Então,

$$\begin{aligned}
\mathrm{E}(\tilde{H}_l) &= H_l \\
\mathrm{Var}(\tilde{H}_l) &= \frac{1}{n^2}\mathrm{E}\left(\sum_j x_{jl}^2 + \sum_j \sum_{j\neq j'} x_{jl}\, x_{j'l}\right) - [\mathrm{E}(\tilde{H}_l)]^2 \\
&= (M_l - H_l^2) + \frac{1}{n}(H_l - M_l)
\end{aligned}$$

Tomando a média sobre todos os m loci

$$\tilde{H} = \frac{1}{m}\sum_l \tilde{H}_l = \frac{1}{nm}\sum_j \sum_l x_{jl}$$

Denote por $M_{ll'}$ a probabilidade de que dois indivíduos escolhidos aleatoriamente da mesma população sejam heterozigotos, um no locus l e outro no locus l':

$$M_{ll'} = \mathrm{E}(x_{jl}\, x_{j'l'})$$

Logo,

$$\begin{aligned}
\mathrm{E}(\tilde{H}) &= \frac{1}{m}\sum_l \mathrm{E}(\tilde{H}_l) = \frac{1}{m}\sum_l H_l \\
\mathrm{Var}(\tilde{H}) &= \frac{1}{m^2 n^2}\mathrm{E}\left(\sum_j \sum_l x_{jl}^2 + \sum_j \sum_{j'\neq j} \sum_l x_{jl}\, x_{j'l}\right. \\
&+ \left.\sum_j \sum_l \sum_{l'\neq l} x_{jl}\, x_{jl'} + \sum_j \sum_{j'\neq j} \sum_l \sum_{l'\neq l} x_{jl}\, x_{j'l'}\right) - H^2 \\
&= \frac{1}{m^2}\left[\sum_l (M_l - H_l^2) + \sum_l \sum_{l'\neq l} (M_{ll'} - H_l H_{l'})\right] \\
&+ \frac{1}{m^2 n}\left[\sum_l (H_l - M_l) + \sum_l \sum_{l'\neq l} (H_{ll'} - M_{ll'})\right] \qquad (1.1)
\end{aligned}$$

Os quatro termos na expressão (1.1) podem ser rearranjados para mostrar como populações, loci e indivíduos contribuem para a variância da heterozigosidade média, colocando estes cálculos num contexto similar ao usado para uma análise de variância. Os

quatro componentes de variância são:

Para o efeito populacional,

$$\sigma_p^2 = \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (M_{ll'} - H_l H_{l'}),$$

Para o efeito do indivíduo dentro da população,

$$\sigma_{i/p}^2 = \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (H_{ll'} - M_{ll'}),$$

Para a interação população por locus,

$$\sigma_{pl}^2 = \frac{1}{m} \sum_l (M_l - H_l^2) - \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (M_{ll'} - H_l H_{l'})$$

e para a interação locus por indivíduo dentro da população,

$$\sigma_{li/p}^2 = \frac{1}{m} \sum_l (H_l - M_l) - \frac{1}{m(m-1)} \sum_l \sum_{l' \neq l} (H_{ll'} - M_{ll'}).$$

Agora, supondo um modelo balanceado, isto é, o número de indivíduos em cada população é o mesmo, adicionamos um índice i na variável indicadora para denotar a população sendo amostrada:

$$x_{ijl} = \begin{cases} 1 \text{ se o indivíduo } j \text{ da população } i \text{ é heterozigoto no locus } l \\ 0 \text{ caso contrário} \end{cases}$$

O quadrado médio (QM) é definido como sendo a soma de quadrados, correspondente a cada fonte de variação, dividido por seus respectivos graus de liberdade (g.l.). O quadro da análise de variância está mostrado na Tabela 1.1 e as esperanças dos quadrados médios são calculadas a partir dos termos:

$$\begin{aligned} \mathrm{E}(x_{ijl}^2) &= \sigma_p^2 + \sigma_{i/p}^2 + H_l^2 + \sigma_{pl}^2 + \sigma_{li/p}^2 \\ \mathrm{E}(x_{ij\cdot}^2) &= m^2\sigma_p^2 + m^2\sigma_{i/p}^2 + \left(\sum_l H_l\right)^2 + m\sigma_{pl}^2 + m\sigma_{li/p}^2 \\ \mathrm{E}(x_{i\cdot l}^2) &= n^2\sigma_p^2 + n\sigma_{i/p}^2 + n^2 H_l^2 + n^2\sigma_{pl}^2 + n\sigma_{li/p}^2 \end{aligned}$$

$$
\begin{aligned}
\mathrm{E}(x_{i..}^2) &= n^2m^2\sigma_p^2 + nm^2\sigma_{i/p}^2 + n^2\left(\sum_l H_l\right)^2 + n^2m\sigma_{pl}^2 + nm\sigma_{li/p}^2 \\
\mathrm{E}(x_{..l}^2) &= rn^2\sigma_p^2 + rn\sigma_{i/p}^2 + r^2n^2H_l^2 + rn^2\sigma_{pl}^2 + rn\sigma_{li/p}^2 \\
\mathrm{E}(x_{...}^2) &= rn^2m^2\sigma_p^2 + rnm^2\sigma_{i/p}^2 + r^2n^2\left(\sum_l H_l\right)^2 + rn^2m\sigma_{pl}^2 + rnm\sigma_{li/p}^2
\end{aligned}
$$

Tabela 1.1: Análise de Variância para Heterozigosidade (Amostra Balanceada)

Fonte de Variação	g.l.	Soma de Quadrados	E(QM)
Populações	$r-1$	$SS_1 - C$	$\sigma_{li/p}^2 + m\sigma_{i/p}^2 + n\sigma_{pl}^2 + mn\sigma_p^2$
Indivíduos dentro de populações	$r(n-1)$	$SS_2 - SS_1$	$\sigma_{li/p}^2 + m\sigma_{i/p}^2$
Loci	$m-1$	$SS_3 - C$	$\sigma_{li/p}^2 + n\sigma_{pl}^2 + L$
Loci por populações	$(r-1)(m-1)$	$SS_4 - SS_1$ $-SS_3 + C$	$\sigma_{li/p}^2 + n\sigma_{lp}^2$
Loci por indiv. dentro de populações	$r(n-1)(m-1)$	$SS_5 - SS_2$ $-SS_4 + SS_1$	$\sigma_{li/p}^2$
Total	$mnr-1$	$SS_5 - C$	

$$
\begin{aligned}
SS_1 &= \frac{1}{nm}\sum_i x_{i..}^2 & SS_2 &= \frac{1}{m}\sum_i\sum_j x_{ij.}^2 & SS_3 &= \frac{1}{rn}\sum_l x_{..l}^2 \\
SS_4 &= \frac{1}{n}\sum_i\sum_l x_{i\cdot l}^2 & SS_5 &= \sum_i\sum_j\sum_l x_{ijl}^2 & C &= \frac{1}{mnr}x_{...}^2 \\
x_{ij.} &= \sum_l x_{ijl} & x_{i\cdot l} &= \sum_j x_{ijl} & x_{i..} &= \sum_j\sum_l x_{ijl} \\
x_{..l} &= \sum_i\sum_j x_{ijl} & x_{...} &= \sum_i\sum_j\sum_l x_{ijl}
\end{aligned}
$$

Portanto,

$$
\mathrm{E}(SS_1) = rnm\sigma_p^2 + rm\sigma_{i/p}^2 + \frac{rn}{m}\left(\sum_l H_l\right)^2 + rn\sigma_{pl}^2 + r\sigma_{li/p}^2
$$

$$
\begin{aligned}
\mathrm{E}(SS_2) &= rnm\sigma_p^2 + rnm\sigma_{i/p}^2 + \frac{rn}{m}\left(\sum_l H_l\right)^2 + rn\sigma_{pl}^2 + rn\sigma_{li/p}^2 \\
\mathrm{E}(SS_3) &= nm\sigma_p^2 + m\sigma_{i/p}^2 + rn\sum_l H_l + nm\sigma_{pl}^2 + m\sigma_{li/p}^2 \\
\mathrm{E}(SS_4) &= rnm\sigma_p^2 + rm\sigma_{i/p}^2 + rn\sum_l H_l + rnm\sigma_{pl}^2 + rm\sigma_{li/p}^2 \\
\mathrm{E}(SS_5) &= rnm\sigma_p^2 + rnm\sigma_{i/p}^2 + rn\sum_l H_l + rnm\sigma_{pl}^2 + rnm\sigma_{li/p}^2 \\
\mathrm{E}(C) &= nm\sigma_p^2 + m\sigma_{i/p}^2 + \frac{rn}{m}\left(\sum_l H_l\right)^2 + n\sigma_{pl}^2 + \sigma_{li/p}^2
\end{aligned}
$$

Estas expressões levam aos valores esperados dos quadrados médios na Tabela 1.1 e

$$L = \frac{1}{m-1}\sum_l (H_l - H)^2, \;\; m > 1$$

Weir (1990) não deixou claro como seria feita a análise de variância e não especificou a estatística de teste para possíveis hipóteses de interesse.

1.3.2 Diversidade Genética

Uma medida de variação alternativa, muitas vezes chamada de heterozigosidade média, mas mais apropriadamente conhecida como *Diversidade Genética*, é formada da soma de quadrados das proporções alélicas. É uma medida mais apropriada de variabilidade para populações endocruzadas, onde há poucos heterozigotos, mas pode haver muitos tipos diferentes de homozigotos. Para populações de cruzamento aleatório, seu valor será bem perto da heterozigosidade.

Seja p_{lu} a proporção do u-ésimo alelo no l-ésimo locus, a diversidade genética neste locus é

$$D_l = 1 - \sum_{u=1}^{U} p_{lu}^2 \tag{1.2}$$

e a média sobre m loci,

$$D = 1 - \frac{1}{m}\sum_{l=1}^{m}\sum_{u=1}^{U} p_{lu}^2. \tag{1.3}$$

Se $\mathbf{N}_l = (N_{l1},\ N_{l2},\ \ldots,\ N_{lU})$ é um vetor aleatório que representa a quantidade de cada alelo no l-ésimo locus, $\mathbf{N}_l$ tem distribuição Multinomial $(2n;\ p_{l1},\ \ldots,\ p_{lU})$, quando temos n indivíduos diplóides na amostra. As proporções alélicas amostrais nos dão estimadores de máxima verossimilhança (EMV) da diversidade genética:

$$\hat{D} = 1 - \frac{1}{m}\sum_{l}\sum_{u}\hat{p}_{lu}^2; \quad \text{onde } \hat{p}_{lu} = \frac{N_{lu}}{2n} \tag{1.4}$$

$$\begin{aligned} E(\hat{D}_l) &= 1 - \sum_u p_{lu}^2 - \frac{1}{2n}\sum_u p_{lu}(1-p_{lu})(1+f) \\ &= \left(1 - \frac{1+f}{2n}\right) D_l \end{aligned}$$

e

$$E(\hat{D}) = \left(1 - \frac{1+f}{2n}\right) D;$$

onde f é uma medida de endocruzamento que está relacionada com o equilíbrio de Hardy-Weinberg (H-W). Quando $f = 0$, a população está em equilíbrio de H-W. Note que há um pequeno vício de $(2n-1)/(2n)$ para populações sem endocruzamento ($f = 0$), mas um vício ainda maior caso contrário. A presença do termo f indica que o valor esperado de $\hat{D}$ depende tanto das proporções genotípicas como das alélicas.

A variância requer a soma das variâncias e covariâncias dos quadrados das freqüências alélicas no mesmo locus. Usando a aproximação de Fisher (Weir, 1990), temos

$$\begin{aligned} \mathrm{Var}(\hat{p}_{lu}^2) &= \frac{1}{n} 2p_{lu}^3(1-p_{lu})(1+f) \\ \mathrm{Cov}(\hat{p}_{lu}^2, \hat{p}_{lv}^2) &= -\frac{1}{n} 2p_{lu}^2 p_{lv}^2(1+f),\ v \neq u \end{aligned}$$

Logo,

$$\begin{aligned} \mathrm{Var}(\hat{D}_l) &= \sum_u \mathrm{Var}(\hat{p}_{lu}^2) + \sum_u \sum_{v \neq u} \mathrm{Cov}(\hat{p}_{lu}^2, \hat{p}_{lv}^2) \\ &= \frac{2(1+f)}{n}\left[\sum_u p_{lu}^3 - \left(\sum_u p_{lu}^2\right)^2\right] \end{aligned}$$

Para múltiplos loci e maiores detalhes ver Weir (1990), capítulo 4.

1.3.3 CATANOVA em Amostras Balanceadas

Para dados categorizados, a média aritmética não é uma medida bem representativa, portanto medidas de variação, tais como a variância, que são bem aplicadas às variáveis contínuas, não devem ser utilizadas. Segundo Pinheiro *et al.* (2000), Gini (1912) encontrou uma maneira alternativa de caracterizar esta variação e desenvolveu uma medida de variação para dados categorizados:

Sejam $X_1,\ X_2,\ \ldots,\ X_N$ variáveis aleatórias independentes. A variância de X ou a soma de quadrados pode ser expressa como

$$SS \ = \ \sum_{j=1}^{N}(X_j - \bar{X})^2 = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i - X_j)^2 = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 \tag{1.5}$$

onde $\bar{X} = \sum_{j=1}^{N} \frac{X_j}{N}$ e $d_{ij} = X_i - X_j$.

Neste caso, cada X_j pertence a uma das C categorias possíveis e portanto $d(X_i, X_j) = d_{ij}$ é definida como

$$d_{ij} = \begin{cases} 1 \text{ se } X_i \text{ e } X_j \text{ pertencem a categorias diferentes} \\ 0 \text{ se } X_i \text{ e } X_j \text{ pertencem a mesma categoria} \end{cases} \tag{1.6}$$

Definição 1.1. A variação entre as respostas categorizadas $X_1,\ \ldots,\ X_N$ é definida por

$$D_N = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \tag{1.7}$$

onde d_{ij} é definida em (1.6).

Como cada resposta assume uma e somente uma de C categorias possíveis, os dados podem ser resumidos no vetor $\mathbf{\Phi} = (n_1, \ldots n_C)$, onde n_i é o número de respostas na i-ésima categoria ($i = 1, \ldots C$), tal que $\sum_{i=1}^{C} n_i = N$. Então a variação nas respostas é expressa como

$$\begin{aligned} D_N \ &= \ \frac{1}{2N}\sum_{i\neq j} n_i n_j = \frac{1}{2N}\left[N^2 - \sum_{i=1}^{C} n_i^2\right] \\ &= \ \frac{N}{2} - \frac{1}{2N}\sum_{i=1}^{C} n_i^2 = \frac{N}{2}\left\{1 - \sum_{i=1}^{C}\left(\frac{n_i}{N}\right)^2\right\} \end{aligned} \tag{1.8}$$

Se $\mathbf{p} = (p_1, \ldots, p_C)$ é o vetor que representa as probabilidades de X_j pertencer a cada uma destas C categorias, o Índice de diversidade ecológica de Simpson é definido como

$$I_s(\mathbf{p}) = 1 - \mathbf{p}'\mathbf{p} = 1 - \sum_{i=1}^{C} p_i^2 \tag{1.9}$$

e seu estimador correspondente é

$$\hat{I}_s(\mathbf{p}) = 1 - \hat{\mathbf{p}}'\hat{\mathbf{p}} = 1 - \sum_{i=1}^{C} \hat{p}_i^2 \tag{1.10}$$

onde $\hat{p}_i = n_i/N$, $i = 1, \ldots, C$ é a proporção amostral. Então,

$$D_N = \frac{N}{2}\hat{I}_s(\mathbf{p})$$

A definição (1.9) é motivada pelas propriedades:

1. A variação entre as N respostas categorizadas é mínima se, e somente se, todas as respostas pertencem a mesma categoria, isto é, $p_i = 1$, para algum i, $i = 1, \ldots, C$.

2. A variação entre as N respostas categorizadas é máxima quando se tem a mesma quantidade de respostas pertencentes a cada uma das C categorias, isto é, $p_i = 1/C$, para todo i, $i = 1, \ldots, C$.

Motivados por essa medida de diversidade para dados categorizados, Pinheiro *et al.* (2000) desenvolveram uma análise de variância para dados categorizados (CATANOVA) para amostras balanceadas. Se $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \ldots, X_{iK}^g)'$ é um vetor aleatório representando a seqüência i do grupo g; $i = 1, \ldots, N$, $k = 1, \ldots, K$, $g = 1, \ldots, G$, então, X_{ik}^g representa a posição k da seqüência i do grupo g. X_{ik}^g é uma variável categorizada que assume C categorias. No caso de comparações feitas em nível de nucleotídeos, $X_{ik}^g \in \{\mathbf{A},\mathbf{C},\mathbf{T},\mathbf{G}\}$, e portanto existem 4 categorias.

Inicialmente, assume-se que há somente uma posição em cada seqüência. Os dados estão resumidos na Tabela 1.2

Agora d_{ij} é definida como

$$d_{ij} = \begin{cases} 1 \text{ se } X_i^g \neq X_j^g \\ 0 \text{ se } X_i^g = X_j^g \end{cases}$$

Tabela 1.2: Tabela de Contingência (1 posição)

	Grupo				
Seqüência	1	2	3	...	G
1	x_1^1	x_1^2	x_1^3	...	x_1^G
2	x_2^1	x_2^2	x_2^3	...	x_2^G
3	x_3^1	x_3^2	x_3^3	...	x_3^G
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	x_N^1	x_N^2	x_N^3	...	x_N^G

O número total de respostas é

$$NG = \sum_{g=1}^{G} n_{\cdot g} = \sum_{c=1}^{C} n_{c\cdot} = \sum_{c=1}^{C}\sum_{g=1}^{G} n_{cg}$$

onde n_{cg} é o número de respostas na categoria c para o grupo g e $N = n_{\cdot g} = \sum_{c=1}^{c} n_{cg}$ é o número de respostas para o grupo g, que, neste caso, é o número de seqüências em cada grupo. O Índice Total de Simpson (TSI) é

$$TSI = 1 - \sum_{c=1}^{C} \left(\frac{n_{c\cdot}}{NG}\right)^2 \tag{1.11}$$

A dispersão dentro do grupo g (entre $x_1^g, x_2^g, \ldots, x_N^g$) é

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cg}}{n_{\cdot g}}\right)^2 \tag{1.12}$$

Então, o Índice de Simpson Intra-Grupos é obtido pela média de (1.12) em relação a todos os grupos:

$$WSI = \frac{1}{G}\sum_{g=1}^{G}\left[1 - \sum_{c=1}^{C}\left(\frac{n_{cg}}{n_{\cdot g}}\right)^2\right] = 1 - G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg}}{NG}\right)^2 \tag{1.13}$$

O Índice de Simpson Entre-Grupos é

$$BSI = TSI - WSI = G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg}}{NG}\right)^2 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot}}{NG}\right)^2 \tag{1.14}$$

Assumindo que existem K posições ao longo de cada seqüência temos
$\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \ldots, X_{iK}^g)'$; $\mathbf{X}^g = (\mathbf{X}_1^g, \mathbf{X}_2^g, \ldots, \mathbf{X}_N^g)'$ e os dados podem ser resumidos como mostra a Tabela 1.3. O número total de respostas é

$$NGK = \sum_{g=1}^{G} n_{\cdot g \cdot} = \sum_{c=1}^{C} n_{c \cdot \cdot} = \sum_{k=1}^{K} n_{\cdot \cdot k} = \sum_{c=1}^{C} \sum_{g=1}^{G} \sum_{k=1}^{K} n_{cgk}$$

O interesse aqui é testar a homogeneidade entre os grupos: a hipótese nula é de que $p_{cgk} = p_{ck}$, onde p_{cgk} é a probabilidade populacional da posição k do grupo g pertencer a categoria c. Um possível argumento é que este é o clássico teste χ^2 de Pearson, mas note que a clássica estatística χ^2 de Pearson para a Tabela 1.3 é

$$\begin{aligned} \chi_P^2 &= \sum_{g=1}^{G} \sum_{c=1}^{C} \sum_{k=1}^{K} \frac{\left(n_{cgk} - \frac{n_{\cdot gk} n_{c \cdot k}}{n_{\cdot \cdot k}} \right)^2}{(n_{\cdot gk} n_{c \cdot k})/n_{\cdot \cdot k}} \\ &= \sum_{g=1}^{G} \sum_{c=1}^{C} \sum_{k=1}^{K} G \frac{\left(n_{cgk} - \frac{n_{c \cdot k}}{G} \right)^2}{n_{c \cdot k}} \end{aligned} \tag{1.15}$$

com K(G-1)(C-1) graus de liberdade. A distribuição limite da estatística χ^2 de Pearson é uma boa aproximação somente quando as freqüências n_{cgk}'s são todas grandes (pelo menos 5). Analizando seqüências genômicas verifica-se que estas condições não são necessariamente satisfeitas. Numa única posição a distribuição de freqüências exibe pouco polimorfismo com freqüências pequenas em algumas caselas. A distribuição exata, sob a hipótese nula, é difícil de ser implementada para pequenos valores de N quando G ou K não é pequeno.

A variação dentro do grupo g na k-ésima posição é

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cgk}}{n_{\cdot gk}} \right)^2 = 1 - \sum_{c=1}^{C} \left(\frac{n_{cgk}}{N} \right)^2; \quad \text{onde } \; n_{\cdot gk} = N.$$

A variação dentro do g-ésimo grupo é

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cg\cdot}}{n_{\cdot g \cdot}} \right)^2 = 1 - \sum_{c=1}^{C} \left(\frac{n_{cg\cdot}}{NK} \right)^2, \quad \text{sendo } \; n_{\cdot g \cdot} = NK.$$

Tabela 1.3: Tabela de Contingência (K posições)

Grupo	Posição	1	2	...	C	Total
1	1	n_{111}	n_{211}	...	n_{C11}	$n_{\cdot 11} = N$
1	2	n_{112}	n_{212}	...	n_{C12}	$n_{\cdot 12} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	K	n_{11K}	n_{21K}	...	n_{C1K}	$n_{\cdot 1K} = N$
Total		$n_{11\cdot}$	$n_{21\cdot}$	...	$n_{C1\cdot}$	$n_{\cdot 1\cdot} = NK$
2	1	n_{121}	n_{221}	...	n_{C21}	$n_{\cdot 21} = N$
2	2	n_{122}	n_{222}	...	n_{C22}	$n_{\cdot 22} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
2	K	n_{12K}	n_{22K}	...	n_{C2K}	$n_{\cdot 2K} = N$
Total		$n_{12\cdot}$	$n_{22\cdot}$	...	$n_{C2\cdot}$	$n_{\cdot 2\cdot} = NK$
...	...	...	...	...	...	...
G	1	n_{1G1}	n_{2G1}	...	n_{CG1}	$n_{\cdot G1} = N$
G	2	n_{1G2}	n_{2G2}	...	n_{CG2}	$n_{\cdot G2} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	K	n_{1GK}	n_{2GK}	...	n_{CGK}	$n_{\cdot GK} = N$
Total		$n_{1G\cdot}$	$n_{2G\cdot}$	...	$n_{CG\cdot}$	$n_{\cdot G\cdot} = KN$
TOTAL		$n_{1\cdot\cdot}$	$n_{2\cdot\cdot}$	...	$n_{C\cdot\cdot}$	$n_{\cdot\cdot\cdot} = NGK$

As medidas de dispersão são então,

$$WSI = \frac{1}{G}\sum_{g=1}^{G}\left[1-\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NK}\right)^2\right] = 1 - G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NGK}\right)^2 \tag{1.16}$$

$$TSI = 1-\sum_{c=1}^{C}\left(\frac{n_{c\cdot\cdot}}{NGK}\right)^2 \tag{1.17}$$

e

$$BSI = TSI - WSI = G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NGK}\right)^2 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot\cdot}}{NGK}\right)^2. \tag{1.18}$$

A hipótese de homogeneidade entre os grupos é $H_0 : p_{cgk} = p_{ck}$, onde p_{cgk} é a probabilidade populacional de pertencer à categoria c na posição k do grupo g. Para testarmos esta hipótese foi proposta a estatística,

$$F_1 = N\frac{BSI}{WSI} \approx N\frac{BSI}{\theta_3^o},$$

onde $\theta_3^o = 1 - \frac{1}{K^2}\sum_{c=1}^{C} p_{c\cdot}^2$.

BSI pode ser escrito de forma matricial,

$$BSI = \mathbf{V}'\mathbf{B}\mathbf{V},$$

onde

$$\mathbf{B} = \frac{1}{G(NK)^2}\left[\left(\mathbf{I}_G \otimes \mathbf{U}_K - \frac{1}{G}\mathbf{U}_{KG}\right)\otimes \mathbf{I}_C\right] = \frac{1}{G(NK)^2}\mathbf{B}^{\diamond}, \tag{1.19}$$

sendo $\mathbf{I}_G$ a matriz identidade de tamanho G, $\mathbf{U}_K$ uma matriz quadrada $K \times K$ de 1's e $\otimes$ o operador de produto de Kronecker (Definição 3.1);

$$\mathbf{V} = (\mathbf{V}_1,\ \mathbf{V}_2,\ \ldots,\ \mathbf{V}_G)',$$

sendo $\mathbf{V}_g = (\mathbf{V}_{g1},\ \mathbf{V}_{g2},\ \ldots,\ \mathbf{V}_{gK})$ e $\mathbf{V}_{gk} = (N_{1gk},\ N_{2gk}\ \ldots,\ N_{Cgk})$, $g = 1,\ \ldots,\ G$.

Note que $\mathbf{V}_{gk}$ tem distribuição Multinomial $(N;\ p_{1gk},\ p_{2gk}\ \ldots,\ p_{Cgk})$ e, como assumimos independência entre os grupos e entre as posições, $\mathbf{V}$ tem distribuição assintótica

normal. Portanto F_1 é uma forma quadrática de variáveis aleatórias normais e tem como distribuição assintótica uma combinação linear de variáveis aleatórias com distribuição χ^2_1 (Pinheiro *et al.*, 2000), isto é,

$$F_1 \sim \frac{N}{\theta_3^o} \sum_{i=1}^{CGK} \lambda_i (\chi^2_1)_i$$

onde $\{\lambda_i : i = 1, \ \ldots, \ CGK\}$ é o conjunto de autovalores de $\dfrac{1}{G(NK)^2}\mathbf{B}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond}$ e $N\boldsymbol{\Sigma}_0^{\diamond}$ é a matriz de covariâncias de $\mathbf{V}$ (Cov($\mathbf{V}$)) sob a hipótese nula de homogeneidade entre os grupos,

$$\boldsymbol{\Sigma}_0^{\diamond} = \boldsymbol{\Sigma}_{01}^{\diamond} \oplus \boldsymbol{\Sigma}_{02}^{\diamond} ... \oplus \boldsymbol{\Sigma}_{0K}^{\diamond},$$

onde $\oplus$ é o operador de soma direta, ou seja, se $\mathbf{D} = \mathbf{E} \oplus \mathbf{F}$ então, $\mathbf{D} = \begin{pmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{pmatrix}$, e $\boldsymbol{\Sigma}_{0k}^{\diamond} = \mathbf{D}_k - \boldsymbol{\mu}_{\diamond k}\boldsymbol{\mu}'_{\diamond k}$, com $\mathbf{D}_k$ sendo uma matriz diagonal $C \times C$ cujos elementos da diagonal são os elementos do vetor $\boldsymbol{\mu}_{\diamond gk} = (p_{1k}, \ \ldots \ p_{Ck})'$.

Capítulo 2

Análise de Variância para Dados Binários em Amostras Não Balanceadas

O interesse do nosso estudo é a comparação de grupos de diferentes tamanhos, ou seja, testar a hipótese de homogeneidade entre grupos em amostras não balanceadas, quando a variável resposta de interesse é categorizada. Em análise de variância clássica esta comparação é feita quando a variável resposta é contínua. Nosso objetivo, neste capítulo, é desenvolver uma análise de variância para dados categorizados, em particular quando a variável resposta é binária, em amostras não balanceadas. Por exemplo, uma das técnicas utilizadas para detectar polimorfismos genéticos, para que se possa comparar grupos e assim verificar se há variabilidade genética entre eles, é a classe de marcadores moleculares RAPD, onde o polimorfismo é detectado através de uma resposta binária (fenótipo dominante ou recessivo).

No Capítulo 1 vimos que Weir (1990), propôs uma medida de diversidade em termos da proporção de heterozigotos, a *heterozigosidade* observada e o quadro da análise de variância foi mostrado para dados binários em amostras balanceadas.

Em nosso caso, os grupos são de diferentes tamanhos. Estendendo o trabalho de Weir (1990), uma análise de variância foi obtida para dados binários em amostras não balanceadas (Seção 2.1). Utilizando uma medida de diversidade similar, em termos das proporções de fenótipos dominantes, foi desenvolvida uma análise de variância para dados binários considerando que os *loci* são amostrados aleatoriamente, como ocorre com os marcadores moleculares da classe RAPD. Pela natureza do marcador RAPD, não faz sentido biológico avaliar a contribuição dos *loci*. Sendo assim, a análise de variância da

Seção 2.2 não considera efeito de *loci* (este efeito está incorporado no termo residual). Na Seção 2.3 foi proposta uma estatística para testarmos a hipótese nula de homogeneidade entre os grupos, assumindo independência entre os *loci*, entre os indivíduos e entre os grupos. Nesta Seção também foi obtida a distribuição assintótica da estatística de teste. Um estudo sobre o poder do teste foi desenvolvido na Seção 2.4.

2.1 Quadro da Análise de Variância

Como foi visto no Capítulo 1, Weir (1990) apresentou o quadro da análise de variância para comparar G grupos com respostas binárias (indivíduo heterozigoto ou não), onde cada grupo possui N indivíduos (amostras balanceadas) e a resposta de cada indivíduo tem K *loci*, ou seja, para cada indivíduo temos como resposta um vetor aleatório $\mathbf{X}_i$, $i = 1, \ldots, N$, de dimensão K, onde cada elemento deste vetor tem o valor 1 (o indivíduo é heterozigoto no *locus* k) ou 0 (o indivíduo não é heterozigoto no *locus* k). Quando as amostras são não balanceadas, neste caso K loci são observados em N_g indivíduos, $g = 1, ..., G$, obtidos em cada um dos G grupos (populações). A variável indicadora X_{gik} é definida como

$$X_{gik} = \begin{cases} 1 \text{ se o indivíduo } i \text{ da população } g \text{ é heterozigoto no locus } k \\ 0 \text{ caso contrário} \end{cases}$$

A Tabela 2.1 mostra o quadro da análise de variância para este caso (delineamento com fatores cruzados e aninhados em dados binários de amostras não balanceadas), considerando que os loci, os indivíduos e as populações são amostrados independentemente e que os efeitos são independentes.

Na Tabela $E(QM)$ é o valor esperado do quadrado médio correspondente à sua fonte de variação, sendo o quadrado médio, a soma de quadrados dividida por seu respectivo grau de liberdade (g.l.) e o número total de indivíduos $N_T = \sum_{g=1}^{G} N_g$.

Agora faremos uma analogia ao quadro da análise de variância para a heterozigosidade com os dados do tipo marcadores RAPD. Neste caso,

$$X_{gik} = \begin{cases} 1 \text{ se o indivíduo } i \text{ da população } g \text{ tem o alelo dominante} \\ \text{no locus } k \text{ (banda presente)} \\ \\ 0 \text{ caso contrário (banda ausente)} \end{cases}$$

Tabela 2.1: Análise de Variância para um Delineamento com Fatores Cruzados e Aninhados em Dados Binários de Amostras Não Balanceadas

Fonte de Variação	g.l.	Soma de Quadrados	E(QM)
Populações	$G-1$	$SS_1 - C$	$E\left(\frac{SS_1-C}{G-1}\right)$
Indivíduos dentro das populações	$\sum_g (N_g - 1)$	$SS_2 - SS_1$	$E\left(\frac{SS_2-SS_1}{N_T-G}\right)$
Loci	$K-1$	$SS_3 - C$	$E\left(\frac{SS_3-C}{K-1}\right)$
Loci por populações	$(G-1)(K-1)$	$SS_4 - SS_1$ $-SS_3 + C$	$E\left(\frac{SS_4-SS_1}{(G-1)(K-1)}\right)$
Loci por indiv. dentro de populações	$(K-1)\sum_g (N_g - 1)$	$SS_5 - SS_2$ $-SS_4 + SS_1$	$E\left(\frac{SS_5-SS_2}{(K-1)(N_T-G)}\right)$
Total	$KN_T - 1$	$SS_5 - C$	

$$SS_1 = \frac{1}{K}\sum_{g=1}^{G}\frac{X_{g\cdot\cdot}^2}{N_g} \qquad SS_2 = \frac{1}{K}\sum_{g=1}^{G}\sum_{i=1}^{N_g} X_{gi\cdot}^2 \qquad SS_3 = \frac{1}{N_T}\sum_{k=1}^{K} X_{\cdot\cdot k}^2$$

$$SS_4 = \sum_{g=1}^{G}\sum_{k=1}^{K}\frac{X_{g\cdot k}^2}{N_g} \qquad SS_5 = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{k=1}^{K} X_{gik}^2 \qquad C = \frac{1}{KN_T}X_{\cdots}^2$$

$$X_{gi\cdot} = \sum_{k=1}^{K} X_{gik} \qquad X_{g\cdot l} = \sum_{i=1}^{N_g} X_{gik} \qquad X_{g\cdot\cdot} = \sum_{i=1}^{N_g}\sum_{k=1}^{K} X_{gik}$$

$$X_{\cdot\cdot l} = \sum_{g=1}^{G}\sum_{i=1}^{N_g} X_{gik} \qquad X_{\cdots} = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{k=1}^{K} X_{gik}$$

Através da técnica de obter marcadores RAPD, descrita no capítulo anterior, pode-se considerar que os *loci* são amostrados independentemente nos indivíduos e nas populações e, como também estes são amostrados de maneira aleatória sem a garantia de que eles são os mesmos para todos os indivíduos, decidimos incorporar o efeito de *locus* no termo residual.

A Tabela 2.2 mostra o quadro de análise de variância para dados de RAPD.

Tabela 2.2: Análise de Variância para Dados de RAPD em Amostras Não Balanceadas

Fonte	g.l.	Soma de Quadrados	E(QM)
Populações	$G-1$	$SS_1 - C = SQP$	$E\left(\frac{SQP}{G-1}\right)$
Indivíduos dentro das populações	$\sum_g (N_g - 1)$	$SS_2 - SS_1 = SQI$	$E\left(\frac{SQI}{N_T - G}\right)$
Resíduo	$(K-1)N_T$	$SS_5 - SS_2 = SQR$	$E\left(\frac{SQR}{(K-1)N_T}\right)$
Total	$KN_T - 1$	$SS_5 - C = SQT$	

2.2 Distribuições Assintóticas e Estatística do Teste

O interesse agora é desenvolver uma estatística para testarmos a hipótese de homogeneidade entre os grupos. Nesta seção vamos nos concentrar em obter as distribuições assintóticas de algumas estatísticas que serão úteis para o desenvolvimento da distribuição da estatística do teste.

Os vetores aleatórios de dados binários em estudo são como as respostas obtidas pelos marcadores RAPD. Considerando que a técnica de RAPD amostra os loci independentemente e assumindo que as respostas são independentes para cada grupo e cada indivíduo, a resposta X_{gik} segue uma distribuição de Bernoulli com os indivíduos sendo identicamente distribuídos, isto é,

$$P(X_{gik} = x_{gik}) = p_{gk}^{x_{gik}}(1 - p_{gk})^{(1-x_{gik})}\mathbf{I}_{\{0,1\}}(x_{gik}),$$

onde p_{gk} é a probabilidade de um indivíduo da população g ter o alelo dominante no locus k, $i = 1, \ldots, N_g$; g $= 1, \ldots, G$; $k = 1, \ldots, K$.

Logo,

$$\mathrm{E}(X_{gik}) = p_{gk} \quad \text{e} \quad \mathrm{Var}(X_{gik}) = p_{gk}(1 - p_{gk}).$$

Para marcadores RAPD temos que $\bar{X}_{gi\cdot}$ representa a proporção, em K loci, de fenótipos dominantes no indivíduo i do grupo g, $\bar{X}_{g\cdot\cdot}$ representa a proporção de fenótipos dominantes no grupo g e $\bar{X}_{\cdots}$ a média geral de fenótipos dominantes:

$$\bar{X}_{gi\cdot} = \frac{X_{gi\cdot}}{K}, \qquad \bar{X}_{g\cdot\cdot} = \frac{X_{g\cdot\cdot}}{KN_g} \qquad \text{e} \qquad \bar{X}_{\cdots} = \frac{X_{\cdots}}{KN_T} = \frac{\sum_g N_g \bar{X}_{g\cdot\cdot}}{N_T}. \tag{2.1}$$

Como nosso objetivo é testar se existe diferença (variabilidade genética) entre os grupos, consideramos $H_0 : p_{gk} = p_k$ para todo g. Para testarmos esta hipótese, observando o quadro da ANOVA (Tabela 2.2), propomos a estatística $F = QMP/QMI$, onde

$$QMP = \frac{SQP}{G-1} \quad \text{e} \quad QMI = \frac{SQI}{N_T - G}, \tag{2.2}$$

sendo SQP a soma de quadrados referente ao efeito populacional, que mede a variabilidade entre populações:

$$\mathrm{SQP} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{k=1}^{K} \left(\bar{X}_{g\cdot\cdot} - \bar{X}_{\cdots}\right)^2 = \sum_{g=1}^{G} KN_g \left(\bar{X}_{g\cdot\cdot} - \bar{X}_{\cdots}\right)^2, \tag{2.3}$$

SQI a soma de quadrados referente ao efeito do indivíduo dentro da população, que mede a variabilidade entre indivíduos dentro do grupo (população):

$$\mathrm{SQI} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \sum_{k=1}^{K} \left(\bar{X}_{gi\cdot} - \bar{X}_{g\cdot\cdot}\right)^2 = \sum_{g=1}^{G} \sum_{i=1}^{N_g} K \left(\bar{X}_{gi\cdot} - \bar{X}_{g\cdot\cdot}\right)^2, \tag{2.4}$$

$G-1$ e $N_T - G$ são, respectivamente, os graus de liberdade referentes às populações e aos indivíduos dentro das populacões.

2.2.1 Distribuição Assintótica de SQP

Para obtermos a distribuição assintótica de F iremos, inicialmente, encontrar a distribuição assintótica da soma de quadrados referente ao efeito da população, então de (2.3) temos,

$$\text{SQP} = \sum_{g=1}^{G} KN_g \left(\bar{X}_{g\cdot\cdot} - \bar{X}_{\cdot\cdot\cdot}\right)^2$$

Para obtermos esta distribuição, inicialmente temos que encontrar a distribuição assintótica do número médio de fenótipos dominantes para o g-ésimo grupo ($\bar{X}_{g\cdot\cdot}$).

Temos que, o número total de fenótipos dominantes no grupo g, é

$$X_{g\cdot\cdot} = \sum_{i=1}^{N_g} \sum_{k=1}^{K} X_{gik}.$$

Assim,

$$\text{E}(X_{g\cdot\cdot}) = N_g \sum_k p_{gk} \qquad \text{e} \qquad \text{Var}(X_{g\cdot\cdot}) = N_g \sum_k p_{gk}(1 - p_{gk}).$$

Como X_{gik} são variáveis aleatórias independentes mas não identicamente distribuídas, utilizaremos o Teorema Central do Limite de Liapunov para variáveis aleatórias independentes:

Teorema 2.1. (Teorema Central do Limite de Liapunov).
Sejam X_1, X_2, ... variáveis aleatórias independentes tais que $\text{E}(X_n) = \mu_n$ e $\text{Var}(X_n) = \sigma_n^2 < \infty$ com pelo menos um $\sigma_n^2 > 0$. Sejam $S_n = \sum_{i=1}^{n} X_i$ e $s_n^2 = \text{Var}(S_n) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$. Se $\exists\ \delta > 0$ tal que

$$\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} \text{E}|X_k - \mu_k|^{2+\delta} \to 0 \quad \text{quando } n \to \infty,$$

então

$$\frac{S_n - \text{E}(S_n)}{s_n} \xrightarrow{D} \text{N}(0,1) \qquad \text{(James, 1996).}$$

■

No nosso caso, para uma dada população g, $X_{g11}, \ldots, X_{g1K}, \ldots, X_{gN_g1} \ldots, X_{gN_gK}$, $g = 1, \ldots, G$, são variáveis aleatórias independentes tais que $\text{E}(X_{gik}) = p_{gk}$, $\text{Var}(X_{gik}) = p_{gk}(1 - p_{gk})$, $S_n = X_{g\cdot\cdot}$ e $s_n^2 = \text{Var}(X_{g\cdot\cdot})$ onde $n = KN_g$. Portanto, gostaríamos de verificar se a condição de Liapunov é satisfeita:

Para $\delta = 1$ e $0 < p_{gk} < 1$, temos

$$\frac{1}{s_n^3}\sum_{i=1}^{N_g}\sum_{k=1}^{K}\mathrm{E}|X_{gik}-\mu_k|^3 = \frac{1}{s_n^3}\sum_{i=1}^{N_g}\sum_{k=1}^{K}\mathrm{E}|X_{gik}-p_{gk}|^3$$

$$= \frac{1}{\left(\sqrt{\sum_{i=1}^{N_g}\sum_{k=1}^{K}p_{gk}(1-p_{gk})}\right)^3}\sum_{i=1}^{N_g}\sum_{k=1}^{K}[p_{gk}^3(1-p_{gk})+(1-p_{gk})^3p_{gk}]$$

$$= \frac{1}{\left(\sqrt{N_g\sum_{k=1}^{K}p_{gk}(1-p_{gk})}\right)^3}N_g\sum_{k=1}^{K}[p_{gk}^3(1-p_{gk})+(1-p_{gk})^3p_{gk}]$$

$$= \frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{N_g\sum_k p_{gk}(1-p_{gk})}\right)}$$

Note que

$$\frac{1}{2} \leq 1-2p_{gk}+2p_{gk}^2 < 1 \Rightarrow \sum_{k=1}^{K}p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2) < \sum_{k=1}^{K}p_{gk}(1-p_{gk})$$

Portanto

$$\frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{N_g\sum_k p_{gk}(1-p_{gk})}\right)} < \frac{\sum_k p_{gk}(1-p_{gk})}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{N_g\sum_k p_{gk}(1-p_{gk})}\right)}$$
$$= \frac{1}{\sqrt{N_g\sum_{k=1}^{K}p_{gk}(1-p_{gk})}}$$

Se $h^\star = \min_k\{p_{gk}(1-p_{gk})\}$, então $N_g\sum_k p_{gk}(1-p_{gk}) \geq KN_gh^\star$ e portanto

$$\frac{1}{\sqrt{\sum_{k=1}^{K}N_gp_{gk}(1-p_{gk})}} \leq \frac{1}{\sqrt{KN_gh^\star}} \to 0 \text{ quando } KN_g \to \infty$$

Como a condição de Liapunov é satisfeita temos que, para K ou N_g suficientemente grandes,

$$X_{g\cdot\cdot} \approx \mathrm{N}\left(N_g \sum_k p_{gk},\ N_g \sum_k p_{gk}(1-p_{gk})\right)$$

Logo,

$$\bar{X}_{g\cdot\cdot} \approx \mathrm{N}\left(\frac{\sum_k p_{gk}}{K}, \frac{\sum_k p_{gk}(1-p_{gk})}{K^2 N_g}\right) \tag{2.5}$$

para K ou N_g suficientemente grandes.

Uma vez que a distribuição assintótica de $\bar{X}_{g\cdot\cdot}$ é normal podemos escrever SQP como uma forma quadrática de variáveis normais.

Note que SQP pode ser escrita de forma matricial como

$$\mathrm{SQP} = \mathbf{H}'\mathbf{F}\mathbf{H} \tag{2.6}$$

onde $\mathbf{H} = (\bar{X}_{1\cdot\cdot},\ \bar{X}_{2\cdot\cdot},\ \ldots,\ \bar{X}_{G\cdot\cdot})'$ e $\mathbf{F} = K\mathbf{F}^\star$, sendo $\mathbf{F}^\star$ uma matriz simétrica $G \times G$ onde os elementos são

$$f^\star(g,g) = N_g\left(1 - \frac{N_g}{N_T}\right) \quad \text{e} \quad f^\star(g,g') = -\frac{N_g N_{g'}}{N_T}, \quad g' \neq g \tag{2.7}$$

Portanto, assintoticamente

$$\mathbf{H} \approx \mathrm{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \tag{2.8}$$

onde

$$\boldsymbol{\mu}_1 = \frac{1}{K}\left(\sum_k p_{1k},\ \sum_k p_{2k},\ \ldots,\ \sum_k p_{Gk}\right)' \quad \text{e} \quad \boldsymbol{\Sigma}_1 = \frac{1}{K^2}\boldsymbol{\Sigma}_1^\star, \tag{2.9}$$

$\boldsymbol{\Sigma}_1^\star$ é uma matriz diagonal $G \times G$ cujos elementos da diagonal são

$$\sigma^\star(g,g) = \frac{\sum_k p_{gk}(1-p_{gk})}{N_g}.$$

Sob a hipótese de homogeneidade entre os grupos, $p_{gk} = p_k$ para todo g, então, de (2.8) temos sob H_0, assintoticamente, que,

$$\mathbf{H} \approx \mathrm{N}\left(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}\right) \tag{2.10}$$

onde $\boldsymbol{\mu}_{01} = \dfrac{\sum_k p_k}{K}\mathbf{u_G}$, sendo $\mathbf{u_G}$ um vetor coluna de 1's de tamanho G e

$$\boldsymbol{\Sigma}_{01} = \frac{\sum_k p_k(1-p_k)}{K^2}\boldsymbol{\Sigma}^{\star}_{01}, \tag{2.11}$$

sendo $\boldsymbol{\Sigma}^{\star}_{01}$ uma matriz diagonal $G \times G$ cujos elementos da diagonal são $\sigma^{\star}_0(g,g) = N_g^{-1}$, $g = 1, \ldots, G$.

Como SQP é uma forma quadrática de váriáveis com distribuição assintótica normal utilizaremos o teorema de Cochram para encontrar a distribuição de SQP sob H_0.

Teorema 2.2. (Cochran).

Se $\{\mathbf{T}_n\}$ é uma seqüência de vetores aleatórios de dimensão p tal que

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$$

e $\{\mathbf{A}_n\}$ uma seqüência de matrizes não estocásticas, semi-definidas positivas. Então

$$Q_n = n(\mathbf{T}_n - \boldsymbol{\theta})'\mathbf{A}_n(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{D} \chi^2_q$$

se e somente se $\mathbf{A}_n$ converge para alguma inversa generalizada de $\boldsymbol{\Sigma}$, isto é, $\lim\limits_{n\to\infty} \mathbf{A}_n = \mathbf{A}$, tal que, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}$ onde $posto(\mathbf{A}) = q \leq p$. (Sen e Singer, 1993) ■

Temos que

$$\mathrm{SQP} = \mathbf{H}'\mathbf{F}\mathbf{H} = K\mathbf{H}'\mathbf{F}^{\star}\mathbf{H}$$

onde $\mathbf{F}^{\star}$ é dada por (2.7).

De (2.10) e (2.11) temos que, para $K \to \infty$,

$$\frac{K}{\sqrt{\sum_k p_k(1-p_k)}}(\mathbf{H} - \boldsymbol{\mu}_{01}) \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\star}_{01}\right). \tag{2.12}$$

Note que, como $\mathbf{\Sigma}_{01}^{\star}$ é uma matriz diagonal cujos elementos da diagonal são todos positivos, $\mathbf{\Sigma}_{01}^{\star}$ é não singular e, portanto, $\mathbf{F}^{\star}$ é inversa generalizada de $\mathbf{\Sigma}_{01}^{\star}$ se e somente se $\mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star}$ é idempotente, isto é,

$$\mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star}\mathbf{F}^{\star} = \mathbf{F}^{\star} \Leftrightarrow \mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star}\mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star} = \mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star}$$

Lema 2.1. $\mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star}$ é idempotente. ∎

Demonstração.

Por (2.7) e (2.11),

$$\mathbf{F}^{\star}\mathbf{\Sigma}_{\mathbf{01}}^{\star} = \mathbf{E_{01}},$$

onde os elementos de $\mathbf{E_{01}}$ são:

$$e_{01}(g,g) = 1 - \frac{N_g}{N_T} \quad e \quad e_{01}(g,g') = -\frac{N_g}{N_T}, \;\; g \neq g', \;\; g, g' = 1, ..., G$$

Os elementos de $\mathbf{E_{01}^2}$ são

$$\begin{aligned} e_{01}^2(g,g) &= \left(1 - \frac{N_g}{N_T}\right)^2 + \frac{N_g}{N_T}(N_T - N_g) = 1 - \frac{N_g}{N_T} \\ e_{01}^2(g,g') &= -\frac{N_g}{N_T}\left(1 - \frac{N_g}{N_T} + 1 - \frac{N_{g'}}{N_T} - \frac{1}{N_T}\sum_{l \neq g,g'} N_l\right) = -\frac{N_g}{N_T} \end{aligned}$$

□

Lema 2.2. Posto($\mathbf{F}^{\star}$) $= G - 1$.

Demonstração. Multiplicar a linha r da matriz $\mathbf{F}^{\star}$ pela constante $\frac{1}{N_r}$, não altera o posto de $\mathbf{F}^{\star}$ (Rao, 1965), o que é equivalente a pre-multiplicar $\mathbf{F}^{\star}$ por matrizes elementares $G \times G$ conhecidas como $\mathbf{\Delta_r}$ de Kronecker, isto é, matrizes diagonais quadradas com elementos não nulos na diagonal, sendo neste caso:

$$\mathbf{\Delta_r} = (\delta_{gg'}): \;\; \delta_{gg} = \begin{cases} \frac{1}{N_g} & \text{se } g = r \\ \\ 1 & \text{se } g \neq r \end{cases}, \quad \delta_{gg'} = 0, \;\; g \neq g' \quad g, \; g' = 1, \ldots, G.$$

Premultiplicando $\mathbf{F}^\star$ por G matrizes de Kronecker $\mathbf{\Delta_r}$, $r = 1, ..., G$ obtemos uma matriz $\mathbf{E_1}$ cujos elementos são:

$$e_1(g,g) = 1 - \frac{N_g}{N_T}; \quad e_1(g,g') = -\frac{N_{g'}}{N_T}, \quad g \neq g', \;\; g, g' = 1, ..., G.$$

Note que $\mathbf{E_1} = \mathbf{E'_{01}}$ e portanto

$$\text{posto}(\mathbf{F}^\star) = \text{posto}(\mathbf{E'_{01}}) = \text{posto}(\mathbf{E_{01}})$$

Como o posto de uma matriz idempotente é igual ao seu traço (Rao, 1965) temos que

$$\text{posto}(\mathbf{F}^\star) = \sum_{g=1}^{G} \left(1 - \frac{N_g}{N_T}\right) = G - 1$$

□

Note que

$$\boldsymbol{\mu}'_{01}\mathbf{F}^\star\boldsymbol{\mu}_{01} = \frac{\left(\sum_k p_k\right)^2}{K^2}\mathbf{u}'_{\mathbf{G}}\mathbf{F}^\star\mathbf{u}_{\mathbf{G}} = 0, \tag{2.13}$$

pois, de (2.7) e sendo $\mathbf{u_G}$ um vetor coluna de 1's de tamanho G, $\mathbf{u}'_{\mathbf{G}}\mathbf{F}^\star$ é um vetor linha de tamanho G cujo elemento i, $i = 1, \ldots, G$ é

$$N_i\left(1 - \frac{N_i}{N_T}\right) - \frac{N_i}{N_T}(N_T - N_i) = 0 \tag{2.14}$$

Então, de (2.12), dos Lemas 2.1 e 2.2, e utilizando o Teorema 2.2 temos, para $K \to \infty$

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H} - \boldsymbol{\mu}_{01})'\mathbf{F}^\star(\mathbf{H} - \boldsymbol{\mu}_{01}) \xrightarrow{D} \chi^2_{G-1} \tag{2.15}$$

Temos que,

$$(\mathbf{H} - \boldsymbol{\mu}_{01})'\mathbf{F}^\star(\mathbf{H} - \boldsymbol{\mu}_{01}) = \mathbf{H}'\mathbf{F}^\star\mathbf{H} - 2\boldsymbol{\mu}'_{01}\mathbf{F}^\star\mathbf{H} + \boldsymbol{\mu}'_{01}\mathbf{F}^\star\boldsymbol{\mu}_{01},$$

logo, de (2.13) e (2.14),

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H} - \boldsymbol{\mu}_{01})'\mathbf{F}^\star(\mathbf{H} - \boldsymbol{\mu}_{01}) = \frac{K^2}{\sum_k p_k(1-p_k)}\mathbf{H}'\mathbf{F}^\star\mathbf{H} = \frac{K}{\sum_k p_k(1-p_k)}SQP.$$

Portanto, sob H_0 e para K suficientemente grande

$$\frac{K}{\sum_k p_k(1-p_k)}SQP \approx \chi^2_{G-1} \tag{2.16}$$

2.2.2 Distribuição Assintótica de SQI

Agora queremos obter a distribuição assintótica da soma de quadrados referente ao efeito do indivíduo dentro da população (SQI), então, de (2.4), temos

$$\text{SQI} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} K \left(\bar{X}_{gi\cdot} - \bar{X}_{g\cdot\cdot} \right)^2,$$

onde $\bar{X}_{gi\cdot}$ representa a proporção, em K loci, de fenótipos dominantes no indivíduo i do grupo g e $\bar{X}_{g\cdot\cdot}$ representa o número médio de fenótipos dominantes no grupo g :

$$\bar{X}_{gi\cdot} = \frac{X_{gi\cdot}}{K} \qquad \text{e} \qquad \bar{X}_{g\cdot\cdot} = \frac{X_{g\cdot\cdot}}{KN_g}$$

Para obtermos a distribuição assintótica de SQI, inicialmente precisamos obter a distribuição assintótica de $\bar{X}_{gi\cdot}$. Note que o número de fenótipos dominantes, em K loci, no indivíduo i da população g é

$$X_{gi\cdot} = \sum_{k=1}^{K} X_{gik}.$$

Logo,

$$\text{E}(X_{gi\cdot}) = \sum_{k} p_{gk} \qquad \text{e} \qquad \text{Var}(X_{gi\cdot}) = \sum_{k} p_{gk}(1 - p_{gk}).$$

Neste caso temos que $X_{gi1}, \ldots, X_{giK}$ são variáveis aleatórias independentes tais que,

$$\text{E}(X_{gik}) = p_{gk}, \;\; \text{Var}(X_{gik}) = p_{gk}(1 - p_{gk}), \;\; S_K = X_{gi\cdot} \;\; \text{e} \;\; s_K = \sqrt{\text{Var}(X_{gi\cdot})}.$$

Para verificarmos se a condição de Liapunov do Teorema 2.1 é satisfeita tomamos $\delta = 1$ e $0 < p_{gk} < 1$, então

$$\begin{aligned}
\frac{1}{s_K^3} \sum_{k=1}^{K} \text{E}|X_{gik} - \mu_k|^3 &= \frac{1}{s_K^3} \sum_{k=1}^{K} \text{E}|X_{gik} - p_{gk}|^3 \\
&= \frac{1}{\left(\sqrt{\sum_{k=1}^{K} p_{gk}(1 - p_{gk})} \right)^3} \sum_{k=1}^{K} [p_{gk}^3(1 - p_{gk}) + (1 - p_{gk})^3 p_{gk}] \\
&= \frac{\sum_k p_{gk}(1 - p_{gk})(1 - 2p_{gk} + 2p_{gk}^2)}{\sum_k p_{gk}(1 - p_{gk}) \left(\sqrt{\sum_k p_{gk}(1 - p_{gk})} \right)}
\end{aligned}$$

Temos que

$$\begin{aligned}\frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{\sum_k p_{gk}(1-p_{gk})}\right)} &< \frac{\sum_k p_{gk}(1-p_{gk})}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{\sum_k p_{gk}(1-p_{gk})}\right)} \\ &= \frac{1}{\sqrt{\sum_{k=1}^K p_{gk}(1-p_{gk})}}.\end{aligned}$$

Se $h^\star = \min_k\{p_{gk}(1-p_{gk})\}$, então $\sum_k p_{gk}(1-p_{gk}) \geq Kh^\star$ e portanto

$$\frac{1}{\sqrt{\sum_{k=1}^K p_{gk}(1-p_{gk})}} \leq \frac{1}{\sqrt{Kh^\star}} \to 0 \text{ quando } K \to \infty$$

o que satisfaz a condição de Liapunov.

Então temos que, quando $K \to \infty$,

$$\bar{X}_{gi\cdot} \xrightarrow{D} \mathrm{N}\left(\frac{\sum_k p_{gk}}{K}, \frac{\sum_k p_{gk}(1-p_{gk})}{K^2}\right) \tag{2.17}$$

Assim como SQP, SQI pode ser escrito como uma forma quadrática de variáveis normais. SQI também pode ser escrita de forma matricial:

$$\mathrm{SQI} = \mathbf{H_2}'\mathbf{F_2}\mathbf{H_2} \tag{2.18}$$

onde $\mathbf{H_2} = (\bar{X}_{11\cdot},\ \bar{X}_{12\cdot},\ \ldots,\ \bar{X}_{1N_1\cdot},\ \bar{X}_{21\cdot},\ \ldots,\ \bar{X}_{2N_2\cdot},\ \ldots,\ \bar{X}_{GN_G\cdot})'$ e $\mathbf{F_2} = K\mathbf{F_2^\star}$, $\mathbf{F_2^\star}$ é uma matriz simétrica $N_T \times N_T$ da forma

$$\mathbf{F}_2^\star = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \mathbf{0} \\ \cdot & \ldots & & \mathbf{0} \\ \cdot & & & \\ \cdot & & & \\ \mathbf{0} & \ldots & & \mathbf{A}_G \end{pmatrix} \tag{2.19}$$

onde $\mathbf{A}_g = (a_g(i,j))\,,\ \ g = 1,\ \ldots,\ G$, é uma matriz $N_g \times N_g$ tal que

$$a_g(i,i) = 1 - \frac{1}{N_g} \quad \text{e} \quad a_g(i,j) = -\frac{1}{N_g},\ \ i,j = 1,\ \ldots,\ N_g \tag{2.20}$$

Portanto, assintoticamente

$$\mathbf{H_2} \approx \mathrm{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right) \tag{2.21}$$

onde

$$\boldsymbol{\mu}_2 = \frac{1}{K}\left(\sum_k p_{1k}\mathbf{u}'_{\mathbf{N_1}},\ \sum_k p_{2k}\mathbf{u}'_{\mathbf{N_2}},\ \ldots,\ \sum_k p_{Gk}\mathbf{u}'_{\mathbf{N_G}}\right)' \quad \text{e} \quad \boldsymbol{\Sigma}_2 = \frac{1}{K^2}\boldsymbol{\Sigma}_2^{\star}, \tag{2.22}$$

$\mathbf{u_{N_g}}$ é um vetor coluna de 1's, de tamanho N_g, $\boldsymbol{\Sigma}_2^{\star}$ é uma matriz $N_T \times N_T$ bloco diagonal cujos blocos diagonais são $\boldsymbol{\Sigma}_{2g}^{\star} = \sum_k p_{gk}(1-p_{gk})\mathbf{I}_{N_g}$, sendo $\mathbf{I}_{N_g}$ matriz identidade $N_g \times N_g$, $g = 1,\ \ldots,\ G$.

De (2.21) temos, sob H_0

$$\mathbf{H_2} \sim \mathrm{N}\left(\boldsymbol{\mu}_{02}, \boldsymbol{\Sigma}_{02}\right) \tag{2.23}$$

onde

$\boldsymbol{\mu}_{02} = \frac{\sum_k p_k}{K}\mathbf{u_{N_T}}$ e $\boldsymbol{\Sigma}_{02}$ é uma matriz diagonal $N_T \times N_T$ da forma $\frac{\sum_k p_k(1-p_k)}{K^2}\mathbf{I}_{N_T}$.

Deste modo,

$$\mathrm{SQI} = \mathbf{H_2}'\mathbf{F_2}\mathbf{H_2} = K\mathbf{H_2}'\mathbf{F_2^{\star}}\mathbf{H_2}$$

Sob H_0,

$$\frac{K}{\sqrt{\sum_k p_k(1-p_k)}}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \mathbf{I}_{N_T}), \qquad \text{quando } \ K \to \infty,$$

pelo Teorema 2.2 e sendo $\mathbf{I}_{N_T}$ uma matriz não singular

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H_2}-\boldsymbol{\mu}_{02})'\mathbf{F}_\mathbf{2}^\star(\mathbf{H_2}-\boldsymbol{\mu}_{02}) \xrightarrow{D} \chi^2_{posto(\mathbf{F}_\mathbf{2}^\star)}$$

se e somente se $\mathbf{F}_\mathbf{2}^\star\mathbf{I}_{N_T} = \mathbf{F}_\mathbf{2}^\star$ é idempotente.

Lema 2.3. $\mathbf{F}_\mathbf{2}^\star$ é idempotente. ■

Demonstração. De (2.19) e (2.20) temos

$$\mathbf{F}_\mathbf{2}^\star = \mathbf{A^\star}_1 + \mathbf{A^\star}_2 + \ldots + \mathbf{A}_\mathbf{G}^\star$$

sendo

$$\mathbf{A^\star}_1 = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \ldots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{0} \end{pmatrix}, \quad \mathbf{A^\star}_2 = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \mathbf{0} \\ \vdots & \ldots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{0} \end{pmatrix}, \ldots$$

$$\ldots, \quad \mathbf{A^\star}_G = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \ldots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{A}_G \end{pmatrix}$$

$$\mathbf{A}_g^2 = \left(a_g^2(i,j)\right), \quad a_g^2(i,i) = \left(1-\frac{1}{N_g}\right)^2 + \sum_{j=1}^{N_g-1}\left(-\frac{1}{N_g}\right)^2 = 1-\frac{1}{N_g} = a_g(i,i)$$

$$a_g^2(i,j) = -\frac{1}{N_g}\left[2\left(1-\frac{1}{N_g}\right) - \frac{N_g-2}{N_g}\right] = -\frac{1}{N_g} = a_g(i,j), \qquad g = 1,\ \ldots,\ G.$$

Portanto, temos que

$(\mathbf{A^\star}_g)^2 = \mathbf{A^\star}_g$ para todo g e $\mathbf{A^\star}_g\mathbf{A^\star}_{g'} = \mathbf{0}$ para todo $g \neq g' \Rightarrow (\mathbf{F}_\mathbf{2}^\star)^2 = \mathbf{F}_\mathbf{2}^\star$ (Rao, 1965)

□

Lema 2.4. Posto($\mathbf{F_2^\star}$) $= N_T - G$.

Demonstração. Como $\mathbf{F_2^\star}$ é idempotente (Lema 2.3),

$$\text{Posto}(\mathbf{F_2^\star}) = \text{Traço}(\mathbf{F_2^\star}) = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\left(1 - \frac{1}{N_g}\right) = \sum_{g}(N_g - 1) = N_T - G$$

□

Temos que

$$\begin{aligned}(\mathbf{H_2} - \boldsymbol{\mu}_{02})'\mathbf{F_2^\star}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) &= \mathbf{H_2}'\mathbf{F_2^\star}\mathbf{H_2} - 2\boldsymbol{\mu}_{02}'\mathbf{F_2^\star}\mathbf{H_2} + \boldsymbol{\mu}_{02}'\mathbf{F_2^\star}\boldsymbol{\mu}_{02} \\ &= \mathbf{H_2}'\mathbf{F_2^\star}\mathbf{H_2} = \frac{SQI}{K} \end{aligned} \tag{2.24}$$

pois

$$\boldsymbol{\mu}_{02}'\mathbf{F}_2^\star\boldsymbol{\mu}_{02} = \frac{\left(\sum_k p_k\right)^2}{K^2}\mathbf{u'_{N_T}}\mathbf{F}_2^\star\mathbf{u_{N_T}} = 0,$$

sendo $\mathbf{u_{N_T}}$ um vetor coluna de 1's de tamanho N_T, $\mathbf{u'_{N_T}}\mathbf{F}_2^\star$ é um vetor linha, $1 \times N_T$ da forma $(\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_G)$ sendo $\mathbf{a}_g$, $g = 1,\ \ldots,\ G$, um vetor linha de tamanho N_g tal que

$$\mathbf{a}_g(i) = 1 - \frac{1}{N_g} - \left(\frac{N_g - 1}{N_g}\right) = 0.$$

Logo, sob H_0, pelo Teorema 2.2, Lemas 2.3 e 2.4 e (2.24), para K suficientemente grande,

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H_2} - \boldsymbol{\mu}_{02})'\mathbf{F_2^\star}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) = \frac{K}{\sum_k p_k(1-p_k)}SQI \approx \chi^2_{N_T - G} \tag{2.25}$$

2.2.3 A Estatística do Teste

Como estamos interessados em comparar grupos de seqüências de respostas binárias, de (2.16) e (2.25), sob a hipótese nula de homogeneidade entre os grupos, temos, assintoticamente ($K \to \infty$), que,

$$\frac{K}{\sum_k p_k(1-p_k)}\text{SQP} \sim \chi^2_{G-1} \quad \text{e} \quad \frac{K}{\sum_k p_k(1-p_k)}\text{SQI} \sim \chi^2_{N_T-G}$$

A hipótese nula pode ser testada através da estatística $F = \dfrac{\text{QMP}}{\text{QMI}}$, onde

$$\begin{aligned} \text{QMP} &= \frac{\text{SQP}}{G-1} \\ \text{QMI} &= \frac{\text{SQI}}{N_T-G}. \end{aligned}$$

Lema 2.5. SQP e SQI são independentes. ∎

Demonstração. De (2.6) e (2.18), SQP e SQI podem ser escritas de forma matricial. Note que

$$\mathbf{H'FH} = (\mathbf{MH_2})'\mathbf{FMH_2}$$

onde $\mathbf{M}$ é uma matriz $G \times N_T$ cujos elementos são:

$$m_{ij} = \begin{cases} \dfrac{1}{N_i} \text{ se } \displaystyle\sum_{l=1}^{i-1} N_l < j \leq \sum_{l=1}^{i} N_l \\ \\ 0 \text{ caso contrário} \end{cases}$$

$i = 1 \ldots G$.

Temos que $\mathbf{H_2}'\mathbf{F_2H_2}$ e $\mathbf{H_2}'\mathbf{M}'\mathbf{FMH_2}$ são independentes se e somente se

$$\mathbf{F_2\Sigma}_{02}\mathbf{M}'\mathbf{FM} = 0 \qquad \text{Searle, 1971}$$

De (2.7) e (2.19),

$$\mathbf{F_2\Sigma}_{02}\mathbf{M}'\mathbf{FM} = \frac{\sum_k p_k(1-p_k)}{K^2}\mathbf{F_2M}'\mathbf{FM} = \sum_k p_k(1-p_k)\mathbf{F_2^\star M}'\mathbf{F}^\star\mathbf{M} = \frac{\sum_k p_k(1-p_k)}{N_T}\mathbf{\Phi}$$

onde os elementos de $\mathbf{\Phi} = (\phi_{ij})$, $i, j = 1 \ldots N_T$, são

$$\phi_{ij} = \begin{cases} \dfrac{1}{N_i}(N_T - N_i)\left[1 - \dfrac{1}{N_i} - \dfrac{N_i - 1}{N_i}\right] = 0 \quad \text{se } \displaystyle\sum_{l=1}^{i-1} N_l < i, j \leq \sum_{l=1}^{i} N_l \\ \\ -\left(1 - \dfrac{1}{N_i}\right) + \dfrac{N_i - 1}{N_i} = 0 \qquad\qquad \text{caso contrário} \end{cases}$$

□

Portanto temos,

$$F = \frac{\text{QMP}}{\text{QMI}} = \frac{[K/\sum_k p_k(1-p_k)]\text{QMP}}{[K/\sum_k p_k(1-p_k)]\text{QMI}} \approx \frac{\left(\frac{\chi^2_{G-1}}{G-1}\right)}{\left(\frac{\chi^2_{N_T-G}}{N_T-G}\right)} \approx \text{F}_{G-1,N_T-G}, \tag{2.26}$$

ou seja, a estatística do teste tem assintoticamente distribuição de *Fisher-Snedecor* com parâmetros $G-1$ e N_T-G.

Quando o vetor de respostas apresentar pequena dimensão, por exemplo, quando os marcadores RAPD apresentarem um número pequeno de *loci*, podemos recorrer a métodos de reamostragem como o *bootstrap*.

2.3 O Poder do Teste

Agora faremos um breve estudo sobre o poder do teste. Vimos em (2.6) e (2.18) que as somas de quadrados referentes aos efeitos da população (SQP) e do indivíduo dentro da população (SQI), respectivamente, podem ser escritas de forma matricial:

$$\text{SQP} = \mathbf{H}'\mathbf{F}\mathbf{H} \quad \text{e} \quad \text{SQI} = \mathbf{H_2}'\mathbf{F_2}\mathbf{H_2}.$$

De (2.8) e (2.21) temos, assintoticamente no número de *loci* K,

$$\mathbf{H} \sim \text{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \quad \text{e} \quad \mathbf{H_2} \sim \text{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right),$$

onde $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}_2$ encontram-se definidos em (2.9) e (2.22).

Como $\mathbf{F}$ é simétrica podemos obter a fatoração:

$$\mathbf{F} = \mathbf{Q_1^\star}\mathbf{D_1^\star}\mathbf{Q_1^{\star\prime}}$$

onde $\mathbf{Q_1^\star}$ é a matriz ortogonal dos autovetores de $\mathbf{F}$ e $\mathbf{D_1^\star}$ é a matriz diagonal dos autovalores de $\mathbf{F}$. Logo,

$$\begin{aligned}
\mathbf{F} &= \mathbf{Q_1^\star}(\mathbf{D_1^\star})^{1/2}(\mathbf{D_1^\star})^{1/2}\mathbf{Q_1^{\star\prime}} \\
&= \mathbf{Q_1^\star}(\mathbf{D_1^\star})^{1/2}[(\mathbf{D_1^\star})^{1/2}]'\mathbf{Q_1^{\star\prime}} \\
&= \mathbf{Q_1^\star}(\mathbf{D_1^\star})^{1/2}[(\mathbf{D_1^\star})^{1/2}\mathbf{Q_1^\star}]' \\
&= (\mathbf{F}^{1/2})'\mathbf{F}^{1/2}
\end{aligned} \tag{2.27}$$

Como $\mathbf{F}$ é semi-definida positiva (prova em A.2), os elementos de $\mathbf{F}^{1/2} \in \mathbb{R}$.

Portanto, podemos escrever SQP como:

$$\text{SQP} = (\mathbf{F}^{1/2}\mathbf{H})'\mathbf{F}^{1/2}\mathbf{H} = \mathbf{X_1'X_1},$$

então temos que

$$\mathbf{X_1} \sim \text{N}\left(\mathbf{F}^{1/2}\boldsymbol{\mu}_1;\ \mathbf{F}^{1/2}\boldsymbol{\Sigma}_1(\mathbf{F}^{1/2})'\right).$$

Teorema 2.3. (Prova em A.1)

Se $\mathbf{X}$ é um vetor aleatório $n \times 1$, $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{V})$, onde $\mathbf{V}$ é uma matriz diagonal não singular e $\mathbf{A}$ uma matriz diagonal $n \times n$ de elementos determinísticos, então,

$$\mathbf{X'AX} \sim \sum_{i=1}^{n} \lambda_i \left(\chi_1^2(\delta_i)\right)_i,$$

onde λ_i são os autovalores da matriz $\mathbf{AV}$ e $\delta_i = \dfrac{\mu_i^2}{2\nu_i}$, sendo μ_i^2 o i-ésimo elemento do vetor $\boldsymbol{\mu}$ e ν_i os autovalores de $\mathbf{V}$.

■

Seja $\mathbf{Q}_1$ uma matriz ortogonal tal que $\mathbf{Q}_1\mathbf{F}^{1/2}\boldsymbol{\Sigma}_1(\mathbf{F}^{1/2})'\mathbf{Q}_1' = \boldsymbol{\Upsilon}_1$, onde $\boldsymbol{\Upsilon}_1$ é uma matriz diagonal.

Se $\mathbf{Y_1} = \mathbf{Q}_1\mathbf{X_1} \Rightarrow \mathbf{X_1} = \mathbf{Q}_1'\mathbf{Y_1}$, então,

$$\mathbf{Y_1} \sim N\left(\mathbf{Q}_1\mathbf{F}^{1/2}\boldsymbol{\mu}_1;\ \boldsymbol{\Upsilon}_1\right)$$

e

$$\mathbf{X_1}'\mathbf{X_1} = \mathbf{Y_1'}\mathbf{Q}_1\mathbf{Q}_1'\mathbf{Y_1} = \mathbf{Y_1'Y_1} \sim \sum_{i=1}^{G} \upsilon_{1i}(\chi_1^2(\delta_{1i}))_i$$

onde $\delta_{1i} = \dfrac{a_{1i}^2}{2\upsilon_{1i}}$, sendo a_{1i} o i-ésimo elemento do vetor $\dfrac{1}{2}\mathbf{Q}_1\mathbf{F}^{-1/2}\boldsymbol{\mu}_1$ e υ_{1i}, $i = 1, \ldots, G$, são os autovalores de $\boldsymbol{\Upsilon}_1$, e portanto são os elementos da diagonal da matriz $\boldsymbol{\Upsilon}_1$ (Teorema 2.3). Note que $\boldsymbol{\Upsilon}_1$ é semi-definida positiva pois é uma matriz de covariâncias de variáveis aleatórias normalmente distribuídas e, portanto, $\upsilon_{1i} \geq 0$.

De 2.27, analogamente, podemos obter

$$\mathbf{F_2} = (\mathbf{F_2}^{1/2})'\mathbf{F_2}^{1/2}$$

e, sendo $\mathbf{F_2}$ semi-definida positiva (prova em A.3), seus elementos $\in \mathbb{R}$.

Desta forma, temos

$$\text{SQI} = ((\mathbf{F_2})^{1/2}\mathbf{H_2})'(\mathbf{F_2})^{1/2}\mathbf{H_2} = \mathbf{X_2'X_2} = \mathbf{Y_2'Q}_2\mathbf{Q}_2'\mathbf{Y_2} = \mathbf{Y_2'Y_2},$$

onde $\mathbf{Q}_2$ é uma matriz ortogonal tal que $\mathbf{Q}_2(\mathbf{F_2})^{1/2}\mathbf{\Sigma}_2((\mathbf{F_2})^{1/2})'\mathbf{Q}_2' = \mathbf{\Upsilon}_2$ é uma matriz diagonal.

Portanto,

$$\mathbf{Y_2} \sim \text{N}\left(\mathbf{Q}_2(\mathbf{F_2})^{1/2}\boldsymbol{\mu}_2;\ \mathbf{\Upsilon}_2\right)$$

e, pelo Teorema 2.3,

$$\text{SQI} = \mathbf{Y_2'Y_2} \sim \sum_{i=1}^{N_T} \upsilon_{2i}(\chi_1^2(\delta_{2i}))_i$$

onde $\delta_{2i} = \dfrac{a_{2i}^2}{2\upsilon_{2i}}$, sendo a_{2i} o i-ésimo elemento do vetor $\dfrac{1}{2}\mathbf{Q}_2(\mathbf{F_2})^{1/2}\boldsymbol{\mu}_2$ e υ_{2i}, $i = 1, \ldots, N_T$, são os autovalores de $\mathbf{\Upsilon}_2$, e portanto são os elementos da diagonal da matriz $\mathbf{\Upsilon}_2$, neste caso $\upsilon_{2i} \geq 0$.

Então, de (2.26), para $u \in \mathbb{R}$,

$$\begin{aligned} \Pr(F \geq u) &= \Pr\left(\frac{\text{QMP}}{\text{QMI}} \geq u\right) \\ &= \Pr\left(\frac{(N_T - G)}{(G-1)}\frac{\text{SQP}}{\text{SQI}} \geq u\right) \\ &= \Pr\left(\frac{\text{SQP}}{\text{SQI}} \geq \frac{G-1}{N_T - G}u\right) \end{aligned} \tag{2.28}$$

Como SQP e SQI são combinações lineares de variáveis aleatórias com distribuição χ_1^2 cujos parâmetros de não centralidade são não negativos e cujos coeficientes da combinação linear também são todos não negativos, SQP/SQI é uma variável aleatória que assume somente valores em $\mathbb{R}_+$.

Portanto, se $N_0 = \min\limits_{0 \leq g \leq G} N_g$, para $N_0 \to \infty$, a probabilidade em (2.28) tende a 1 indicando que o poder do teste converge para 1 pois,

$$\Pr\left(\frac{\text{SQP}}{\text{SQI}} \geq \frac{G-1}{N_T - G}u\right) \longrightarrow \Pr\left(\frac{\text{SQP}}{\text{SQI}} \geq 0\right) = 1.$$

Capítulo 3

Análise de Variância para Dados Categorizados em Amostras Não Balanceadas

Nosso interesse continua sendo comparar grupos com diferentes tamanhos de amostras. No capítulo anterior esta comparação foi feita quando as variáveis apresentam respostas binárias. Agora, queremos desenvolver uma análise de variância para dados categorizados em geral, com mais de duas categorias de resposta. Por exemplo, gostaríamos de verificar se existe variabilidade entre grupos de seqüências de DNA. Neste caso, a variável resposta é politômica com quatro categorias possíveis de resposta: **A**, **C**, **T** ou **G**.

Com base em uma medida de variação para dados categorizados, em termos de freqüências para cada categoria, Light & Margolin (1971) desenvolveram uma análise de variância para dados categorizados (CATANOVA). As propriedades dos componentes de variação são investigadas através do modelo multinomial, o que permite comparar a variabilidade da variável resposta intra e entre grupos. Esta análise foi desenvolvida com aplicações para seqüências com uma única posição. No caso de seqüências de DNA, uma única posição não fornece informação suficiente. Pensando nisso, Pinheiro *et al.* (2000), utilizando uma medida de variação similar, estenderam a CATANOVA considerando várias posições ao longo das seqüências em amostras balanceadas.

No nosso caso estamos considerando diferentes números de seqüências em cada grupo (i.e., as amostras são não balanceadas). De acordo com os fatores considerados por Pinheiro *et al.* (2000), foram obtidas medidas de diversidade para amostras não balanceadas na Seção 3.1. Assumindo independência entre as posições, foi estudado o modelo probabilístico assintótico da estatística do teste sob a hipótese nula de homogeneidade entre os

grupos (Seções 3.2 3.3 e 3.4) e, finalmente, foi feito um estudo sobre o poder do teste na Seção 3.5.

3.1 Medidas de Diversidade em Amostras Não Balanceadas

Quando o número de indivíduos não é necessariamente o mesmo em cada grupo temos que $\mathbf{X_i^g} = (X_{i1}^g,\ X_{i2}^g,\ \ldots X_{iK}^g)'$ é um vetor aleatório representando a seqüência i do grupo g, onde X_{ik}^g representa a categoria presente na posição k da seqüência i do grupo g, $i = 1, \ldots, N_g$, $k = 1, \ldots, K$ e $g = 1 \ldots, G$. Note que no caso de estarmos trabalhando com seqüências em nível de nucleotídeos $X_{ik}^g \in \{\mathbf{A},\ \mathbf{C},\ \mathbf{T},\ \mathbf{G}\}$.

Iniciaremos com o caso de uma única posição, ou seja, $K = 1$. As Tabelas 3.1 e 3.2 resumem os dados para este caso.

Tabela 3.1: Resumo dos Dados (1 posição)

	Grupo				
Sequência	1	2	3	...	G
1	x_1^1	x_1^2	x_1^3	...	x_1^G
2	x_2^1	x_2^2	x_2^3	...	x_2^G
3	x_3^1	x_3^2	x_3^3	...	x_3^G
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N_g	$x_{N_1}^1$	$x_{N_2}^2$	$x_{N_3}^3$	...	$x_{N_G}^G$

Define-se

$$d_{ij} = \begin{cases} 1 \text{ se } X_i^g \neq X_j^g \\ 0 \text{ se } X_i^g = X_j^g \end{cases}$$

A variação entre as respostas categorizadas $X_1^g,\ \ldots,\ X_{N_g}^g$ é definida como sendo

$$D_{N_g} = \frac{1}{2N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} d_{ij}^2 = \frac{1}{2N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} d_{ij} \tag{3.1}$$

Tabela 3.2: Tabela de Contingência (1 posição)

Grupo	Categoria 1	2	...	C	Total
1	n_{11}	n_{21}	...	n_{C1}	$n_{\cdot 1} = N_1$
2	n_{12}	n_{22}	...	n_{C2}	$n_{\cdot 2} = N_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	n_{1G}	n_{2G}	...	n_{CG}	$n_{\cdot G} = N_G$
Total	$n_{1\cdot}$	$n_{2\cdot}$	...	$n_{C\cdot}$	$n_{\cdot\cdot} = \sum_{g=1}^{G} N_g = N_T$

Desta forma o número total de respostas é

$$N_T = \sum_{g=1}^{G} N_g = \sum_{g=1}^{G} n_{\cdot g} = \sum_{c=1}^{C} n_{c\cdot} = \sum_{g=1}^{G}\sum_{c=1}^{C} n_{cg};$$

onde n_{cg} é o número de respostas na categoria c para o grupo g e $N_g = n_{\cdot g} = \sum_{c=1}^{C} n_{cg}$ é o número de seqüências em cada grupo.

Então, a variação nas respostas, dentro do grupo g, pode ser expressa como:

$$D_{N_g} = \frac{1}{2N_g}\sum_{c\neq c'} n_{cg}n_{c'g} = \frac{1}{2N_g}\left[N_g^2 - \sum_{c=1}^{C} n_{cg}^2\right] = \frac{N_g}{2} - \frac{1}{2N_g}\sum_{c=1}^{C} n_{cg}^2 \tag{3.2}$$

$$= \frac{N_g}{2}\left\{1 - \sum_{c=1}^{C}\left(\frac{n_{cg}}{N_g}\right)^2\right\} \tag{3.3}$$

Seja $\mathbf{p} = (p_{1g}, \ldots, p_{Cg})'$ o vetor das probabilidades de X_i^g, $i = 1, \ldots, N_g$; $g = 1, \ldots, G$, pertencer a estas C categorias, o Índice de diversidade ecológica de Simpson é definido como

$$I_s(\mathbf{p}) = 1 - \mathbf{p}'\mathbf{p} = 1 - \sum_{c=1}^{C} p_{cg}^2 \tag{3.4}$$

Então, o Índice de Simpson Total (IST) é

$$IST = 1 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot}}{N_T}\right)^2. \tag{3.5}$$

A dispersão dentro do grupo g é

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cg}}{n_{\cdot g}} \right)^2 . \tag{3.6}$$

O Índice de Simpson Intra-Grupos (ISW), que é a média de (3.6) para todos os grupos, torna-se

$$ISW = \frac{1}{G} \sum_{g=1}^{G} \left[1 - \sum_{c=1}^{C} \left(\frac{n_{cg}}{n_{\cdot g}} \right)^2 \right] = 1 - \frac{1}{G} \sum_{g=1}^{G} \sum_{c=1}^{C} \left(\frac{n_{cg}}{N_g} \right)^2 . \tag{3.7}$$

Portanto, o Índice de Simpson Entre-Grupos (ISB) é

$$ISB = IST - ISW = \sum_{c=1}^{C} \left[\frac{1}{G} \sum_{g=1}^{G} \left(\frac{n_{cg}}{N_g} \right)^2 - \left(\frac{n_{c\cdot}}{N_T} \right)^2 \right] . \tag{3.8}$$

Suponha agora que existem K posições ao longo de cada seqüência. A Tabela 3.3 resume os dados quando estamos diante desta situação. Neste caso, o número total de respostas é

$$KN_T = \sum_{g=1}^{G} n_{\cdot g \cdot} = \sum_{c=1}^{C} n_{c \cdot \cdot} = \sum_{k=1}^{K} n_{\cdot \cdot k} = \sum_{c=1}^{C} \sum_{g=1}^{G} \sum_{k=1}^{K} n_{cgk};$$

onde n_{cgk} é o número de respostas pertencentes à categoria c, na posição k, para o grupo g.

A variação dentro do grupo g na posição k é

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cgk}}{n_{\cdot gk}} \right)^2 = 1 - \sum_{c=1}^{C} \left(\frac{n_{cgk}}{N_g} \right)^2 ;$$

pois $n_{\cdot gk} = N_g$.

A variação total dentro do grupo g é

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cg\cdot}}{n_{\cdot g \cdot}} \right)^2 = 1 - \sum_{c=1}^{C} \left(\frac{n_{cg\cdot}}{KN_g} \right)^2 ;$$

pois $n_{\cdot g \cdot} = KN_g$.

Tabela 3.3: Tabela de Contingência (K posições)

Grupo	Posição	1	2	...	C	Total
1	1	n_{111}	n_{211}	...	n_{C11}	$n_{\cdot 11} = N_1$
1	2	n_{112}	n_{212}	...	n_{C12}	$n_{\cdot 12} = N_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	K	n_{11K}	n_{21K}	...	n_{C1K}	$n_{\cdot 1K} = N_1$
Total		$n_{11\cdot}$	$n_{21\cdot}$	...	$n_{C1\cdot}$	$n_{\cdot 1\cdot} = KN_1$
2	1	n_{121}	n_{221}	...	n_{C21}	$n_{\cdot 21} = N_2$
2	2	n_{122}	n_{222}	...	n_{C22}	$n_{\cdot 22} = N_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
2	K	n_{12K}	n_{22K}	...	n_{C2K}	$n_{\cdot 2K} = N_2$
Total		$n_{12\cdot}$	$n_{22\cdot}$	...	$n_{C2\cdot}$	$n_{\cdot 2\cdot} = KN_2$
...	...	...	...	...	...	...
G	1	n_{1G1}	n_{2G1}	...	n_{CG1}	$n_{\cdot G1} = N_G$
G	2	n_{1G2}	n_{2G2}	...	n_{CG2}	$n_{\cdot G2} = N_G$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
G	K	n_{1GK}	n_{2GK}	...	n_{CGK}	$n_{\cdot GK} = N_G$
Total		$n_{1G\cdot}$	$n_{2G\cdot}$	...	$n_{CG\cdot}$	$n_{\cdot G\cdot} = KN_G$
TOTAL		$n_{1\cdot\cdot}$	$n_{2\cdot\cdot}$	...	$n_{C\cdot\cdot}$	$n_{\cdots} = \sum_{g=1}^{G} KN_g = KN_T$

Desta forma, as medidas de diversidade são

$$IST = 1 - \sum_{c=1}^{C} \left(\frac{n_{c\cdot\cdot}}{KN_T} \right)^2 ; \tag{3.9}$$

$$ISW = 1 - \frac{1}{G} \sum_{g=1}^{G} \sum_{c=1}^{C} \left(\frac{n_{cg\cdot}}{KN_g} \right)^2 ; \tag{3.10}$$

$$ISB = IST - ISW = \sum_{c=1}^{C} \left[\frac{1}{G} \sum_{g=1}^{G} \left(\frac{n_{cg\cdot}}{KN_g} \right)^2 - \left(\frac{n_{c\cdot\cdot}}{KN_T} \right)^2 \right]. \tag{3.11}$$

3.2 O Modelo Probabilístico

Denote por N_{cgk} o número de respostas pertencente à categoria c, na posição k para o grupo g, e p_{cgk} a probabilidade populacional de pertencer à categoria c na posição k do grupo g. Note que para cada grupo e cada posição, o vetor de respostas $(N_{1gk}, N_{2gk}, ..., N_{Cgk})$ segue uma distribuição multinomial:

$$Pr\{N_{1gk} = n_{1gk}, N_{2gk} = n_{2gk}, ..., N_{Cgk} = n_{Cgk}\} = \frac{N_g!}{\prod_{c=1}^{C} n_{cgk}!} \prod_{c=1}^{C} (p_{cgk})^{n_{cgk}}$$

onde $\sum_{c=1}^{C} n_{cgk} = N_g$, $\sum_{c=1}^{C} p_{cgk} = 1$, $p_{cgk} > 0$, $c = 1, ..., C$, $\forall\ k = 1, ..., K$ e $g = 1, ..., G$.

Logo,

$$\mathrm{E}(N_{cgk}) = N_g p_{cgk}, \qquad \mathrm{Var}(N_{cgk}) = N_g p_{cgk}(1 - p_{cgk}) \quad \text{e} \quad \mathrm{Cov}(N_{c_1gk}, N_{c_2gk}) = N_g p_{c_1gk} p_{c_2gk}$$

Para o caso geral, assumindo indepêndencia entre os grupos e entre as posições, temos que

$$\mathrm{Cov}(N_{c_1g_1k_1}, N_{c_2g_2k_2}) = -\delta N_g p_{c_1g_1k_1} p_{c_2g_2k_2}, \qquad \text{onde}$$

$$\delta = \begin{cases} 1 & \text{se } g_1 = g_2 = g \text{ e } k_1 = k_2 \\ 0 & \text{caso contrário} \end{cases} ;$$

O modelo então será

$$\begin{aligned} & \prod_{k=1}^{K}\prod_{g=1}^{G} Pr\{N_{1gk}=n_{1gk}, N_{2gk}=n_{2gk}, ..., N_{Cgk}=n_{Cgk}\} \\ = & \prod_{k=1}^{K}\prod_{g=1}^{G} \frac{N_g!}{\prod_{c=1}^{C} n_{cgk}!} \prod_{c=1}^{C} (p_{cgk})^{n_{cgk}}. \end{aligned}$$

Se $\mathbf{V}_{gk} = (N_{1gk},\ \ldots,\ N_{Cgk})'$ e $\mathbf{P}_{gk} = \boldsymbol{\mu}_{\diamond gk} = (p_{1gk},\ \ldots\ p_{Cgk})'$, $k = 1, \ldots, K$; $g = 1, \ldots, G$, são vetores $C \times 1$ podemos escrever

$$\mathrm{E}(\mathbf{V}_{gk}) = N_g \mathbf{P}_{gk} \quad e \quad \mathrm{Cov}(\mathbf{V}_{gk}) = N_g \boldsymbol{\Sigma}^{\diamond}_{gk}$$

onde

$$\boldsymbol{\Sigma}^{\diamond}_{gk} = \mathbf{D}_{gk} - \boldsymbol{\mu}_{\diamond gk}\boldsymbol{\mu}'_{\diamond gk}, \tag{3.12}$$

com $\mathbf{D}_{gk}$ sendo uma matriz diagonal $C \times C$ cujos elementos da diagonal são $p_{1gk},\ ...,\ p_{Cgk}$.

Se $\mathbf{V}_g = (\mathbf{V}_{g1},\ \mathbf{V}_{g2},\ \ldots,\ \mathbf{V}_{gK})'$ e $\mathbf{P}_g = (\mathbf{P}_{g1},\ \ldots,\ \mathbf{P}_{gK})'$ são vetores $CK \times 1$ temos que

$$\mathrm{E}(\mathbf{V}) \equiv \boldsymbol{\mu} = (N_1\mathbf{P}_1,\ N_2\mathbf{P}_2,\ \ldots,\ N_G\mathbf{P}_G)', \tag{3.13}$$

onde $\mathbf{V} = (\mathbf{V}_1,\ \mathbf{V}_2,\ \ldots,\ \mathbf{V}_G)'$ é um vetor $GCK \times 1$.

Se $\oplus$ é o operador de soma direta, ou seja, se $\mathbf{D} = \mathbf{E} \oplus \mathbf{F}$ então, $\mathbf{D} = \begin{pmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{pmatrix}$ e temos que

$$\begin{aligned} \mathrm{Cov}(\mathbf{V}) \equiv \boldsymbol{\Sigma} &= \boldsymbol{\Sigma}_{11} \oplus \boldsymbol{\Sigma}_{12} \oplus ... \oplus \boldsymbol{\Sigma}_{1K} \oplus \boldsymbol{\Sigma}_{21} \oplus ... \oplus \boldsymbol{\Sigma}_{2K} \oplus ... \oplus \boldsymbol{\Sigma}_{GK} \\ &= N_1(\boldsymbol{\Sigma}^{\diamond}_{11} \oplus ... \oplus \boldsymbol{\Sigma}^{\diamond}_{1K}) \oplus N_2(\boldsymbol{\Sigma}^{\diamond}_{21} \oplus ... \oplus \boldsymbol{\Sigma}^{\diamond}_{2K}) \oplus ... \oplus \\ &\quad \oplus N_G(\boldsymbol{\Sigma}^{\diamond}_{G1} \oplus ... \oplus \boldsymbol{\Sigma}^{\diamond}_{GK}), \end{aligned} \tag{3.14}$$

3.3 Momentos das Medidas de Diversidade

Nesta seção escreveremos as somas de quadrados (medidas de diversidade), de forma matricial afim de calcularmos os respectivos momentos.

Definição 3.1. Sejam $\mathbf{A} = (a_{ij})$ e $\mathbf{B} = (b_{ij})$ matrizes $m \times n$ e $p \times q$ respectivamente. Então o *produto de Kronecker*

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B})$$

é uma matriz $mp \times nq$ expressa como uma matriz particionada com $a_{ij}B$ como sendo a (i, j)-ésima partição, $i = 1, ..., m$ e $j = 1, ..., n$. ■

Seja

$$\mathbf{T} = \frac{1}{(KN_T)^2}\mathbf{U}_{KG} \otimes \mathbf{I}_C \tag{3.15}$$

onde $N_T = \sum_g N_g$, $\mathbf{U}_{KG}$ é uma matriz $KG \times KG$ de $1's$, $\mathbf{I}_C$ é uma matriz identidade $C \times C$.

Assim a expressão (3.9) pode ser expressa de forma matricial como,

$$IST = 1 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot\cdot}}{KN_T}\right)^2 = 1 - \mathbf{V}'\mathbf{T}\mathbf{V}; \tag{3.16}$$

Seja $\mathbf{M}$ uma matriz diagonal $G \times G$ cujos elementos da diagonal $M_{gg} = Gn_{\cdot g\cdot}^2 = G(KN_g)^2$. Então $\mathbf{M}^{-1}$ é uma matriz diagonal $G \times G$ com elementos da diagonal $M_{gg}^{-1} = \dfrac{1}{G(KN_g)^2}$.

Podemos então definir uma matriz $\mathbf{W}$ da seguinte forma:

$$\mathbf{W} = [(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C]. \tag{3.17}$$

Desta forma também podemos escrever a expressão (3.10) de forma matricial:

$$ISW \;=\; 1 - \frac{1}{G}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{KN_g}\right)^2 = 1 - \mathbf{V}'\mathbf{W}\mathbf{V}. \tag{3.18}$$

Portanto,

$$\begin{aligned} ISB &= IST - ISW \\ &= -\mathbf{V}'\mathbf{T}\mathbf{V} + \mathbf{V}'\mathbf{W}\mathbf{V} = \mathbf{V}'(-\mathbf{T} + \mathbf{W})\mathbf{V} \\ &= \mathbf{V}'\mathbf{B}\mathbf{V}; \end{aligned} \tag{3.21}$$

onde

$$\mathbf{B} = -\mathbf{T} + \mathbf{W} = \left(-\frac{1}{(KN_T)^2}\mathbf{U}_{KG} + \mathbf{M}^{-1} \otimes \mathbf{U}_K\right) \otimes \mathbf{I}_C. \tag{3.22}$$

Logo, de acordo com resultados clássicos de Modelos Lineares (Searle, 1971), temos

$$\mathrm{E}(\mathbf{V}'\mathbf{T}\mathbf{V}) = \text{traço}(\mathbf{T}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu}$$

e

$$\mathrm{E}(\mathbf{V}'\mathbf{W}\mathbf{V}) = \text{traço}(\mathbf{W}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu}.$$

Assim,

$$\begin{aligned} E(IST) &= 1 - \text{traço}(\mathbf{T}\boldsymbol{\Sigma}) - \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu} \\ &= 1 - \frac{1}{KN_T} + \frac{1}{(KN_T)^2}\sum_{c=1}^{C}\left[\sum_{g=1}^{G}\sum_{k=1}^{K} N_g p_{cgk}^2 - \left(\sum_g N_g p_{cg\cdot}\right)^2\right]; \\ E(ISW) &= 1 - \text{traço}(\mathbf{W}\boldsymbol{\Sigma}) - \boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu} \\ &= 1 - \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{N_g} + \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\left[\sum_{k=1}^{K}\frac{p_{cgk}^2}{N_g} - p_{cg\cdot}^2\right]; \\ E(ISB) &= E(IST) - E(ISW) \\ &= -\frac{1}{KN_T} + \frac{1}{(KN_T)^2}\sum_{c=1}^{C}\left[\sum_{g=1}^{G}\sum_{k=1}^{K} N_g p_{cgk}^2 - \left(\sum_g N_g p_{cg\cdot}\right)^2\right] \\ &+ \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{N_g} - \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\left[\sum_{k=1}^{K}\frac{p_{cgk}^2}{N_g} - p_{cg\cdot}^2\right]. \end{aligned}$$

Define-se a variação da população dentro do grupo g na k-ésima posição como

$$I_S(\mathbf{p}_{gk}) = 1 - \sum_{c=1}^{C} p_{cgk}^2 \, . \tag{3.23}$$

Como nosso interesse está em verificar se há homogeneidade entre os grupos, a hipótese nula é $H_0 : p_{cgk} = p_{ck}$, onde p_{cgk} é a probabilidade populacional de pertencer à categoria c na posição k do grupo g. Sob a hipótese nula, para todo g, temos que

$$I_S(\mathbf{p}_{1k}) = I_S(\mathbf{p}_{2k}) = ... = I_S(\mathbf{p}_{Gk}) = I_S(\mathbf{p}_k) \, ,$$

ou seja, a variação dentro do grupo na k-ésima posição é a mesma em todos os grupos, onde $\mathbf{p}_{gk} = (p_{1gk}, \ p_{2gk}, \ ..., p_{Cgk})'$ é um vetor $C \times 1$ representando as probabilidades associadas às categorias $c = 1, ...C$ do grupo g na posição k.

Sob a hipótese nula, as esperanças das medidas de diversidade são:

$$\begin{aligned}
\mathrm{E}_0(IST) &= 1 - \frac{1}{KN_T} + \frac{1}{(KN_T)^2} \sum_{c=1}^{C} \left[\sum_{g=1}^{G} \sum_{k=1}^{K} N_g p_{ck}^2 - \left(\sum_g N_g p_{c.} \right)^2 \right] \\
&= 1 - \frac{1}{KN_T} + \frac{1}{(KN_T)^2} \sum_{c=1}^{C} \left[\left(\sum_{g=1}^{G} N_g \right) \sum_{k=1}^{K} p_{ck}^2 - \left(p_{c.} \sum_g N_g \right)^2 \right] \\
&= 1 - \frac{1}{KN_T} + \frac{1}{K^2 N_T} \sum_{c=1}^{C} \left[\sum_{k=1}^{K} p_{ck}^2 - N_T p_{c.}^2 \right] ; \qquad (3.24)
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}_0(ISW) &= 1 - \frac{1}{GK} \sum_{g=1}^{G} \frac{1}{N_g} + \frac{1}{GK^2} \sum_{c=1}^{C} \sum_{g=1}^{G} \left[\sum_{k=1}^{K} \frac{p_{ck}^2}{N_g} - p_{c.}^2 \right] \\
&= 1 - \frac{1}{GK} \sum_{g=1}^{G} \frac{1}{N_g} + \frac{1}{GK^2} \sum_{c=1}^{C} \left[\left(\sum_{g=1}^{G} \frac{1}{N_g} \right) \sum_{k=1}^{K} p_{ck}^2 - G p_{c.}^2 \right] ; \qquad (3.25)
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}_0(ISB) &= -\frac{1}{KN_T} + \frac{1}{K^2 N_T} \sum_{c=1}^{C} \left[\sum_{k=1}^{K} p_{ck}^2 - N_T p_{c.}^2 \right] \\
&+ \frac{1}{GK} \sum_{g=1}^{G} \frac{1}{N_g} - \frac{1}{GK^2} \sum_{c=1}^{C} \left[\left(\sum_{g=1}^{G} \frac{1}{N_g} \right) \sum_{k=1}^{K} p_{ck}^2 - G p_{c.}^2 \right] . \qquad (3.26)
\end{aligned}$$

Como $\mathbf{V}$ segue uma distribuição multinomial, podemos utilizar o Teorema Central do Limite, e

$$\mathbf{V} \xrightarrow{D} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{quando} \quad N_0 \to \infty, \tag{3.27}$$

onde $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são dados em (3.13) e (3.14), respectivamente, e $N_0 = \min_{1 \le g \le G} N_g$.

Sob H_0, para $g = 1, ..., G$,

$$\boldsymbol{\Sigma}_{gk}^{\diamond} = \boldsymbol{\Sigma}_{0k}^{\diamond} \qquad \text{e} \qquad \boldsymbol{\Sigma}_{g}^{\diamond} = \boldsymbol{\Sigma}_{0}^{\diamond} = \boldsymbol{\Sigma}_{01}^{\diamond} \oplus \boldsymbol{\Sigma}_{02}^{\diamond} ... \oplus \boldsymbol{\Sigma}_{0K}^{\diamond}; \tag{3.28}$$

onde $\boldsymbol{\Sigma}_{0k}^{\diamond}$ é uma matriz $C \times C$, da forma

$$\boldsymbol{\Sigma}_{0k}^{\diamond} = \mathbf{D}_k - \boldsymbol{\mu}_{\diamond k} \boldsymbol{\mu}_{\diamond k}', \tag{3.29}$$

onde $\mathbf{D}_k$ é uma matriz diagonal $C \times C$ cujos elementos da diagonal são $p_{1k}, \ \ldots, \ p_{Ck}$ e $\boldsymbol{\mu}_{\diamond k} = (p_{1k}, \ \ldots, \ p_{Ck})'$.

Então, sob H_0

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^{\diamond} \tag{3.30}$$

onde $\boldsymbol{\eta}$ é uma matriz diagonal $G \times G$ cujos elementos da diagonal são $\eta_{gg} = N_g$ e $\boldsymbol{\Sigma}_0^{\diamond}$ dada em (3.28) .

Portanto, sob H_0, assintoticamente,

$$\mathbf{V} \xrightarrow{D} N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0); \tag{3.31}$$

onde

$$\boldsymbol{\mu}_0 = ((N_1, \ N_2, \ \ldots, \ N_G) \otimes \mathbf{P}_0)' \tag{3.32}$$

com $\mathbf{P}_0 = (p_{11}, \ \ldots, \ p_{C1}, \ p_{12}, \ \ldots, \ p_{C2}, \ \ldots, \ p_{1K}, \ \ldots, \ p_{CK})'$.

3.4 Distribuições Assintóticas e Estatística do Teste

Estamos interessados em obter uma estatística para testar a hipótese de homogeneidade entre os grupos. Para isto, iremos propor uma estatística de teste em função dos Índices de Simpson e, portanto, encontraremos a distribuição assintótica das estatísticas $\mathbf{V}'\mathbf{BV}$,$\mathbf{V}'\mathbf{TV}$ e $\mathbf{V}'\mathbf{WV}$ que estão relacionadas com os Índices de Simpson ISB, IST e ISW, respectivamente.

3.4.1 Distribuição Assintótica de V′BV

Recorde de (3.21) que ISB pode ser escrita da seguinte forma,

$$ISB = \mathbf{V}'\mathbf{B}\mathbf{V} = \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{N_{cg\cdot}}{KN_g}\right)^2 - \left(\frac{N_{c\cdot\cdot}}{KN_T}\right)^2\right]. \tag{3.33}$$

Seja $\theta_{cgk} = N_{cgk} - E_0(N_{cgk}) = N_{cgk} - N_g p_{ck}$. Então,

$$\theta_{c\cdot\cdot} = \sum_{g=1}^{G}\sum_{k=1}^{K}\theta_{cgk} = N_{c\cdot\cdot} - \sum_{g=1}^{G}N_g\sum_{k=1}^{K}p_{ck}. \tag{3.34}$$

Lembrando que $N_0 = \min_{1\le g\le G} N_g$, sob H_0,

$$\boldsymbol{\theta} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_0), \quad \text{quando } N_0 \to \infty, \tag{3.35}$$

onde $\boldsymbol{\theta} = (\theta_{111}\ \ldots\ \theta_{Cg1}\ \ldots\ \theta_{CGK})'$ e $\boldsymbol{\Sigma_0}$ é dada em (3.30).

Então, podemos escrever

$$\begin{aligned}
ISB &= \mathbf{V}'\mathbf{B}\mathbf{V} \\
&= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\theta_{cg\cdot}+N_g p_{c\cdot}}{KN_g}\right)^2 - \left(\frac{\theta_{c\cdot\cdot}+p_{c\cdot}N_T}{KN_T}\right)^2\right] \\
&= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\theta_{cg\cdot}}{KN_g}\right)^2 - \left(\frac{\theta_{c\cdot\cdot}}{KN_T}\right)^2 + 2\frac{p_{c\cdot}}{K^2}\sum_{g=1}^{G}\frac{\theta_{cg\cdot}}{GN_g} - 2\frac{p_{c\cdot}}{K^2}\frac{\theta_{c\cdot\cdot}}{N_T}\right] \\
&= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\theta_{cg\cdot}}{KN_g}\right)^2 - \left(\frac{\theta_{c\cdot\cdot}}{KN_T}\right)^2 + 2\frac{p_{c\cdot}}{K^2}\left(\sum_{g=1}^{G}\frac{\theta_{cg\cdot}}{GN_g} - \frac{\theta_{c\cdot\cdot}}{N_T}\right)\right] \\
&= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + 2\sum_{c=1}^{C}\frac{p_{c\cdot}}{K^2}\left(\sum_{g=1}^{G}\frac{\theta_{cg\cdot}}{GN_g} - \frac{\theta_{c\cdot\cdot}}{N_T}\right) \\
&= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \sum_{c=1}^{C}\sum_{g=1}^{G}\frac{2p_{c\cdot}}{GK^2}\left(\frac{\theta_{cg\cdot}}{N_g} - \frac{\theta_{c\cdot\cdot}}{N_T}\right) \\
&= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta};
\end{aligned} \tag{3.36}$$

onde $\mathbf{A} = (a_1\mathbf{A}^\star\ a_2\mathbf{A}^\star\ \ldots\ a_G\mathbf{A}^\star) = \mathbf{a}\otimes\mathbf{A}^\star$ é um vetor $1\times CGK$, $\mathbf{a} = (a_g)$ é um vetor $1\times G$ sendo $a_g = \frac{1}{GN_g} - \frac{1}{N_T}$, e $\mathbf{A}^\star$ é um vetor $1\times CK$ da forma

$$\mathbf{A}^\star = \frac{2}{GK^2}(p_{1\cdot},\ \ldots,\ p_{C\cdot},\ p_{1\cdot},\ \ldots,\ p_{C\cdot},\ \ldots,\ p_{1\cdot},\ \ldots,\ p_{C\cdot}) \tag{3.37}$$

Como já vimos, $\boldsymbol{\theta}$ tem distribuição assintótica normal, então

$$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_i \left(\chi_1^2\right)_i ; \quad \text{quando } N_0 \to \infty \text{ (Searle, 1971)} \tag{3.38}$$

onde $(\chi_1^2)_i' s$ são variáveis aleatórias independentes com distribuição qui-quadrado com 1 grau de liberdade e $\{\lambda_i,\ i = 1,\ ...\ CGK\}$ é um conjunto de autovalores de

$$\mathbf{B}\boldsymbol{\Sigma}_0 = \mathbf{B}\left(\boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^{\diamond}\right) = \left[\left(-\frac{1}{(KN_T)^2}\mathbf{U}_{KG} + \left(\mathbf{M}^{-1} \otimes \mathbf{U}_K\right)\right) \otimes \mathbf{I}_C\right]\left(\boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^{\diamond}\right)$$

por (3.22) e (3.30).

Note também que, de (3.35),

$$\mathbf{A}\boldsymbol{\theta} \xrightarrow{D} N\left(0, \mathbf{A}\boldsymbol{\Sigma}_0^{\diamond}\mathbf{A}'\right); \tag{3.39}$$

com

$$\begin{aligned}
& Var\left(\mathbf{A}\boldsymbol{\theta}\right) = \mathbf{A}\boldsymbol{\Sigma}_0^{\diamond}\mathbf{A}' \\
= \ & \frac{4}{(GK^2)^2}\sum_{g=1}^{G}\left[\left(\frac{1}{N_g} - \frac{1}{N_T}\right)^2 \sum_{c=1}^{C}\left(\sum_{k=1}^{K} p_{c.}^2 Var(\theta_{cgk}) + \right.\right. \\
+ \ & \left.\left.\sum_{c' \neq c=1}^{C}\sum_{k=1}^{K} p_{c.}p_{c'.}Cov(\theta_{cgk}, \theta_{c'gk})\right)\right] \\
= \ & \frac{4}{(GK^2)^2}\sum_{g=1}^{G}\left[\left(\frac{1}{N_g} - \frac{1}{N_T}\right)^2 \sum_{c=1}^{C}\left(\sum_{k=1}^{K} p_{c.}^2\left[N_g p_{cgk}(1 - p_{cgk})\right] + \right.\right. \\
+ \ & \left.\left.\sum_{c' \neq c=1}^{C}\sum_{k=1}^{K} -N_g p_{c.}p_{c'.}p_{cgk}p_{c'gk}\right)\right].
\end{aligned}$$

Então, podemos dizer que, sob H_0,

$$ISB = \mathbf{V}'\mathbf{B}\mathbf{V} = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_i \left(\chi_1^2\right)_i + N\left(0, \mathbf{A}\boldsymbol{\Sigma}_0^{\diamond}\mathbf{A}'\right); \tag{3.40}$$

ou seja, ISB é a soma de uma combinação linear de variáveis aleatórias com distribuição χ_1^2 e outra com distribuição normal. Pelo Lema 3.1 temos que $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $\mathbf{A}\boldsymbol{\theta}$ não são independentes e portanto a distribuição de $\mathbf{V}'\mathbf{B}\mathbf{V}$ não é uma convolução destas variáveis aleatórias.

Lema 3.1. $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $\mathbf{A}\boldsymbol{\theta}$ não são independentes. (Prova em B.1.) ∎

Uma maneira alternativa de se obter a distribuição de $\mathbf{V}'\mathbf{B}\mathbf{V}$ é a seguinte:

Seja $\mathbf{R}$ uma matriz $CGK \times CGK$ tal que $\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}' = \mathbf{I}_{CGK}$ e

$$\mathbf{Y} = \mathbf{R}\mathbf{V} \Rightarrow \mathbf{V} = \mathbf{R}^{-1}\mathbf{Y}.$$

Logo, para $N_0 \to \infty$,

$$\mathbf{Y} \xrightarrow{D} N(\mathbf{R}\boldsymbol{\mu}, \mathbf{I}_{CGK})$$

e

$$\mathbf{V}'\mathbf{B}\mathbf{V} = \mathbf{Y}'(\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}\mathbf{Y} \equiv \mathbf{Y}'\mathbf{C}\mathbf{Y},$$

onde $\mathbf{C} = (\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}$.

Seja $\mathbf{P}$ uma matriz ortogonal tal que $\mathbf{P}\mathbf{C}\mathbf{P}'$ seja uma matriz diagonal e

$$\mathbf{Y}^\star = \mathbf{P}\mathbf{Y} \Rightarrow \mathbf{Y} = \mathbf{P}^{-1}\mathbf{Y}^\star = \mathbf{P}'\mathbf{Y}^\star.$$

Portanto,

$$\mathbf{Y}^\star \xrightarrow{D} N(\mathbf{P}\mathbf{R}\boldsymbol{\mu}, \mathbf{I}_{CGK});$$

$$\mathbf{V}'\mathbf{B}\mathbf{V} = \mathbf{Y}'\mathbf{C}\mathbf{Y} = (\mathbf{Y}^\star)'\mathbf{P}\mathbf{C}\mathbf{P}'\mathbf{Y}^\star = (\mathbf{Y}^\star)'\mathbf{C}^\star\mathbf{Y}^\star;$$

onde $\mathbf{C}^\star \equiv \mathbf{P}(\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}\mathbf{P}'$.
Logo,

$$ISB = \mathbf{V}'\mathbf{B}\mathbf{V} = (\mathbf{Y}^\star)'\mathbf{C}^\star\mathbf{Y}^\star \xrightarrow{D} \sum_{i=1}^{CGK} c_i^\star \left(\chi_1^2(\delta_{1i})\right); \tag{3.41}$$

onde $\delta_{1i} \equiv \dfrac{(\mu_i^\star)^2}{2}$, sendo $\mu_i^\star$ o i-ésimo elemento do vetor $\boldsymbol{\mu}^\star = \mathbf{P}\mathbf{R}\boldsymbol{\mu}$ e $c_i^\star$'s os elementos da diagonal de $\mathbf{C}^\star$, $c_i^\star \in \mathbb{R}$. (Ver prova em A.1).

3.4.2 Autovalores da Matriz $\mathbf{C}^\star$: Um exemplo utilizando dados reais.

Nesta seção iremos mostrar, através de um conjunto de dados reais, possíveis valores de $c_i^\star$, os autovalores da matriz diagonal $\mathbf{C}^\star$, que estão definidos em (3.41).

O conjunto de dados que iremos utilizar consiste de 48 seqüências do DNA mitocondrial de cágados da espécie *Hydromedusa maximiliani*. Estas seqüências se encontram divididas em três grupos onde cada grupo possui 30, 7 e 11 seqüências. Este conjunto de seqüências são provenientes de duas regiões distintas do DNA mitocondrial denominadas gene citocromo b e região de controle (maiores detalhes podem ser vistos no Capítulo 4). Iremos então obter dois conjuntos de autovalores de $\mathbf{C}^\star$ correspondentes aos dados destas duas regiões do DNA mitocondrial dos cágados.

Seguindo os passos, da seção anterior, para obtermos o resultado em (3.41), inicialmente encontramos a matriz $\mathbf{R}$ tal que $\mathbf{\Sigma} = \mathbf{R}^{-1}(\mathbf{R}')^{-1}$. Pelo fato da matriz de covariâncias $\mathbf{\Sigma}$ ser singular, não é possível utilizar a fatorização de Cholesky para obtermos $\mathbf{R}$. Mas como $\mathbf{\Sigma}$ é simétrica, foi possível fazer a decomposição

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{R}^{-1}(\mathbf{R}')^{-1} = (\mathbf{Q}\mathbf{D}^{\mathbf{1/2}})(\mathbf{Q}\mathbf{D}^{\mathbf{1/2}})' \\ &= \mathbf{Q}\mathbf{D}^{\mathbf{1/2}}(\mathbf{D}^{\mathbf{1/2}})'\mathbf{Q}' \\ &= \mathbf{Q}\mathbf{D}\mathbf{Q}', \end{aligned} \tag{3.42}$$

onde $\mathbf{Q}$ é uma matriz ortogonal dos autovetores de $\mathbf{\Sigma}$, e $\mathbf{D}$ é uma matriz diagonal dos autovalores de $\mathbf{\Sigma}$ (elementos da diagonal). Como $\mathbf{\Sigma}$ é semi-definida positiva, seus autovalores são números não negativos e portanto $\mathbf{D}^{\mathbf{1/2}}$ é uma matriz diagonal cujos elementos $\in \mathbb{R}_+$.

Em seguida, após encontramos $\mathbf{C} = (\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}$, obtivemos a matriz diagonal $\mathbf{C}^\star = \mathbf{P}\mathbf{C}\mathbf{P}'$ através do mesmo procedimento em (3.42), $\mathbf{C} = \mathbf{P}'\mathbf{C}^\star\mathbf{P}$, onde $\mathbf{P}'$ é a matriz dos autovetores de $\mathbf{C}$, ortogonal, e $\mathbf{C}^\star$ a matriz diagonal contendo os autovalores de $\mathbf{C}$ (elementos da diagonal). O procedimento numérico para realizar esta fatorização foi realizado pela função *EIG* do software *MATLAB*.

A Tabela 3.4 mostra os autovalores não nulos, correspondentes às seqüências do gene citocromo b e da região de controle. Os valores, na tabela abaixo, correspondem aos primeiroas elementos da diagonal na matriz $\mathbf{C}^\star$. Foram observadas 50 posições ao longo de cada seqüência do gene citocromo b e 61 posições em cada seqüência da região de controle. Portanto, existem $CGK = 4 \times 3 \times 50 = 600$ e $4 \times 3 \times 61 = 732$ autovalores de $\mathbf{C}$, respectivamente, para as duas regiões.

Tabela 3.4: Autovalores.

Elementos	Gene citocromo b (10^{-5})	Região de controle (10^{-5})
1	$0,7104$	$0,5646$
2	$0,5558$	$0,4417$
3	$0,1236$	$0,1379$
4	$0,0967$	$0,1079$
5	$-0,0200$	$-0,0159$
6	$0,0363$	$-0,0039$
7	$0,0284$	$-$
8	$-0,0035$	$-$
9	$-0,0010$	$-$

3.4.3 Distribuição Assintótica de V′TV

Lembremos que,

$$IST = 1 - \mathbf{V}'\mathbf{T}\mathbf{V}.$$

Sob H_0 $\mathbf{V} \xrightarrow{D} N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Como $\mathbf{T}\boldsymbol{\Sigma}_0$ não é idempotente, a distribuição de $\mathbf{V}'\mathbf{T}\mathbf{V}$ não é $\chi^2_{(posto(\mathbf{T}),\boldsymbol{\mu}_0'\mathbf{T}\boldsymbol{\mu}_0)}$. No entanto,

$$\begin{aligned}
\mathbf{V}'\mathbf{T}\mathbf{V} &= \sum_{c=1}^{C}\left(\frac{N_{c\cdot\cdot}}{K\sum_{g=1}^{G}N_g}\right)^2 \\
&= \sum_{c=1}^{C}\left(\frac{\theta_{c\cdot\cdot}+\sum_{g=1}^{G}N_g p_{c\cdot}}{K\sum_{g=1}^{G}N_g}\right)^2 \\
&= \sum_{c=1}^{C}\left[\frac{\theta_{c\cdot\cdot}^2+2\theta_{c\cdot\cdot}\sum_{g=1}^{G}N_g p_{c\cdot}+\left(\sum_{g=1}^{G}N_g p_{c\cdot}\right)^2}{K\sum_{g=1}^{G}N_g}\right]
\end{aligned} \tag{3.43}$$

$$
\begin{aligned}
&= \sum_{c=1}^{C} \left(\frac{\theta_{c\cdot\cdot}}{KN_T} \right)^2 + \frac{2}{K^2 N_T} \sum_{c=1}^{C} \theta_{c\cdot\cdot} p_{c\cdot} + \frac{1}{K^2} \sum_{c=1}^{C} p_{c\cdot}^2 \\
&= \boldsymbol{\theta}' \mathbf{T} \boldsymbol{\theta} + \mathbf{A_2} \boldsymbol{\theta} + \delta; \qquad (3.44)
\end{aligned}
$$

onde $\mathbf{A_2}' = (\mathbf{A_2^\star}\ \mathbf{A_2^\star}\ ...\ \mathbf{A_2^\star})' = (\mathbf{1}_G \otimes \mathbf{A_2^\star})'$ é um vetor $CGK \times 1$, $\mathbf{1}_G$ é um vetor linha de 1's de tamanho G e $\mathbf{A_2^\star}$ é um vetor $1 \times CK$ da forma

$$
\mathbf{A_2^\star} = \frac{2}{K^2 N_T} \left(p_{1\cdot},\ \ldots,\ p_{C\cdot},\ p_{1\cdot},\ \ldots,\ p_{C\cdot},\ \ldots,\ p_{1\cdot},\ \ldots,\ p_{C\cdot} \right).
$$

De acordo com o Lema 3.2 temos que $\boldsymbol{\theta}' \mathbf{T} \boldsymbol{\theta}$ e $\mathbf{A_2} \boldsymbol{\theta}$ não são independentes e neste caso $\mathbf{V}'\mathbf{T}\mathbf{V}$ não será a convolução destas variáveis.

Lema 3.2. $\boldsymbol{\theta}' \mathbf{T} \boldsymbol{\theta}$ e $\mathbf{A_2} \boldsymbol{\theta}$ não são independentes. (Prova em B.2.) ∎

Neste caso , sob H_0,

$$
\boldsymbol{\theta}' \mathbf{T} \boldsymbol{\theta} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_{2i} \left(\chi_1^2 \right)_i \quad e \quad \mathbf{A_2} \boldsymbol{\theta} \xrightarrow{D} N(0, \mathbf{A}_2 \boldsymbol{\Sigma}_0 \mathbf{A}_2'), \quad \text{quando } N_0 \to \infty.
$$

Portanto, assintoticamente, podemos dizer que $\mathbf{V}'\mathbf{T}\mathbf{V}$ é a soma de uma combinação linear de variáveis aleatórias com distribuição χ_1^2 e de uma variável aleatória normal.

$$
\mathbf{V}'\mathbf{T}\mathbf{V} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_{2i} \left(\chi_1^2 \right)_i + N(0, \mathbf{A}_2 \boldsymbol{\Sigma}_0 \mathbf{A}_2') + \delta;
$$

onde $\lambda_{2i}, i = 1, ..., CGK$ é o conjunto de autovalores de

$$
\mathbf{T}\boldsymbol{\Sigma}_0 = \frac{1}{(KN_T)^2} \mathbf{T}^\diamond \boldsymbol{\Sigma}_0 = \frac{1}{(KN_T)^2} \mathbf{T}^\diamond \boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^\diamond.
$$

Alternativamente, temos que a distribuição de $\mathbf{V}'\mathbf{T}\mathbf{V}$ é análoga à obtida em (3.41):

$$
\mathbf{V}'\mathbf{T}\mathbf{V} = (\mathbf{Y_2^\star})' \mathbf{C_2^\star} \mathbf{Y_2^\star} \xrightarrow{D} \sum_{i=1}^{CGK} c_{2i}^\star \left(\chi_1^2(\delta_{2i}) \right), \quad \text{com } \delta_{2i} \equiv \frac{(\mu_{2i}^\star)^2}{2}; \qquad (3.45)
$$

onde $c^{\star}_{2i}$'s , $c^{\star}_{2i} \in \mathbb{R}$, são os elementos da diagonal da matriz diagonal

$$\mathbf{C_2^{\star}} \equiv \mathbf{P_2}(\mathbf{R_2}^{-1})'\mathbf{T}\mathbf{R_2}^{-1}\mathbf{P_2},$$

onde $\mathbf{P_2}$ é uma matriz ortogonal e $\mu^{\star}_{2i}$ o i-ésimo elemento do vetor $\boldsymbol{\mu}^{\star}_2 = \mathbf{P_2}\mathbf{R_2}\boldsymbol{\mu}$.

Note que

$$\mathbf{Y}^{\star}_{\mathbf{2}} \xrightarrow{D} N(\boldsymbol{\mu}^{\star}_2, \mathbf{I}_{CGK});$$

onde $\mathbf{R_2}$ é uma matriz $CGK \times CGK$ tal que $\mathbf{R_2}\boldsymbol{\Sigma}\mathbf{R}'_{\mathbf{2}} = \mathbf{I}_{CGK}$.

3.4.4 Distribuição Assintótica de V′WV

No caso do Índice de Simpson Intra-Grupos, temos que

$$ISW = 1 - \mathbf{V}'\mathbf{W}\mathbf{V},$$

sob H_0,

$$\begin{aligned}
\mathbf{V}'\mathbf{W}\mathbf{V} &= \frac{1}{G}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{N_{cg\cdot}}{KN_g}\right)^2 \\
&= \frac{1}{G}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\theta_{cg\cdot} + N_g p_{c\cdot}}{KN_g}\right)^2 \\
&= \frac{1}{GK^2}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\theta^2_{cg\cdot} + 2\theta_{cg\cdot}N_g p_{c\cdot} + (N_g p_{c\cdot})^2}{N_g^2}\right) \\
&= \frac{1}{GK^2}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\theta^2_{cg\cdot}}{N_g^2} + \frac{2\theta_{cg\cdot}p_{c\cdot}}{N_g} + p^2_{c\cdot}\right) \\
&= \frac{1}{G}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\theta_{cg\cdot}}{N_g}\right)^2 + \frac{2}{G}\sum_{g=1}^{G}\sum_{c=1}^{C}\frac{\theta_{cg\cdot}p_{c\cdot}}{K^2N_g} + \sum_{c=1}^{C}\frac{p^2_{c\cdot}}{K^2} \\
&= \boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta} + \mathbf{A_3}\boldsymbol{\theta} + \delta; \qquad (3.46)
\end{aligned}$$

onde $\mathbf{A_3}' = (\mathbf{A_3^{\star}}\ \mathbf{A_3^{\star}}\ ...\ \mathbf{A_3^{\star}})' = (\mathbf{n_3} \otimes \mathbf{A_3^{\star}})'$ é um vetor $CGK \times 1$, $\mathbf{n_3} = \left(\frac{1}{N_1}\ \ldots\ \frac{1}{N_g}\right)$ é um vetor linha de tamanho G e $\mathbf{A_3^{\star}}$ é um vetor $1 \times CK$ da forma

$$\mathbf{A_3^{\star}} = \frac{2}{K^2G}\left(p_{1\cdot},\ \ldots,\ p_{C\cdot},\ p_{1\cdot},\ \ldots,\ p_{C\cdot},\ \ldots,\ p_{1\cdot},\ \ldots,\ p_{C\cdot}\right).$$

Mostramos também (Lema 3.3) que $\mathbf{V'WV}$ não é uma convolução de $\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta}$ e $\mathbf{A_3}\boldsymbol{\theta}$.

Lema 3.3. $\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta}$ e $\mathbf{A_3}\boldsymbol{\theta}$ não são independentes. (Prova em B.3.) ∎

Desta forma temos,

$$\mathbf{V'WV} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_{3i} \left(\chi_1^2\right)_i + N(0, \mathbf{A}_3\boldsymbol{\Sigma}_0\mathbf{A}_3') + \delta;$$

onde $\lambda_{3i}, i = 1, ..., CGK$ é o conjunto de autovalores de

$$\mathbf{W}\boldsymbol{\Sigma}_0 = [(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C]\boldsymbol{\Sigma}_0.$$

Em outras palavras, $\mathbf{V'WV}$ é a soma de uma combinação linear de variáveis aleatórias com distribuição χ_1^2, uma variável aleatória com distribuição normal e uma constante.

Da mesma forma que em (3.41) obtemos a distribuição de $\mathbf{V'WV}$ como sendo

$$\mathbf{V'WV} = (\mathbf{Y_3^\star})'\mathbf{C_3^\star}\mathbf{Y_3^\star} \xrightarrow{D} \sum_{i=1}^{CGK} c_{3i}^\star \left(\chi_1^2(\delta_{3i})\right) \quad \text{com } \delta_{3i} \equiv \frac{(\mu_{3i}^\star)^2}{2} \tag{3.47}$$

onde $c_{3i}^\star$'s, $c_{3i}^\star \in \mathbb{R}$, são os elementos da diagonal da matriz diagonal

$$\mathbf{C_3^\star} \equiv \mathbf{P_3}(\mathbf{R_3}^{-1})'\mathbf{T}\mathbf{R_3}^{-1}\mathbf{P_3'},$$

onde $\mathbf{P_3}$ é uma matriz ortogonal e $\mu_{3i}^\star$ o i-ésimo elemento do vetor $\boldsymbol{\mu}_3^\star = \mathbf{P_3}\mathbf{R_3}\boldsymbol{\mu}$.

Note que

$$\mathbf{Y_3^\star} \xrightarrow{D} N(\boldsymbol{\mu}_3^\star, \mathbf{I}_{CGK});$$

onde $\mathbf{R_3}$ é uma matriz $CGK \times CGK$ tal que $\mathbf{R_3}\boldsymbol{\Sigma}\mathbf{R_3'} = \mathbf{I}_{CGK}$.

3.4.5 A Estatística do Teste

Para testarmos se há homogeneidade entre os grupos queremos obter a distribuição assintótica de uma estatítisca que seja função de ISB/ISW. Para isto precisamos estudar as ordens de convergência de algumas estatísticas ou de seus momentos.

Seja $N_0 = \min_{1 \le g \le G} N_g$. Temos, de (3.26), (3.24) e (3.25), respectivamente,

$$
\begin{aligned}
\theta_1 &\equiv E_0(ISB) = -\frac{1}{KN_T} + \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{N_g} + \frac{1}{K^2N_T}\sum_{c=1}^{C}\left(\sum_{k=1}^{K}p_{ck}^2 - N_T p_{c.}^2\right) + \\
&- \frac{1}{GK^2}\sum_{c=1}^{C}\left(\sum_{g=1}^{G}\frac{1}{N_g}p_{ck}^2 - G p_{c.}^2\right) \\
&= -\frac{1}{KN_T} + \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{N_g} + \frac{1}{K^2N_T}\sum_{c=1}^{C}\sum_{k=1}^{K}p_{ck}^2 - \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\frac{1}{N_g}p_{ck}^2 \\
\theta_2 &\equiv E_0(IST) = 1 + \frac{1}{KN_T}\left(\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{p_{ck}^2}{K} - 1\right) - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2} \\
&= 1 + O(N_0^{-1}) - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2} \\
\theta_3 &\equiv E_0(ISW) = 1 - \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{N_g} + \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\frac{1}{N_g}p_{ck}^2 - \frac{p_{c.}^2}{K^2} \\
&= 1 + O(N_0^{-1}) - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2}
\end{aligned}
$$

Definamos agora,

$$
\begin{aligned}
T_1 &\equiv ISB - \theta_1; \\
T_2 &\equiv IST - \theta_2; \\
T_3 &\equiv ISW - \theta_3.
\end{aligned}
$$

Se $N_g = O(N_0)\ \forall g, g = 1, ..., G$ temos que

(i) $\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} \sim \sum_{i=1}^{CGK}\lambda_{2i}\left(\chi_1^2\right)_i = O_p(N_0^{-1})$,
pois $\{\lambda_{2i}, i = 1, ..., CGK\}$ é o conjunto de autovalores de
$\mathbf{T}\boldsymbol{\Sigma}_0 = O(N_0^{-2})O(N_0) = O(N_0^{-1})$;

(ii) $\mathbf{A_2}\boldsymbol{\theta} = O_p(N_0^{-1/2})$, pois $\mathbf{A_2}\boldsymbol{\theta} \sim N(0, \mathbf{A}_2\boldsymbol{\Sigma}_0\mathbf{A}_2')$, $\mathbf{A_2} = O(N_0^{-1})$
e $\boldsymbol{\Sigma}_0 = O(N_0)$;

$$(iii) \quad \delta = \frac{1}{K^2}\sum_{c=1}^{C} p_{c.}^2 = O(1).$$

Analogamente temos,

$$(i) \quad \boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta} = \sum_{i=1}^{CGK} \lambda_{3i}\left(\chi_1^2\right)_i = O_p(N_0^{-1});$$

$$(ii) \quad \mathbf{A_3}\boldsymbol{\theta} = O_p(N_0^{-1/2}).$$

Então,

$$\begin{aligned} T_2 &= IST - \theta_2 = 1 - \mathbf{V}'\mathbf{T}\mathbf{V} - \theta_2 \\ &= 1 - \left(O_p(N_0^{-1}) + O_p(N_0^{-1/2}) + \frac{1}{K^2}\sum_{c=1}^{C} p_{c.}^2 \right) \\ &\quad - \left(1 + O(N_0^{-1}) - \frac{1}{K^2}\sum_{c=1}^{C} p_{c.}^2 \right) \\ &= O_p(N_0^{-1/2}) \end{aligned}$$

e

$$\begin{aligned} T_3 &= ISW - \theta_3 = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} - \theta_3 \\ &= 1 - \left(O_p(N_0^{-1}) + O_p(N_0^{-1/2}) + \frac{1}{K^2}\sum_{c=1}^{C} p_{c.}^2 \right) \\ &\quad - \left(1 + O(N_0^{-1}) - \frac{1}{K^2}\sum_{c=1}^{C} p_{c.}^2 \right) \\ &= O_p(N_0^{-1/2}). \end{aligned}$$

De modo que,

$$ISB = \mathbf{V}'\mathbf{B}\mathbf{V} = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} = O_p(N_0^{-1}) + O_p(N_0^{-1/2}) = O_p(N_0^{-1/2}).$$

Então, para testar a hipótese de homogeneidade entre os grupos, iremos propor a seguinte estatística:

$$F_1 \equiv N_0^{1/2}\left(\frac{ISB}{ISW}\right); \tag{3.48}$$

Logo,

$$\begin{aligned} F_1 = N_0^{1/2}\left(\frac{ISB}{ISW}\right) &= N_0^{1/2}\left(\frac{ISB}{T_3+\theta_3}\right) \\ &= N_0^{1/2}\frac{ISB}{\theta_3}\left(1-\frac{T_3}{\theta_3+T_3}\right) \\ &= N_0^{1/2}\frac{ISB}{\theta_3} + O(N_0^{-1/2}) = O(1); \end{aligned}$$

pois $N_0^{1/2}ISB = O_p(1)$, $\theta_3 + T_3 = O(1) + O_p(N_0^{-1/2}) = O_p(1)$ e

$$\begin{aligned} \theta_3 &= 1 - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2} + O(N_0^{-1}) \\ &= \theta_3^0 + O(N_0^{-1}). \end{aligned}$$

Desta forma, por (3.41), temos que, para $N_g = O(N_0)$ e $N_0 \to \infty$, F_1 é expressa através de uma combinação linear de variáveis aleatórias com distribuição de qui-quadrado.

$$F_1 = N_0^{1/2}\frac{ISB}{\theta_3^0} \sim \sum_{i=1}^{CGK}\frac{N_0^{1/2}}{\theta_3^0}c_i^\star\left(\chi_1^2(\delta_i)\right), \tag{3.49}$$

onde $c_i^\star$ e δ_i estão obtidos de acordo com (3.41).

Quando o tamanho das amostras (N_g) for pequeno recorreremos a métodos de reamostragem como o *bootstrap* e assim obtemos a distribuição empírica de F_1.

3.5 Poder do Teste

Na seção anterior obtivemos a distribuição assintótica da estatística do teste para a hipótese nula de homogeneidade entre grupos. Agora iremos considerar as seguintes hipóteses alternativas para estudarmos o poder do teste:

$$H_1 : p_{cgk} = \frac{1}{\sqrt{N_g}}\gamma_{cgk} + p_{ck}$$

para todo $g = 1, \ldots, G$.

Estas hipóteses alternativas se referem à *alternativas de Pitman* (Pinheiro *et al.*, 2000). Desse modo, também pode-se verificar o comportamento do poder do teste para alternativas que se tornam cada vez mais próximas da hipótese nula conforme o aumento de N_0.

Estamos interessados no caso em que $\gamma_{cgk} \neq 0$. Sob a hipótese alternativa temos:

$$\begin{aligned}
\theta_{cgk} &= N_{cgk} - E(N_{cgk}) = N_{cgk} - N_g\left(\frac{\gamma_{cgk}}{\sqrt{N_g}} + p_{ck}\right), \\
\theta_{cg\cdot} &= N_{cg\cdot} - N_g\left(\frac{\gamma_{cg\cdot}}{\sqrt{N_g}} + p_{c\cdot}\right) \text{ e} \\
\theta_{c\cdot\cdot} &= N_{c\cdot\cdot} - \sum_g N_g\left(\frac{\gamma_{cg\cdot}}{\sqrt{N_g}} + p_{c\cdot}\right).
\end{aligned}$$

Agora,

$$\begin{aligned}
ISB &= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{n_{cg\cdot}}{KN_g}\right)^2 - \left(\frac{n_{c\cdot\cdot}}{KN_T}\right)^2\right] \\
&= \sum_{c=1}^{C}\sum_{g=1}^{G}\left[\frac{1}{G}\left(\frac{n_{cg\cdot}}{KN_g}\right)^2 - \frac{1}{G}\left(\frac{n_{c\cdot\cdot}}{KN_T}\right)^2\right] \\
&= \sum_{c=1}^{C}\sum_{g=1}^{G}\left[\frac{1}{G}\left(\frac{\theta_{cg\cdot} + N_g\left(\frac{\gamma_{cg\cdot}}{\sqrt{N_g}} + p_{c\cdot}\right)}{KN_g}\right)^2 - \frac{1}{G}\left(\frac{\theta_{c\cdot\cdot} + \sum_g N_g\left(\frac{\gamma_{cg\cdot}}{\sqrt{N_g}} + p_{c\cdot}\right)}{KN_T}\right)^2\right] \\
&= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\theta_{cg\cdot} + N_g p_{c\cdot}}{KN_g}\right)^2 - \left(\frac{\theta_{c\cdot\cdot} + p_{c\cdot}N_T}{KN_T}\right)^2\right] \\
&+ \sum_{c=1}^{C}\sum_{g=1}^{G}\left\{\frac{1}{G}\left[2\frac{\gamma_{cg\cdot}}{\sqrt{N_g}}\frac{N_g(\theta_{cg\cdot} + N_g p_{c\cdot})}{(KN_g)^2}\right] - \left[2\frac{\sum_g N_g\gamma_{cg\cdot}}{\sqrt{N_g}}\frac{(\theta_{c\cdot\cdot} + N_T p_{c\cdot})}{(KN_T)^2}\right]\right.
\end{aligned}$$

$$+ \quad \left(\frac{N_g\gamma_{cg\cdot}}{KN_g\sqrt{N_g}}\right)^2 - \left(\frac{\sum_g N_g\gamma_{cg\cdot}}{KN_T\sqrt{N_g}}\right)^2\Bigg\}.$$

De (3.36) temos

$$\begin{aligned} ISB &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} + \sum_{c=1}^{C}\sum_{g=1}^{G}\left\{\frac{1}{G}\left[2\frac{\gamma_{cg\cdot}}{\sqrt{N_g}}\frac{N_g(\theta_{cg\cdot}+N_gp_{c\cdot})}{(KN_g)^2}\right]\right. \\ &- \left.\left[2\frac{\sum_g N_g\gamma_{cg\cdot}}{\sqrt{N_g}}\frac{(\theta_{c\cdot\cdot}+N_Tp_{c\cdot})}{(KN_T)^2}\right] + \left(\frac{N_g\gamma_{cg\cdot}}{KN_g\sqrt{N_g}}\right)^2 - \left(\frac{\sum_g N_g\gamma_{cg\cdot}}{KN_T\sqrt{N_g}}\right)^2\right\} \\ &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} + \sum_c\sum_g\left[\frac{2\sqrt{N_g}\gamma_{cg\cdot}}{K^2}\left(\frac{\theta_{cg\cdot}}{GN_g^2} - \frac{\theta_{c\cdot\cdot}}{N_T^2}\right)\right] \\ &+ \sum_c\sum_g\left[\frac{2p_{c\cdot}\sqrt{N_g}\gamma_{cg\cdot}}{K^2}\left(\frac{1}{GN_g} - \frac{1}{N_T}\right)\right] \\ &+ \sum_c\sum_g\frac{1}{GK^2}\left[\left(\frac{\gamma_{cg\cdot}}{\sqrt{N_g}}\right)^2 - \left(\frac{\sum_g\sqrt{N_g}\gamma_{cg\cdot}}{N_T}\right)^2\right] \\ &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} + \mathbf{A_4}\boldsymbol{\theta} + \delta^{\star\star} \\ &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + (\mathbf{A}+\mathbf{A_4})\boldsymbol{\theta} + \delta^{\star\star}, \end{aligned}$$

onde $\mathbf{A}$ é um vetor $1\times CGK$ definido em (3.36) e (3.37), e $\mathbf{A_4} = (\mathbf{A}_4^\star,\ \mathbf{A}_4^\star,\ \ldots,\ \mathbf{A}_4^\star)$, um vetor $1\times CGK$, sendo $\mathbf{A}_4^\star = (\mathbf{A}_{41}^\star,\ \mathbf{A}_{42}^\star,\ \ldots,\ \mathbf{A}_{4G}^\star)$, $1\times CG$, com $\mathbf{A}_{4g}^\star = (\mathbf{a}_{41g}^\star,\ \mathbf{a}_{42g}^\star,\ \ldots,\ \mathbf{a}_{4Cg}^\star)$, $1\times C$,

$$\mathbf{a}_{4cg}^\star = \frac{2\sqrt{N_g}}{K^2}\left(\frac{1}{GN_g^2} - \frac{1}{N_T}\right)\gamma_{cg} - \sum_{g'\neq g}\frac{2\sqrt{N_{g'}}}{(KN_T)^2}\gamma_{cg'}.$$

Como $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $\mathbf{A}\boldsymbol{\theta}$ não são independentes (Lema 3.1), $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $(\mathbf{A}+\mathbf{A_4})\boldsymbol{\theta}$ também não são independentes. Então a distribuição de $\mathbf{V}'\mathbf{B}\mathbf{V}$ não é uma convolução de uma combinação linear de variáveis aleatórias com distribuição χ^2 e de uma variável aleatória com distribuição normal.

Se $\mathbf{A}^{\star\star} = \mathbf{A} + \mathbf{A_4}$, então

$$\begin{aligned}\mathbf{V}'\mathbf{B}\mathbf{V} &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}^{\star\star}\boldsymbol{\theta} + \delta^{\star\star} \\ &= (\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}^{\star\star})'(\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}) - \frac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star} \\ &= \mathbf{X}'\mathbf{X} - \frac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star},\end{aligned}$$

onde $\mathbf{X} = \mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}$.

Supondo $\mathbf{B}$ semi-definida positiva temos os seus elementos $\in \mathbb{R}$, como ocorre no caso de amostras balanceadas (Pinheiro *el al.*, 2000). Neste caso, se $\boldsymbol{\Gamma} = \mathbf{B}^{1/2}\boldsymbol{\Sigma}(\mathbf{B}^{1/2})'$ e $\boldsymbol{\mu}^{\star\star} = \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}$, então

$$\mathbf{X} \sim N(\boldsymbol{\mu}^{\star\star}; \boldsymbol{\Gamma}).$$

Seja $\mathbf{P}_4$ uma matriz ortogonal tal que $\mathbf{P}_4\boldsymbol{\Gamma}(\mathbf{P}_4)' = \boldsymbol{\Lambda}$, onde $\boldsymbol{\Lambda}$ é uma matriz diagonal. Se $\mathbf{Y} = \mathbf{P}_4\mathbf{X} \Rightarrow \mathbf{X} = \mathbf{P}_4'\mathbf{X}$, então,

$$\mathbf{Y} \sim N(\mathbf{P}_4\boldsymbol{\mu}^{\star\star}; \boldsymbol{\Lambda})$$

e

$$\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{P}_4\mathbf{P}_4'\mathbf{Y} \sim \sum_{i=1}^{CGK} \lambda_i(\chi_1^2(\delta_i))_i \tag{3.50}$$

onde λ_i são os autovalores de $\boldsymbol{\Lambda}$, neste caso são os elementos da diagonal da matriz $\boldsymbol{\Lambda}$. Note que $\boldsymbol{\Lambda}$ é semi-definida positiva e portanto $\lambda_i \geq 0$. $\delta_i = \dfrac{a_i^2}{2\lambda_i}$, sendo a_i o i-ésimo elemento do vetor $\frac{1}{2}\mathbf{P}_4\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}$, que é uma combinação linear dos γ_{cgk}'s.

Se a constante $c = -\dfrac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star}$, então,

$$\Pr(F_1 \geq u) = \Pr\left(\frac{\sqrt{\mathrm{N}_0}(\mathbf{X}'\mathbf{X} + \mathrm{c})}{\theta_3^{\mathrm{o}}} \geq \mathrm{u}\right) = \Pr(\sqrt{\mathrm{N}_0}\mathbf{X}'\mathbf{X} \geq \theta_3^{\mathrm{o}}\mathrm{u} - \sqrt{\mathrm{N}_0}\mathrm{c}). \tag{3.51}$$

Mas,

$$\begin{aligned}c = -\frac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star} &= O(N_0^{-3/2}N_0^2N_0^{-3/2}) + O(N_0^{-1} + N_0^{-1/2}) \\ &= O(N_0^{-1} + N_0^{-1} + N_0^{-1/2}) \\ &= O(N_0^{-1/2}),\end{aligned}$$

desde que $N_g = O(N_0)$. Então, $\sqrt{N_0}c = O(1)$, com o aumento do parâmetro de não centralidade δ_i a distribuição da variável aleatória χ^2 vai tendendo para a direita e, para $N_0 \to \infty$, a probabilidade em (3.51) tende a 1, indicando que o poder do teste converge para 1.

Nos casos em que $\mathbf{B}$ não é semi-definida positiva, mais estudos precisam ser feitos sobre o poder do teste. No entanto, nestes casos, pode-se calcular o poder numericamente.

Capítulo 4

Aplicações

Como foi visto, uma possível aplicação das análises de variâncias desenvolvidas nos capítulos 2 e 3, seria no estudo de diferenças genéticas entre organismos (genética populacional). Serão apresentados, neste capítulo, conjuntos de dados onde serão aplicados nossos resultados . Na maioria dos casos, o tamanho da amostra não é grande o suficiente para a aplicação dos resultados assintóticos. Para estes dados será utilizado o método *bootstrap* de reamostragem.

O primeiro conjunto de dados representa uma população de cágados da espécie *Hydromedusa maximiliani*. Segundo Souza *et al.* (2000) esta espécie habita rios e riachos de águas rasas e claras de regiões montanhosas e é endêmica da floresta Atlântica. Estes ecossistemas apresentam uma das mais elevadas taxas de destruição e fragmentação e, como esta espécie é altamente sensível a efeitos estocásticos demográficos e ambientais, esta encontra-se ameaçada de extinção. A variabilidade genética entre as populações é o que permite a adaptação à mudanças ambientais.

A segunda aplicação analisa a variabilidade genética da espécie *Dermatobia hominis* (mosca do berne). Esta espécie parasita feridas em animais vivos, principalmente em rebanhos de gado. Os ferimentos provocados por esta mosca na derme do hospedeiro podem causar infecções, prejudicando a saúde destes animais e, consequentemente, diminuindo a produção de leite e carne. As cicatrizes causadas pelos ferimentos na pele do gado também prejudicam o comércio do couro. O estudo da variabilidade genética dessa espécie permite o desenvolvimento de novas formas de controle, como a utilização de inseticidas mais apropriados à particularidades regionais.

4.1 Analisando a Variabilidade entre Marcadores RAPD

O nosso primeiro interesse é comparar grupos de cágados da espécie *Hydromedusa maximiliani* através de marcadores moleculares da classe RAPD, visando investigar diversidade genética nessas populações.

Os cágados em nosso estudo foram coletados em microbacias do Parque Estadual Carlos Botelho, região sul do estado de São Paulo. Devido ao procedimento de coleta destes animais e obtenção de seus marcadores moleculares da classe RAPD, não foi possível obter o mesmo número de indivíduos para os diferentes grupos. Este conjunto possui um total de 44 indivíduos, com 10 loci em cada. Inicialmente foram comparados três grupos correspondentes às microbacias I, II, e III. Cada grupo possui respectivamente 25, 8 e 11 indivíduos. A microbacia I pode ser subdividida em três repartições: 1, 2 e 3. Estes três subgrupos também foram comparados posteriormente. Em cada subgrupo da microbacia I existem respectivamente, 4, 12 e 9 indivíduos.

Sob a hipótese de homogeneidade entre grupos temos $H_0 : p_{1k} = p_{2k} = p_{3k} = p_k$, onde p_k é a probabilidade de aparecimento da banda (fenótipo dominante) na posição k. Como o número de *loci* não é grande o suficiente para aplicarmos o teste assintótico, foi necessário gerar a distribuição empírica da estatística do teste utilizando métodos de reamostragem. Utilizamos o *bootstrap* e o procedimento (para os dados provenientes de marcadores da classe RAPD dos cágados) está descrito abaixo:

Passo 1: Estima-se p_k através dos dados, isto é, $\hat{p}_k = \frac{x_{\cdot\cdot k}}{N_T}$, que é a proporção observada de bandas na posição k para a amostra combinada de N_T indivíduos, e calcula-se o valor observado da estatística F (F_{obs}).

Passo 2: Geram-se $N_T = 44$ vetores aleatórios de dimensão $K = 10$, onde cada um dos K elementos é gerado a partir da distribuição Bernoulli ($\hat{p}_k$).

Passo 3: Calcula-se o valor da estatística F através dos dados gerados.

Passo 4: Repetem-se os passos 2 e 3 10.000 vezes.

Usando o procedimento descrito acima, geramos a distribuição empírica de F com o auxílio do software *MATLAB*. O $p - valor$ será o número total das estatísticas F cujos

valores são maiores que o valor observado, divido por 10.000, isto é,

$$p - valor = \frac{\#F's \geq F_{obs}}{10.000}.$$

A Figura 4.1 mostra o comportamento assintótico da distribuição da estatística F, dada em (2.26), para comparar as microbacias I, II, e III.

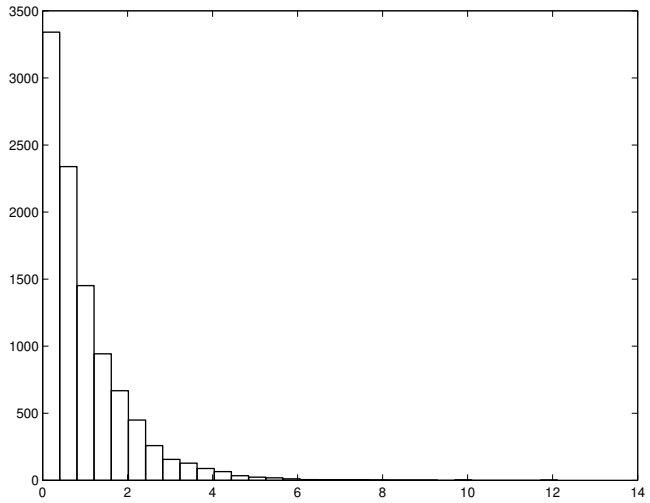

Figura 4.1: Distribuição empírica de F gerada a partir dos dados de RAPD dos cágados (Microbacias I, II, e III).

Através dos valores observados obtivemos uma estimativa $F_{obs} = 0,2702$, e um $p - valor = 0,7625$, indicando-nos que não há diferença entre as microbacias.

Para as 3 repartições dentro da microbacia I obtivemos $F_{obs} = 0,5251$ e $p - valor = 0,5839$. Este resultado mostra que também não existe diferença nas proporções de fenótipos dominantes entre os cágados das repartições 1, 2 e 3. O comportamento da distribuição neste caso pode ser observado na Figura 4.2.

O outro conjunto de dados analisado constitui-se de seis populações de moscas do berne. Um total de 62 indivíduos foi obtido em diferentes regiões dos estados de São Paulo (SP), Goiás (GO), Minas Gerais (MG) e Paraná (PR). O número de indivíduos amostrados de acordo com cada localidade foi:

7 – em Presidente Prudente (SP),

11 – em Pirassununga (SP),

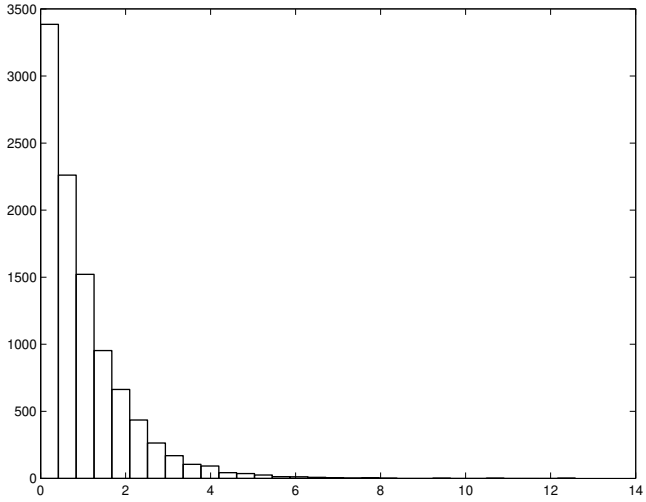

Figura 4.2: Distribuição empírica de F gerada a partir dos dados de RAPD dos cágados (Repartições 1, 2 e 3 da Microbacia I).

11 – em São Carlos (SP),
11 – em Anápolis (GO),
11 – em Brasópolis (MG) e
11 – em Ponta Grossa (PR).

Para cada indivíduo na amostra foram isolados 123 *loci*.

O mesmo procedimento realizado com os marcadores moleculares da classe RAPD dos cágados foi utilizado para este exemplo. Neste caso, $N_T = 62$ e $K = 123$.

Para compararmos estes seis grupos também observamos o comportamento da distribuição da estatística F (Figura 4.3) e obtivemos $F_{obs} = 0,5416$ e um $p - valor = 0,7574$. Para este conjunto de dados também aplicamos diretamente nossos resultados teóricos assintóticos onde a estatística do teste $F \sim F_{5,56}$ e portanto $p - valor = \Pr(F > 0,5416) \approx 0,744$. Logo, não existem evidências para rejeitarmos H_0.

De acordo com os resultados, neste exemplo também não rejeitamos a hipótese de homogeneidade entre as seis populações de moscas do berne.

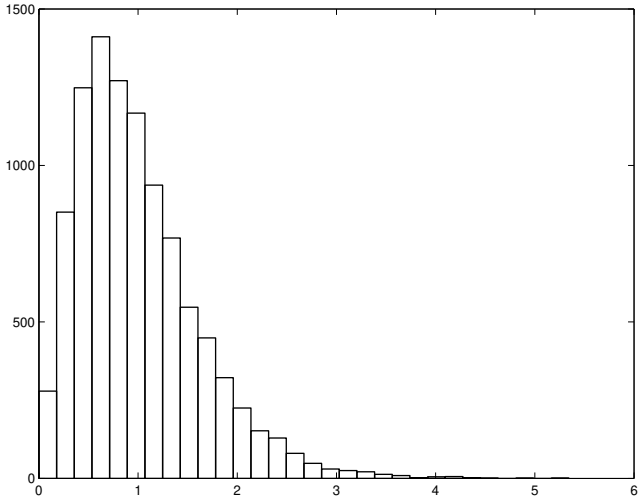

Figura 4.3: Distribuição empírica de F gerada a partir dos dados de RAPD das moscas do berne.

4.2 Analisando a Variabilidade entre Seqüências de DNA

Agora, iremos comparar as populações de cágados, as mesmas utilizadas na análise de marcadores moleculares da classe RAPD, através de seqüências de DNA. A molécula de DNA é constituída por duas cadeias complementares de nucleotídeos, por isso dizemos que as cadeias se encontram em orientação oposta. No processo de extração do DNA o seqüenciamento de nucleotídeos é feito nos dois sentidos. Podemos dizer que o seqüenciamento de uma cadeia dá início, a partir de uma posição, da esquerda para a direita, enquanto que na cadeia complementar, o seqüenciamento se origina da direita para a esquerda. Portanto, neste processo foram obtidas duas seqüências de nucleotídeos do DNA mitocondrial dos cágados, as quais representam duas regiões diferentes do genoma, o gene citocromo b e a região de controle.

O conjunto de dados possui 48 seqüências do DNA mitocondrial sendo que os cágados capturados nas Microbacias I, II, e III possuem 30, 7 e 11 seqüências, respectivamente. A diferença no tamanho das amostras deve-se ao fato de que algumas seqüências foram perdidas tanto no processo de seqüênciamento do DNA quanto no de RAPD . Foram identificadas 262 posições no gene citocromo b e 413 na região de controle. Como as

seqüências representam duas regiões diferentes do genoma mitocondrial iremos analisá-las separadamente. Sob $H_0 : p_{cgk} = p_{ck}$, onde p_{ck} é a proporção de seqüências que na posição k tem a categoria c, onde $c \in \{\mathbf{A},\ \mathbf{T},\ \mathbf{C},\ \mathbf{G}\}$. Inicialmente, comparamos os grupos correspondentes às três microbacias de seqüências do gene citocromo b e, em seguida, analisamos as seqüências da região de controle. Estas comparações também foram feitas para as três repartições dentro da Microbacia I. Tendo em vista o pequeno tamanho de amostra, recorremos à métodos de reamostragem afim de gerar a distribuição empírica da estatítica F_1 sob H_0. O procedimento consiste no seguinte:

Passo 1: Estima-se p_{ck} através dos dados, isto é, $\hat{p}_{ck} = \dfrac{n_{c1k} + n_{c2k} + n_{c3k}}{N_T}$, que é a proporção observada de seqüências que na posição k tem a categoria c em toda a amostra (desconsiderando os grupos), e calcula-se o valor observado da estatística F_1 (F_{1obs}).

Passo 2: Geram-se $N_T = 48$ seqüências com $K = 262$ (no caso das seqüências do gene citocromo b) e 413 (no caso das seqüências da região de controle) posições cada a partir de uma distribuição Multinomial $(48;\ \hat{p}_{1k},\ \hat{p}_{2k}, \hat{p}_{3k}, \hat{p}_{4k})$.

Passo 3: Calcula-se o valor da estatística F_1 através dos dados gerados.

Passo 4: Repetem-se os passos 2 e 3 10.000 vezes.

A Tabela 4.1 mostra os valores observados da estatística F_1, dada em (3.49), e os respectivos $p - valores$ correspondentes às seqüências do gene citocromo b e da região de controle do DNA mitocontrial das populações de cágados.

Tabela 4.1: Valores observados de F_1 e p-valores : populações de cágados nas Microbacias I, II e III.

Seqüência	F_{1obs}	$p - valor$
Gene citocromo b	$0,0000$	$0,4923$
Região de controle	$0,0003$	$0,0025$

Os histogramas da Figura 4.4 mostram o comportamento da distribuição da estatística F_1.

Na Figura 4.5, observamos o comportamento da distribuição da estatítica F_1 para as

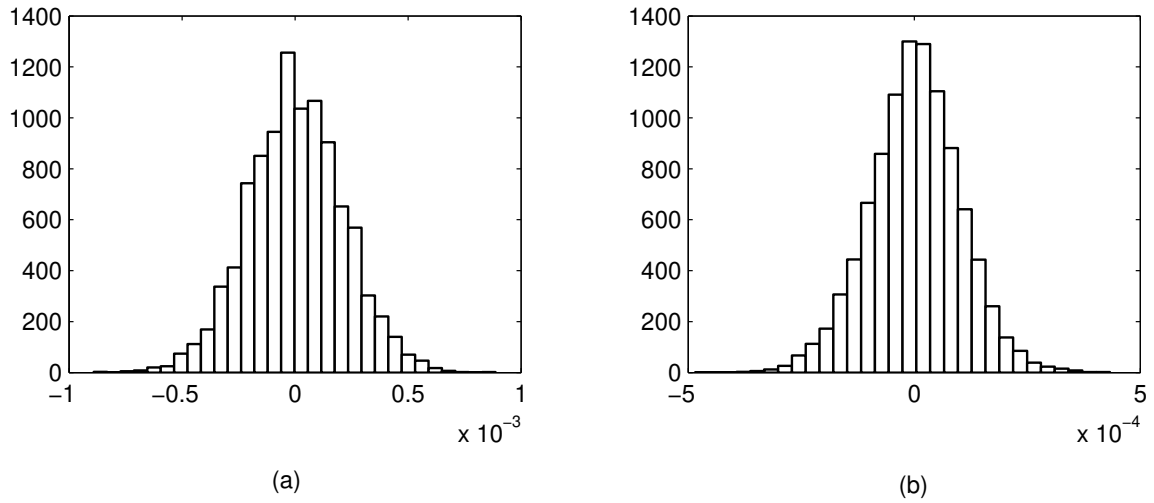

Figura 4.4: Distribuição empírica de F_1: Seqüências de DNA das populações de cágados. (a) Gene citocromo b. (b) Região de controle.

subpopulações de cágados dentro da Microbacia I. Os valores observados da estatística F_1 e os respectivos $p - valores$ correspondentes às seqüências do gene citocromo b e da região de controle estão na Tabela 4.2.

Tabela 4.2: Valores observados de F_1 e p-valores: subpopulações de cágados dentro da Microbacia I.

Seqüência	F_{1obs}	$p - valor$
Gene citocromo b	$0,0000$	$0,8112$
Região de controle	$-0,0003$	$0,0002$

Para o nível se significância de 5%, rejeitamos a hipótese de homogeneidade contra a hipótese de heterogeneidade entre os grupos se $p - valor \leq 0.05$. Se o valor observado da estatística F_1 for negativo então $p - valor = 2P(F_1 \leq F_{1obs})$, caso contrário, $p - valor = 2P(F_1 \geq F_{1obs})$.

Entre as três microbacias encontramos fortes evidências para rejeitarmos a hipótese de homogeneidade entre os grupos, isto é, analisando a seqüência da região de controle dos cágados, encontramos evidências estatísticas de que há variação genética entre as três

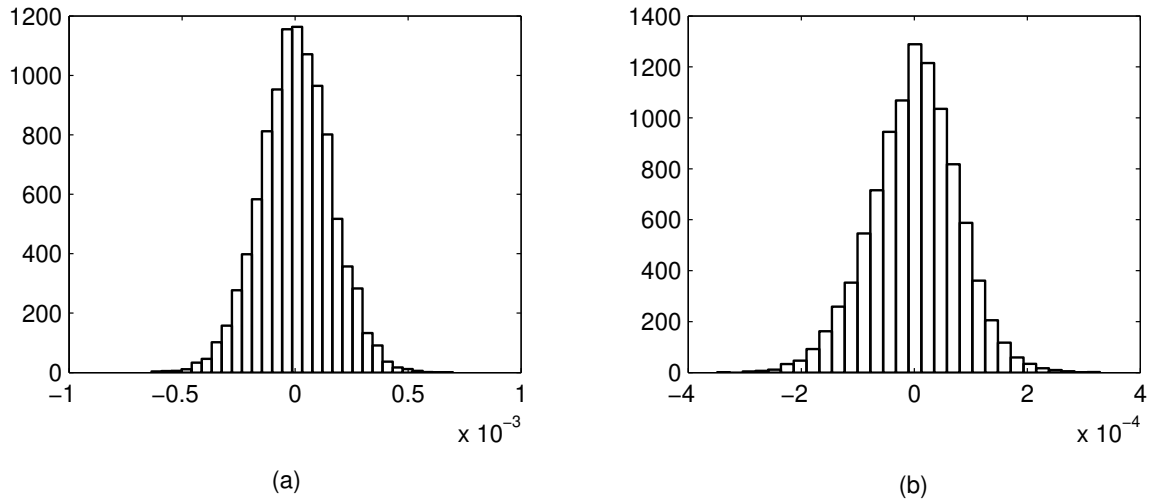

Figura 4.5: Distribuição empírica de F_1: Seqüências de DNA das subpopulações de cágados da Microbacia I. (a) Gene citocromo *b*. (b) Região de controle.

populações de cágados das Microbacias I, II e III.

Comparando as seqüências de DNA da região de controle das subpopulações de cágados dentro da Microbacia I, também obtivemos fortes evidências para rejeitarmos a hipótese nula de homogeneidade entre os grupos. Portanto, ao nível de seqüências de DNA foi possível observar variação genética entre os subgrupos de cágados da Microbacia I.

Uma análise mais específica, comparando os grupos dois a dois, nos mostra em quais microbacias as populações de cágados se diferem geneticamente, assim como, quais os diferentes subgrupos de cágados correspondentes às repartições 1, 2 e 3, dentro da Microbacia I. As Tabelas 4.3 e 4.4 mostram os valores observados da estatística F_1 e os p-valores referentes às Microbacias e às repartições, respectivamente.

Tabela 4.3: Valores observados de F_1 e p-valores: populações de cágados das Microbacias I, II e III, comparadas duas a duas.

Microbacias	F_{1obs}	$p - valor$
I e II	$0,0004$	$0,0070$
I e III	$0,0001$	$0,1955$
II e III	$0,0001$	$0,0270$

Tabela 4.4: Valores observados de F_1 e p-valores: subpopulações de cágados da Microbacia I comparadas duas a duas.

Repartições	F_{1obs}	$p - valor$
1 e 2	$-0,0005$	$< 0,0002$
1 e 3	$-0,0002$	$0,0136$
2 e 3	$-0,0001$	$0,1330$

Para que possamos obter as afirmações referentes a quais grupos se diferenciam geneticamente, comparando-os dois a dois, devemos utilizar a correção de Bonferroni. Para estas comparações, a um nível de significância de 5%, rejeitamos a hipótese de homogeneidade entre os grupos se $p - valor < (0,05/3) \approx 0,017$.

Portanto, há fortes evidências de que existe diversidade genética entre as microbacias I e II. A diversidade também é evidente entre as repartições 1 e 2 e entre 1 e 3.

O comportamento da distribuição de F_1 para estes casos pode ser observado nas Figuras 4.6 e 4.7.

Os resultados mostrados nesta seção também foram gerados no software *MATLAB*.

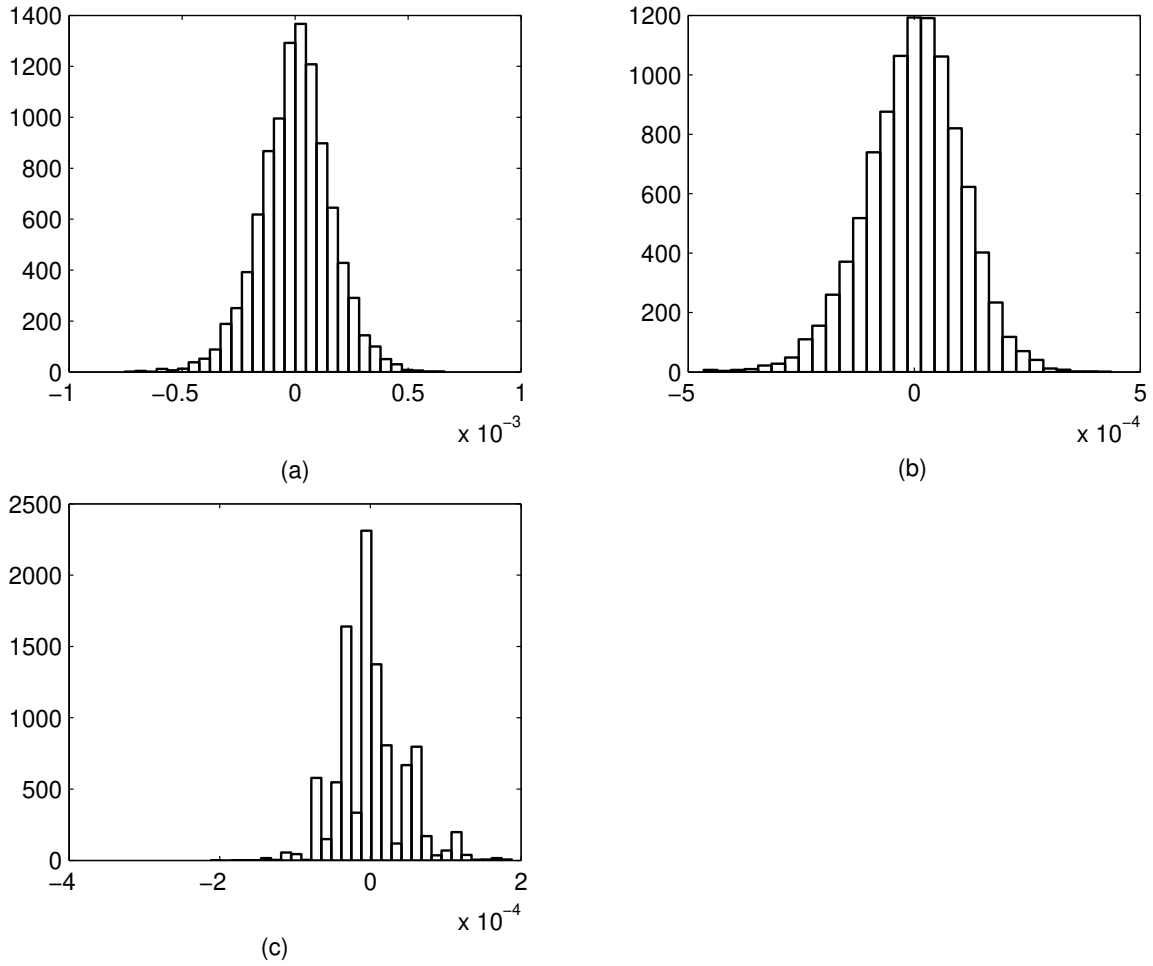

Figura 4.6: Distribuição empírica de F_1: Seqüências de DNA da região de controle das populações de cágados. (a) Microbacias I e II. (b) Microbacias I e III. (c) Microbacias II e III.

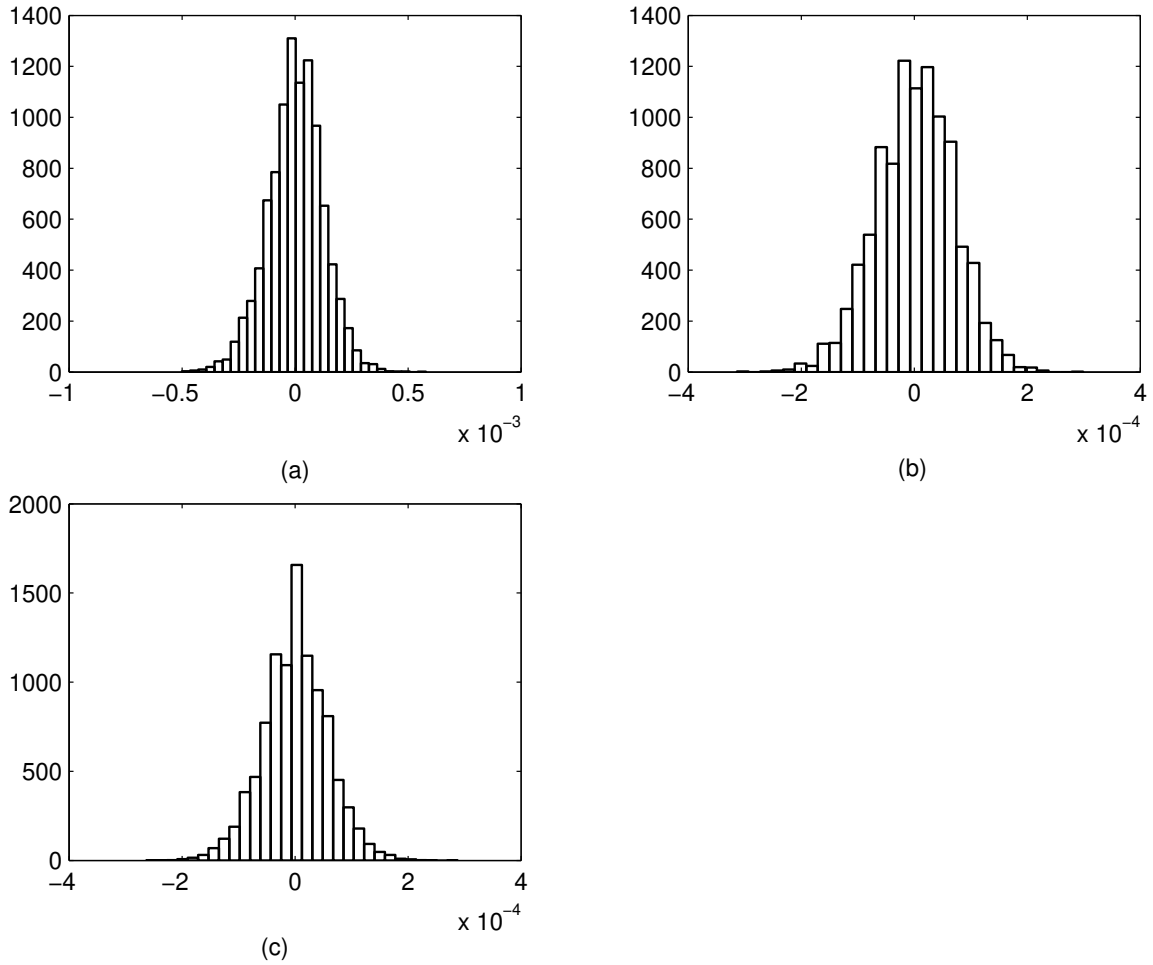

Figura 4.7: Distribuição empírica de F_1: Seqüências de DNA da região de controle das subpopulações de cágados da Microbacia I. (a) Repartições 1 e 2. (b) Repartições 1 e 3. (c) Repartições 2 e 3.

Capítulo 5

Considerações Finais

Através das medidas de variação entre dados categorizados, a *heterozigosidade* e o Índice de Simpson, foi possível desenvolver análises de variâncias e testar a hipótese de homogeneidade entre grupos em amostras não balanceadas.

No Capítulo 2 obtivemos a estatística do teste e sua distribuição assintótica ($K \to \infty$) considerando independência entre as populações, entre os indivíduos e entre os *loci*. Este teste se torna mais poderoso quanto maior o número de indivíduos nos grupos ($N_0 \to \infty$). O efeito do *locus* foi incorporado no termo residual pois, no presente contexto, não faz sentido biológico testá-lo. No entanto, em outras ocasiões, este efeito pode ter grande importância, o que poderá ser pesquisado futuramente.

Na Seção 3.4.2 mostramos dois conjuntos de autovalores correspondentes às seqüências de DNA de duas regiões do DNA mitocondrial (gene citocromo b e região de controle) dos cágados das Microbacias I, II e III. Para obtermos os autovalores, foram consideradas somente algumas posições ao longo de cada seqüência, 50 no gene citocromo b e 61 na região de controle, pois o procedimento de obtenção dos autovalores exige um custo computacional alto, o que inviabiliza a utilização de todas as posições seqüenciadas, 262 e 413, respectivamente. Portanto, mesmo que tivéssemos números de indivíduos grandes o suficiente para aplicarmos os resultados assintóticos do Capítulo 3, poderíamos ter problemas com o método numérico que utilizamos na obtenção dos autovalores. Sendo assim, recomandamos a utilização de técnicas de reamostragem e, o *bootstrap* mostrou-se adequado. O estudo de métodos numéricos mais apropriados pode ser uma alternativa para este problema.

Na Seção 3.5 investigamos o poder do teste para o caso particular em que a matriz $\mathbf{B}$ é semi-definida positiva, o que garante sua fatoração em matrizes de elementos reais. Para o caso em que as amostras são não balanceadas não, necessariamente, temos esta condição,

como ocorre nos casos onde as amostras são balanceadas (Pinheiro *et al.*, 2000). Portanto, ao se fazer a fatoração $\mathbf{B} = (\mathbf{B}^{1/2})'\mathbf{B}^{1/2}$, $\mathbf{B}^{1/2}$ poderá apresentar elementos complexos, o que não garante o resultado em (3.50). Como foi enfatizado, estudos numéricos do poder devem ser feitos, nestes casos. Também podemos pensar no estudo de variáveis aleatórias normais complexas como uma possível solução para este problema.

Os resultados apresentados no Capítulo 4 mostram, entre outros, os comportamentos das distribuições empíricas das estatísticas F e F_1 propostas, respectivamente, nos Capítulos 2 e 3, para testarmos a hipótese nula de homogeneidade entre grupos em amostras não balanceadas, não contradizendo nossos resultados teóricos.

Observou-se que, ao comparar grupos de marcadores moleculares da classe RAPD, não foram encontradas evidências suficientes para rejeitarmos a hipótese nula. Quando comparamos as seqüências a nível de nucleotídeo, evidenciamos a variabilidade genética entre as seqüências da região de controle do DNA mitocondrial dos cágados. Isto comprova o fato de que, embora nos permitam detectar polimorfismos genéticos e por um custo menor, os marcadores de RAPD são menos abrangentes que as seqüências de nucleotídeos.

Apêndice A

A.1 Prova do Teorema 2.3

A função geratriz de momentos de uma variável aleatória Y com distribuição χ^2 não-central, com n graus de liberdade e parâmetro de não-centralidade δ, ou seja, $Y \sim \chi^2_n(\delta)$ é dada por:

$$M_Y(t) = (1-2t)^{-\frac{1}{2}n} e^{-\delta[1-(1-2t)^{-1}]} \qquad \text{(Searle, 1971, p49)}$$

Segundo Searle (1971, p57), se $\mathbf{X}$ é um vetor aleatório $n \times 1$, $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{V}$ não singular, e $\mathbf{A}$ uma matriz $n \times n$ de elementos determinísticos, então a função geratriz de momentos da variável aleatória $\mathbf{X'AX}$ é dada por:

$$M_{\mathbf{X'AX}}(t) = \prod_{i=1}^{n}(1-2t\lambda_i)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'\left[-\sum_{k=1}^{\infty}(2t)^k(\mathbf{AV})^k\right]\mathbf{V}^{-1}\boldsymbol{\mu}\right\}.$$

Sendo $\mathbf{A}$ e $\mathbf{V}$ matrizes diagonais, $\mathbf{AV}$ é uma matriz diagonal cujos elementos da diagonal são λ_i, $i = 1, \ldots, n$, portanto $(\mathbf{AV})^k$ é uma matriz diagonal cujos elementos da diagonal são λ_i^k. Então, $-\sum_{k=1}^{\infty}(2t)^k(\mathbf{AV})^k$ é também uma matriz diagonal onde os elementos da diagonal são

$$-\sum_{k=1}^{\infty}(2t\lambda_i)^k = 1-(1-2t\lambda_i)^{-1}, \quad \text{desde que } |t\lambda_i| < 1, \quad i = 1, \ldots, n.$$

E portanto, como $\mathbf{V}$ é não singular, cujos elementos da diagonal são ν_i, $i = 1, \ldots, n$,

$$\boldsymbol{\mu}'\left[-\sum_{k=1}^{\infty}(2t)^k(\mathbf{AV})^k\right]\mathbf{V}^{-1}\boldsymbol{\mu} = \sum_{i=1}^{n}\frac{\mu_i^2}{\nu_i}[1-(1-2t\lambda_i)^{-1}].$$

$$
\begin{aligned}
M_{\mathbf{X}'\mathbf{A}\mathbf{X}}(t) &= \prod_{i=1}^{n}(1-2t\lambda_i)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{\nu_i}[1-(1-2t\lambda_i)^{-1}]\right\} \\
&= \prod_{i=1}^{n}M_{Y_i}(t\lambda_i) = \prod_{i=1}^{n}M_{\lambda_i Y_i}(t),
\end{aligned}
$$

onde $Y_i \sim \chi_1^2(\delta_i)$, sendo $\delta_i = \dfrac{\mu_i^2}{2\nu_i}$.

■

A.2 Prova: F é semi-definida positiva

Demonstração. De (2.7) temos que $\mathbf{F} = K\mathbf{F}^{\star}$, onde $\mathbf{F}^{\star}$ é uma matriz simétrica $G \times G$ cujos elementos são

$$
f^{\star}(g,g) = N_g\left(1-\frac{N_g}{N_T}\right) \quad \text{e} \quad f^{\star}(g,g') = -\frac{N_g N_{g'}}{N_T} \tag{1.1}
$$

ou seja, $\mathbf{F} = \dfrac{K}{N_T}\mathbf{F}^{\star\star}$, onde $\mathbf{F}^{\star\star}$ é uma matriz simétrica de elementos

$$
f^{\star\star}(g,g) = N_g\sum_{j\neq g}N_j \quad \text{e} \quad f^{\star\star}(g,g') = -N_g N_{g'} \quad g' \neq g \tag{1.2}
$$

Seja $\mathbf{x} = (x_1, x_2, \ldots, x_G)'$, $\mathbf{x} \neq \mathbf{0}$, um vetor coluna de tamanho G, então

$$
\mathbf{x}'\mathbf{F}\mathbf{x} = \frac{K}{N_T}\mathbf{x}'\mathbf{F}^{\star\star}\mathbf{x},
$$

$\mathbf{x}'\mathbf{F}^{\star\star}$ é um vetor linha de tamanho G cujo elemento i, $i = 1, \ldots, G$, é :

$$
x_i N_i\sum_{j\neq i}N_j - N_i\sum_{j\neq i}x_j N_j
$$

Portanto,

$$
\begin{aligned}
\mathbf{x}'\mathbf{F}^{\star\star}\mathbf{x} &= \sum_i x_i \left(x_i N_i \sum_{j\neq i} N_j - N_i \sum_{j\neq i} x_j N_j \right) \\
&= \sum_i \sum_{j\neq i} \left(N_i N_j x_i^2 - N_i N_j x_i x_j \right) \\
&= \sum_i \sum_{j\neq i} N_i N_j (x_i^2 - x_i x_j) \\
&= \sum_i \sum_{j>i} N_i N_j (x_i - x_j)^2 \geq 0
\end{aligned}
$$

□

A.3 Prova: $\mathbf{F_2}$ é semi-definida positiva

Demonstração. Temos que $\mathbf{F_2} = K\mathbf{F_2^\star}$ é uma matriz simétrica, $N_T \times N_T$, onde $\mathbf{F_2^\star}$ é definida em (2.19). Seja $\mathbf{x} = (x_1, x_2, \ldots, x_{N_T})'$, $\mathbf{x} \neq \mathbf{0}$, um vetor coluna de tamanho N_T. Do Lema (2.3) temos que $\mathbf{F_2^\star}$ é idempotente, portanto,

$$
\begin{aligned}
\mathbf{x}'\mathbf{F_2}\mathbf{x} &= K\mathbf{x}'\mathbf{F_2^\star}\mathbf{x} = K\mathbf{x}'\mathbf{F_2^\star}\mathbf{F_2^\star}\mathbf{x} \\
&= K\mathbf{x}'(\mathbf{F_2^\star})'\mathbf{F_2^\star}\mathbf{x} = K(\mathbf{F_2^\star}\mathbf{x})'(\mathbf{F_2^\star}\mathbf{x}) \geq 0
\end{aligned}
$$

□

Apêndice B

B.1 Prova do Lema 3.1

$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $\mathbf{A}\boldsymbol{\theta}$ seriam independentes se e somente se

$$\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{B} = 0 \quad \text{(Searle, 1971).}$$

$$\begin{aligned}
\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{B} &= \mathbf{A}(\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^{\diamond})\left[-\frac{1}{(KN_T)^2}\mathbf{U}_{KG}+(\mathbf{M}^{-1}\otimes\mathbf{U}_K\right]\otimes\mathbf{I}_C \\
&= -\frac{1}{(KN_T)^2}\mathbf{A}(\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^{\diamond})(\mathbf{U}_{KG}\otimes\mathbf{I}_C)+\mathbf{A}(\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^{\diamond})(\mathbf{M}^{-1}\otimes\mathbf{U}_K\otimes\mathbf{I}_C) \\
&= -\frac{1}{(KN_T)^2}\mathbf{A}\left[\boldsymbol{\eta}\mathbf{U}_G\otimes\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right]+\mathbf{A}\left[\boldsymbol{\eta}\mathbf{M}^{-1}\otimes\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&= -\frac{1}{(KN_T)^2}(\mathbf{a}\otimes\mathbf{A}^{\star})\left[\boldsymbol{\eta}\mathbf{U}_G\otimes\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&+ (\mathbf{a}\otimes\mathbf{A}^{\star})\left[\boldsymbol{\eta}\mathbf{M}^{-1}\otimes\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&= -\frac{1}{(KN_T)^2}\left[\mathbf{a}\boldsymbol{\eta}\mathbf{U}_G\otimes\mathbf{A}^{\star}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&+ \left[\mathbf{a}\boldsymbol{\eta}\mathbf{M}^{-1}\otimes\mathbf{A}^{\star}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&= \left[\mathbf{a}\boldsymbol{\eta}\left(-\frac{1}{(KN_T)^2}\mathbf{U}_G+\mathbf{M}^{-1}\right)\right]\otimes\left[\mathbf{A}^{\star}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right]
\end{aligned}$$

Seja $\mathbf{a}^{\star} = (p_{1\cdot}\ldots p_{C\cdot})$. Lembremos que $\boldsymbol{\Sigma}_0^{\diamond} = \boldsymbol{\Sigma}_{01}^{\diamond}\oplus\boldsymbol{\Sigma}_{02}^{\diamond}\oplus\ldots\oplus\boldsymbol{\Sigma}_{0K}^{\diamond}$.

$$\mathbf{A}^{\star}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C) = \frac{2}{GK^2}\left(\mathbf{a}^{\star}\Sigma_{01}^{\diamond}\ \mathbf{a}^{\star}\Sigma_{02}^{\diamond}\ \ldots\ \mathbf{a}^{\star}\Sigma_{0K}^{\diamond}\right)(\mathbf{U}_K\otimes\mathbf{I}_C).$$

Para cada k, de (3.29) temos

$$\mathbf{a}^{\star}\Sigma_{0k}^{\diamond} = \left(p_{1k}\left(p_{1\cdot} - \sum_{c=1}^{C} p_{c\cdot}p_{ck}\right)\ p_{2k}\left(p_{2\cdot} - \sum_{c=1}^{C} p_{c\cdot}p_{ck}\right)\ \ldots\ p_{Ck}\left(p_{C\cdot} - \sum_{c=1}^{C} p_{c\cdot}p_{ck}\right)\right).$$

O primeiro elemento do vetor $\mathbf{A}^{\star}\mathbf{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)$ é

$$\frac{2}{GK^2}\left(p_{11}\left(p_{1\cdot} - \sum_{c} p_{c\cdot}p_{c1}\right) + p_{12}\left(p_{1\cdot} - \sum_{c} p_{c\cdot}p_{c2}\right) + \ldots + p_{1K}\left(p_{1\cdot} - \sum_{c} p_{c\cdot}p_{cK}\right)\right)$$

$$= \frac{2}{GK^2}\left(\sum_{k} p_{1k}\left(p_{1\cdot} - \sum_{c} p_{c\cdot}p_{ck}\right)\right) = \frac{2}{K^2}\left(p_{1\cdot} - \sum_{c} p_{c\cdot}\sum_{k} p_{ck}\right) \neq 0.$$

Como $\mathbf{a}\boldsymbol{\eta}\left(-\frac{1}{(KN_T)^2}\mathbf{U}_G + \mathbf{M}^{-1}\right) \neq 0 \Rightarrow \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $\mathbf{A}\boldsymbol{\theta}$ não são independentes.

■

B.2 Prova do Lema 3.2

$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ e $\mathbf{A_2}\boldsymbol{\theta}$ seriam independentes se e somente se

$$\mathbf{A_2}\mathbf{\Sigma}_0\mathbf{T} = 0 \quad \text{(Searle, 1971).}$$

$$\begin{aligned}
\mathbf{A_2}\mathbf{\Sigma}_0\mathbf{T} &= \frac{1}{(KN_T)^2}\mathbf{A_2}(\boldsymbol{\eta} \otimes \mathbf{\Sigma}_0^{\diamond})(\mathbf{U}_{KG} \otimes \mathbf{I}_C) \\
&= \frac{1}{(KN_T)^2}\mathbf{A_2}(\boldsymbol{\eta} \otimes \mathbf{\Sigma}_0^{\diamond})(\mathbf{U}_G \otimes \mathbf{U}_K \otimes \mathbf{I}_C) \\
&= \frac{1}{(KN_T)^2}\mathbf{A_2}[\boldsymbol{\eta}\mathbf{U}_G \otimes \mathbf{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)] \\
&= \frac{1}{(KN_T)^2}(\mathbf{1}_G \otimes \mathbf{A_2^{\star}})[\boldsymbol{\eta}\mathbf{U}_G \otimes \mathbf{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)] \\
&= \frac{1}{(KN_T)^2}(\mathbf{1}_G\boldsymbol{\eta}\mathbf{U}_G \otimes \mathbf{A_2^{\star}}\mathbf{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)).
\end{aligned}$$

Do Lema 3.1 temos

$$\mathbf{A_2^{\star}}\mathbf{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)) = \frac{2}{K^2N_T}\left(\mathbf{a}^{\star}\Sigma_{01}^{\diamond}\ \mathbf{a}^{\star}\Sigma_{02}^{\diamond}\ \ldots\ \mathbf{a}^{\star}\Sigma_{0K}^{\diamond}\right)(\mathbf{U}_K \otimes \mathbf{I}_C) \neq 0.$$

Como $\mathbf{1}_G\boldsymbol{\eta}\mathbf{U}_G \neq 0 \Rightarrow \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ e $\mathbf{A_2}\boldsymbol{\theta}$ não são independentes.

■

B.3 Prova do Lema 3.3

$\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta}$ e $\mathbf{A_3}\boldsymbol{\theta}$ seriam independentes se e somente se

$$\mathbf{A_3}\boldsymbol{\Sigma}_0\mathbf{W} = 0 \quad \text{(Searle, 1971).}$$

$$\begin{aligned}
\mathbf{A_3}\boldsymbol{\Sigma}_0\mathbf{W} &= \mathbf{A_3}(\boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^{\diamond})[(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C] \\
&= (\mathbf{n_3} \otimes \mathbf{A_3^{\star}})(\boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^{\diamond})(\mathbf{M}^{-1} \otimes \mathbf{U}_K \otimes \mathbf{I}_C) \\
&= (\mathbf{n_3} \otimes \mathbf{A_3^{\star}})[\boldsymbol{\eta}\mathbf{M}^{-1} \otimes \boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)] \\
&= \mathbf{n_3}\boldsymbol{\eta}\mathbf{M}^{-1} \otimes \mathbf{A_3^{\star}}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C).
\end{aligned}$$

Do lema 3.1 temos

$$\mathbf{A_3^{\star}}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K \otimes \mathbf{I}_C)) = \frac{2}{K^2 G}\left(\mathbf{a}^{\star}\Sigma_{01}^{\diamond}\ \mathbf{a}^{\star}\Sigma_{02}^{\diamond}\ \ldots\ \mathbf{a}^{\star}\Sigma_{0K}^{\diamond}\right)(\mathbf{U}_K \otimes \mathbf{I}_C) \neq 0.$$

Como $\mathbf{n_3}\boldsymbol{\eta}\mathbf{M}^{-1} \neq 0 \Rightarrow \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ e $\mathbf{A_3}\boldsymbol{\theta}$ não são independentes.

■

Bibliografia

[1] Andrade M. e Pinheiro H. P. (2002). *Métodos Estatísticos Aplicados em Genética Humana*. 15 SINAPE - ABE-Associação Brasileira de Estatística.

[2] Farah S. B. (1997). *DNA Segredos & Mistérios*. Sarvier Editora de Livros Médicos Ltda.

[3] Ferreira M. E. e Grattapaglia D. (1998) *Introdução ao Uso de Marcadores Moleculares em Análise Genética*. EMBRAPA-CENARGEN.

[4] Gini C. W. (1912). Variabilita e Mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari* 3(2), 3-159.

[5] Gonick L. e Wheelis M. (1991). *The Cartoon Guide to Genetics*. Harper Collins Publishers, New York.

[6] Hartl D. L. (2000). *A Primer Of Population Genetics*. Sinaur Associates.

[7] Hoeffding W. (1948). A class of statistics with asymptotically normal distribution. *Ann. of Math. Stat.* 19, 293-325.

[8] James B. R. (1996). *Probabilidade: um curso em nível intermediário*. Projeto Euclides.

[9] Light R. J. e Margolin B. H. (1971). An Analysis of Variance for Categorical Data. *Jornal of the American Statistical Associates* 66, 534-544.

[10] Lodish H., Berk A., Zipursky S. L., Matsudaira P., Baltimore D. e Darnell J. (2000). *Molecular Cell Biology*. W. H. Freeman and Company.

[11] Neter J., Wasserman W., Kutner M. H. (1990). *Applied Linear Statistical Models (Regression, Analysis of Variance, and Experimental Designs)*. Irwin.

[12] Pinheiro H. P., Seillier-Moiseiwitsch F. e Sen P.K. (2001). *Analysis of Variance for Hamming Distances Applied to Unbalanced Designs*. Relatório de Pesquisa 13, IMECC-UNICAMP.

[13] Pinheiro H. P. (1997). *Modelling Variability in the HIV Genome*. Tese de PhD, University of North Carolina, Mimeo Series No. 2186T

[14] Pinheiro H. P., Seillier-Moiseiwitsch F., Sen P. K. e Eron Jr. J. (2000). Genomic Sequences and Quasi-Multivariate CATANOVA. *Handbook of Statistics* 18, 713-746.

[15] Randles R. H. e Wolfe D. A. (1991). *Introduction to The Theory of Nonparametric Statistics*. Krieger Publishing Company.

[16] Rao C. R. (1973) *Linear Statistical Inference And Its Aplications*. John While & Sons.

[17] Searle S. L. (1971) *Linear Models*. John While & Sons.

[18] Sen P. K. e Singer J. M. (1993). *Large Sample Methods In Statistics. An Introduction Whith Aplication*. Chapman & Hall.

[19] Simpson E. H. (1949). The Measurement of Diversity. *Nature* 163, 688.

[20] Souza F. L., Cunha A. F., Oliveira M. A., Pereira G. A. G., Pinheiro H. P. e Reis S. F. (2002). Partitioning of molecular variation on at local spatial scales in the vulnerable neotropical freshwater turtle, *Hydromedusa maximiliani* (Testudines, Chelidae): implications for the conservation of aquatic organisms in natural hierarchical systems. *Biological Conservation* 104, 119-126.

[21] Weir B. (1990). *Genetic Data Analysis*. Sinauer