Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Análise de variação e estrutura populacional em *loci* de microsatélites baseada em distâncias genéticas

Tatiana Buratto Bordin Taglianetti
Orientador: Profa. Dra. Hildete Prisco Pinheiro

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Estatística.

Campinas - SP
2007

Análise de variação e estrutura populacional em *loci* de microsatélites baseada em distâncias genéticas

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Tatiana Buratto Bordin e aprovada pela comissão julgadora.

Campinas, 13 de março de 2007.

Hildete Pinheiro

Profa. Dra. Hildete Prisco Pinheiro
Orientadora

Banca examinadora:

1. Profa. Dra Hildete Prisco Pinheiro (Orientador) - IMECC/UNICAMP
2. Profa. Dra. Júlia Pavan Soler - IME/USP
3. Prof. Dr. Filidor Vilca Labra - IMECC/UNICAMP

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requesito parcial para obtenção do Título de Mestre em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Miriam Cristina Alves – CRB8a / 5094

Taglianetti, Tatiana Buratto Bordin

T128a Análise de variação e estrutura populacional em *loci* de microsatélites baseadas em distâncias genéticas/Tatiana Buratto Bordin Taglianetti -- Campinas, [S.P. :s.n.], 2007.

Orientadora : Hildete Prisco Pinheiro

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Análise de variância. 2. Microsatélites (Genética). 3. População – Aspectos genéticos. 4. Diversidade. 5. Genética. I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

(mca/imecc)

Título em inglês: Analysis of variation and population structure in microsatellite loci based on genetics distances

Palavras-chave em inglês (Keywords): 1. Analysis of variance. 2.Microsatellites (Genetics). 3. Population – Aspects of genetics. 4. Diversity. 5. Genetic.

Área de concentração: Bioestatística

Titulação: Mestrado em Estatística

Banca examinadora: Profa. Dra. Hildete Prisco Pinheiro (IMECC-Unicamp)
Profa. Júlia Pavan Soler (IME-USP)
Prof. Dr. Filidor Vilca Labra (IMECC-Unicamp)

Data da defesa: 13/02/2007

Programa de Pós-Graduação: Mestre em Estatística

Dissertação de Mestrado defendida em 13 de fevereiro de 2007 e aprovada pela Banca Examinadora composta pelos Profs. Drs.

Hildete Pinheiro

Prof (a). Dr (a). HILDETE PRISCO PINHEIRO

Julia Maria Pavan Soler

Prof (a). Dr (a). JULIA MARIA PAVAN SOLER

Prof (a). Dr (a). FILIDOR EDILFONSO VILCA LABRA

Aos meus avôs Joaquim e Beatriz (in memorian),
aos meus pais, Benjamin e Maria
e ao meu marido Sérgio.

Agradecimetos

Agradeço ao meus pais Benjamin e Maria José pelo apoio e incentivo. Ao meu irmão Tiago. Ao meu marido Sérgio. Aos meus avôs, Joaquim e Beatriz, que sempre me apoiaram e compreenderam as minhas decisões e estão juntos com Deus.

À minha orientadora Profa. Dra. Hildete Prisco Pinheiro, pela atenção, dedicação e paciência. Seus conhecimentos e sua paciência foram muito importantes para o desenvolvimento deste trabalho. Os seus conselhos serão levados por mim à vida toda.

À Fapesp pelo apoio financeiro, que foi muito importante para o desenvolvimento desse projeto e para minha formação.

Ao prof. Dr. Aluísio Pinheiro, pelas suas contribuições de grande importância na parte teórica. Ao prof. Dr. Sérgio Furtado dos Reis pelo carinho e paciência. Ao prof. Dr. Edmundo Capelas pela ajuda matemática.

Aos professores do IMECC, por todo conhecimento transmitido durante esses anos. Em especial aos professores Dr. Mauro Sérgio de Freitas Marques, que me incentivou e de quem eu guardo admiração e carinho, Dra. Nancy Lopes Garcia, que eu admiro pela sua capacidade e sua simplicidade.

Aos meus amigos, que me acompanharam durante a graduação e mestrado, dentre eles, Gabriel Coelho Gonçalves de Abreu que me apoiou deste os tempos de graduação em Odontologia. Agradeço em especial aos amigos do curso de Estatística: Daniela Fonsechi, Rafael Moraes, Camila Estevam, Clarice Mendes, Marina Canina, Carolina Barbosa, Samara Khiil, Rosana Lin, Tsai Yun Hui e Benilton. Agradeço muito às minhas amigas Cássia Lagrotta Brigagão e Claúdia Affonso e à Deus por ter colocado pessoas tão maravilhosas na minha vida.

Aos membros da banca examinadora, profa. Dra. Júlia Pavan Soler e prof. Dr. Filidor Vilca Labra, pelas correções e sugestões.

Do not worry about your difficulties in mathematics.
I can assure you that mine are still greater.
Albert Einstein

Resumo

Neste trabalho, o principal interesse é estudar as medidas de distância genética para *loci* de microsatélites baseadas nos desvios absolutos e quadráticos sob o modelo de mutação "*stepwise*".

Os estudos em microsatélites têm sido cada vez mais freqüentes devido a sua importância na aplicação em mapeamento genético. Desta forma, surgeriu-se um modelo para explicar a mutação nas seqüências de repetições nos *loci* de microsatélites, que é conhecido por modelo de mutação "*stepwise*". Nesse modelo supõe-se que a cada geração, cada alelo pode sofrer mutação para outra classe alélica. Na sua forma mais simples, que é o modelo mutacional de um passo o alelo pode sofrer mutação, aumentando ou diminuindo em um estado com probabilidade β. Vamos assumir o modelo de mutação "*stepwise*" de um passo para desenvolver as medidas de distância baseadas nos desvios absolutos e quadráticos.

Propõe-se dois testes de homogeneidade, um baseado na medida de distância dos desvios quadráticos e outro na dos devios absolutos. Suas distribuições assintóticas são estudadas utilizando-se a teoria de Estatística U.

Para verificar os resultados analíticos com respeito a distribuição assintótica, um modelo de simulação foi aplicado baseado no modelo de mutação "*stepwise*" de um passo e na teoria de coalescência.

Os testes de homogeneidade são aplicados a dados reais com o interesse de verificar se existe ou não diferença na variação do número de repetições para os grupos definidos pela etnia e o índice de alcolismo (ALDX1) em um determinado *locus*.

Abstract

In this work, the main interest is to study the measures of genetic distance for microsatellite loci based on the absolute and the quadratic differences under the "it stepwise" mutation model.

The study in microsatellite has become the mainstay due to its importance to develop genetic map. Therefore, one suggests a model to explain the mutation that occurs in the repeated sequence (microsatellite loci), called "*stepwise*" mutation model. The model assumes that in each generation, each allele can mutate to another allelic class. In the simplest case, which we call the "one-step model", ones assumes that the allele can increase or decrease by one unit with probability β. We assume the one-step model to develop the measures of genetic distance based on absolute and quadratic differences.

We suggest two types of homogeneity tests, one based in the measure of quadratic distance and the other based in the absolute distance. Its asymptotic distributions are going to be study using U-statistics theory.

In order to certify the analytical results about the asymptotic distribution, a simulation study based on one-step mutation model and coalescence theory was employed.

An application using real microsatellite data was performed in order to verify if there are differences in the distribution in the repeat sequence among groups defined by ethnicity and alcoholism index (ALDX1) in a determined locus using the homogeneity tests.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Um dos objetivos em estudos genéticos é descrever a quantidade de variação genética entre os indivíduos dentre e entre populações. As medidas de variação genética devem refletir mudanças temporal e espacial, por exemplo, mutações que ocorrem ao longo do tempo e mutações que ocorrem devido ao meio ambiente. Além disso, segundo Chakraborty et al. (1991), essas medidas devem estar fortemente relacionadas com diferentes fatores evolucionários, como mutação e seleção natural.

Duas classes de medidas são consideradas:

- variabilidade genética dentre populações;
- variabilidade genética entre populações.

Antes da descoberta de métodos sorológicos para detectar variabilidade genética, variação morfológica era a maneira mais popular de se avaliar a variabilidade em quase todos os organismos. Embora a variação morfológica seja a maneira mais conveniente de se identificar diferenças entre indivíduos, segundo Chakraborty et al. (1991), o principal problema de medir variação dessa maneira é que não temos certeza sobre o que é realmente

variabilidade genética. Isso se dá devido a fatores ambientais que têm papel importante na variação morfológica e esta variação é difícil de quantificar.

Com o advento de métodos bioquímicos, e com a descoberta de tecnologia de DNA recombinante, as dificuldades apresentadas anteriormente foram superadas, pois a variação genética pode ser detectada a nível molecular.

A origem de estudos sobre distância genética predata da descoberta de marcadores genéticos e a observação de diferenças nas freqüências dos alelos.

Com a descoberta de polimorfismo em microsatélites (VNTR-*variable number of tandem repeat*), medidas de distância genética foram sugeridas por diversos autores para a análise de variação genética com base no número de repetições. Os microsatélites, também denominados de repetições de seqüências simples, compreendem uma classe de DNA repetitivo composto de dois a seis pares de base e encontram-se dispersos no genoma da grande maioria dos organismos estudados (Freimer & Slatkin, 1996). O polimorfismo gera variabilidade genética. Fondon & Garner (2004) sugerem por meio de estudos comparativos entre a morfologia e o genoma que a variação do tamanho de *loci* de microsatélites são a maior fonte de variação morfológica. Os *locus* (ou no plural *loci*) é um local onde se encontra uma seqüência de repetição (mais detalhes ver Seção 1.2). Esses *loci* são definidos pelas seqüências que os flanqueiam, essas seqüências estão dispostas no cromossomo antes e depois de cada *locus* de microsatélites. Dependendo da técnica laboratorial utilizada, o tamanho alélico inclui, além do comprimento das unidades de repetição, o comprimento das seqüências flanqueadoras. Os alelos são definidos pelo número de repetições de nucleotídeos, que é o material genético que compõe o DNA (mais detalhes ver Seção 1.2).

A alta taxa de mutação dos *loci* de microsatélites, comparado com outras regiões do DNA (como as regiões SNP "*single nucleotide polymorphism*"), gera uma grande variação genética. O SNP é uma variação na seqüência de DNA que ocorre quando um único nucleodídeo (A, T, C ou G) no genoma difere entre membros de uma espécie ou entre pares de cromossomos no indivíduo. Por exemplo, duas seqüências de fragmentos de DNA de diferentes indivíduos, AAGCCTA e AAGCTTA, contem uma diferença em um único

nucleotídeo. Neste caso é o C e T.

Os *loci* de microsatélites podem ser mais informativos em estudos da relação entre espécies mais proximamente relacionadas, como também entre subpopulações de uma única espécie (Goldstein et al., 1995), ou seja, estes detectam pequena variação existente entre espécies. Embora os microsatélites sejam muito úteis para determinar a estrutura populacional e as relações entre espécies mais proximamente relacionadas, eles podem ser menos informativos para relação entre espécies mais distantes geneticamente. Isso se deve ao fato da variação nos números de repetições ser muito grande (Goldstein et al., 1995).

A instabilidade genética que ocorre em certas repetições pode implicar na herança de certas doenças congênitas como, por exemplo, a distrofia spino-bulbo muscular e a Síndrome do retardamento mental. O mecanismo para instabilidade das repetições é desconhecido, estudos sugerem que a instabilidade aumenta com o tamanho do alelo (Valdes & Slatkin, 1993).

O processo de mutação nesses *loci* não apresenta falta de memória, ou seja, quando uma mutação ocorre, o novo alelo tem informação sobre o estado do alelo ancestral. Neste caso, a diferença no comprimento entre alelos contém informação filogenética. As estatísticas desenvolvidas por Goldstein et al.(1995) e por Slatkin (1995) incluem essa informação e são equivalentes.

Um grande número de medidas de distância genética pode ser aplicado aos microsatélites, mas existem poucos estudos teóricos e inferenciais sobre a confiabilidade dos resultados obtidos.

Alelos de muitos desses *loci* podem estar envolvidos pelo processo de mutação "*stepwise*", pelo qual os alelos sofrem mutação para mais ou menos uma ou duas unidades de repetição. O modelo de mutação de um passo, pelo qual os alelos sofrem mudança em uma unidade de repetição, é freqüentemente mais utilizado. Embora Valdes (1993) tenha mostrado que esse modelo de mutação é consistente com a distribuição dos alelos nos *loci* de microsatélites, Di Rienzo et al. (1994) fornece evidências de que este pode não ser suficiente para modelar a distribuição das freqüências alélicas nos *loci* de microsatélites.

Essa é uma questão aberta e mais estudos sobre esse modelo devem ser feitos.

O objetivo da dissertação é estudar as medidas de distância genética para *loci* de microsatélites baseadas nos desvios absolutos e quadráticos sob o modelo de mutação "*stepwise*". Para isso, na Seção 1.2 é feita uma introdução de conceitos básicos de biologia molecular e genética populacional, para melhor entendimento do texto, como também introduziremos medidas de distância muito utilizadas em genética populacional, e que são importantes para o estudo de variabilidade dentre e entre populações. Na Seção 1.3 é introduzida a teoria de Processo de coalescência, pois esta será importante para o estudo de medidas de distância genética baseada em *loci* de microsatélites.

No Capítulo 2 é feita uma revisão do modelo de mutação "*stepwise*" aplicados a *loci* de microsatélites. Sob este modelo, a distribuição das freqüências alélicas não apresenta distribuição limite e desta forma é necessário ser desenvolvida uma outra estatística para trabalhar com esse modelo. Uma alternativa é utilizar o momento das freqüências alélicas, que é a probabilidade de que dois genes, que são aleatoriamente amostrados na população, se diferenciem de j unidades de repetição.

No Capítulo 3 faz-se uma extensão da teoria de coalescência introduzida na Seção 1.3 para *loci* de microsatélites, a fim de estudar os estimadores das medidas de distância genética desenvolvidas por Slatkin (1995) e encontrar suas distribuições.

No Capítulo 4 propomos medidas de distância genética dentre e entre populações baseadas na medida desenvolvida por Shriver et al. (1995). Propomos essas medidas como forma de encontrar medidas mais robustas do que as medidas propostas por Slatkin (1995).

No Capítulo 5 fazemos a aplicação a dados reais das medidas propostas. Fazemos também um modelo de simulação para verificar os resultados teóricos obtidos.

1.2 Conceitos básicos de biologia molecular

Os estudos genéticos tornaram-se de grande importância nos dias de hoje. Através deles foi possível detectar diferenças genéticas entre os organismos e mesmo entre as mesmas espécies. Atualmente, diferenças genéticas entre organismos são freqüentemente encontradas diretamente da análise molecular do DNA ou das proteínas. A análise genética é possível em qualquer organismo. Por esse motivo, os conceitos e experimentos em genética populacional têm se tornado importantes para toda área da biologia moderna (Hartl, 2000).

A genética populacional é o estudo da ocorrência natural de diferenças genéticas entre organismos. Diferenças genéticas que são comuns entre organismos da mesma espécie são chamadas em genética de **polimorfismo**, enquanto que diferenças genéticas que acumulam entre espécies são chamadas de divergência genética (Hartl, 2000).

Gene é um termo geral para uma entidade física transmitida dos pais para os seus descendentes durante um processo de reprodução, que influencia a característica hereditária. O conjunto de genes presente num indivíduo constitui o seu **genótipo**. A expressão física do genótipo é chamada de **fenótipo**.

Os genes podem existir em diferentes formas e estados e essas formas alternativas do gene são chamadas de **alelos**. Do ponto de vista bioquímico, no organismo eucarionte, no qual o material genético está envolto por uma membrana (envoltório nuclear), o gene corresponde a uma seqüência específica de nucleotídeos ao longo da molécula de DNA. Diferentes seqüências de nucleotídeos que ocorrem no gene representam os alelos. O DNA é o material genético e existem 4 tipos de nucleotídeos no DNA, que são reconhecidos de acordo com as bases de nitrogenadas, **Adenina (A), Citosina (C), Guanina (G) e Timina (T)**. No caso de microsatélites, os alelos são definidos pelo número de repetições de pares de bases de nitrogênio.

O DNA (ácido desoxiribonucléico) é o responsável pelo armazenamento e transmissão da informação genética (Junqueira & Carneiro, 1991). A molécula de DNA consiste em duas cadeias de nucleotídeos dispostas em hélice. Ao contrário do DNA, a molécula de

RNA (ácido ribonucléico) é um filamento único. No entanto, o RNA passa a informação genética do DNA para as proteínas celulares.

As tecnologias de análise molecular da variabilidade do DNA permitem determinar pontos de referência nos cromossomos, tecnicamente denominados **marcadores moleculares**. Por **marcador molecular** define-se todo e qualquer fenótipo molecular oriundo de um gene expresso ou de um segmento específico de DNA.

No texto fazemos a distinção entre *locus* (*loci* no plural) e **sítio**. Em ambos os casos são locais marcados no cromossomo. No entanto, consideramos que o sítio determina um local com uma base de nitrogênio e o *locus* determina uma seqüência de repetição. O esquema a seguir ilustra esse fato.

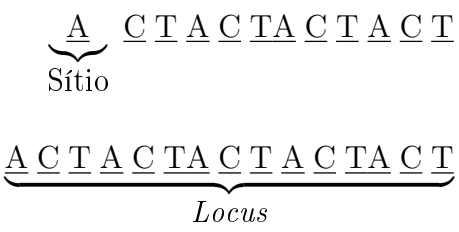

Figura 1.1: Esquema ilustrando a diferença entre *locu* e sítio.

Para uma identificação do genótipo existem diversas técnicas e uma delas é conhecida por **eletroforese**, que é a separação da carga da molécula em campo elétrico. Fragmentos de DNA de tamanhos diferentes se movem no gel em resposta a uma carga elétrica em diferentes taxas e cada fragmento migra numa distância proporcional ao logarítmo do seu tamanho. A posição de cada tamanho do fragmento pode ser localizada no gel visualizando-o sob a luz ultravioleta, onde a concentração de cada tamanho do fragmento produz uma banda fluorescente através do gel (Hartl, 2000).

Um dos atributos universais da população natural é a diversidade fenotípica. Para muitas características, muitos fenótipos diferentes podem ser encontrados entre indivíduos de uma mesma população. Entre os homens, por exemplo, existem diversidades com respeito ao peso, a cor do cabelo, a cor de pele, etc. A genética populacional deve saber

lidar com as diversidades fenotípicas, especialmente com a porção da diversidade que é causada por diferenças no genótipo entre indivíduos. Variação genética, na forma de alelos múltiplos em muitos genes, existe na maioria da população natural.

Um gene **polimórfico** é tipicamente definido como sendo um gene para o qual existe uma probabilidade maior que 99% de observar mais que um alelo numa amostra de 100 genes (50 indivíduos diplóides). De outra forma, um gene **monomórfico** é aquele que não é polimórfico, mais detalhes na Seção 1.2.1.

Alelos na população natural usualmente diferem em freqüência de um alelo para outro. A freqüência alélica de um dado alelo entre um grupo de indivíduos é definida como sendo a proporção de todos os alelos no *locus* que são de um dado tipo. A proporção de um dado alelo na amostra é igual à 2 vezes o número de genótipos homozigóticos para o alelo (porque cada homozigoto carrega 2 cópias do alelo) mais o número de genótipos heterozigóticos por alelo (pois cada heterozigoto carrega uma cópia), dividido por duas vezes o tamanho populacional na amostra (pois cada indivíduo diplóide carrega dois alelos no *locus*), como pode ser visto pela relação abaixo

$$p_A = \frac{2n_{AA} + n_{Aa}}{2n}.$$

Em genética populacional, a palavra *população* não se refere à espécie como um todo e sim a um grupo de indivíduos de uma mesma espécie vivendo numa área geográfica, de forma que cada membro pode potencialmente se acasalar com qualquer outro membro. No caso, a população a ser considerada, será aquela que tem a capacidade de deixar descendentes. Desta forma, será considerado o tamanho populacional efetivo, N_e, ou seja, o tamanho da população de indivíduos que tem capacidade de deixar descendentes. Uma hipótese freqüente nos modelos de estatística genética é que a população seja panmítica, aquela em que não se admite migração e na qual o acasalamento é aleatório.

Recombinação genética é um fenômeno que está intimamente ligado com a meiose celular, que é um processo de divisão celular pelo qual as células diplóides de linhagem germinativa dão origem a gametas haplóides. Uma das causas do aumento da variabilidade genética é a ocorrência de recombinação genética.

Novas variações genéticas são geradas por mutações no material genético. Assim, a mutação é a fonte fundamental de variação genética. O termo mutação é utilizado no sentido de todas as mudanças genéticas, incluindo substituições de bases na molécula de DNA, mudanças na localização de elementos genéticos e reorganização dos cromossomos. A mutação seletivamente neutra é uma situação na qual diferentes tipos de alelos de um determinado gene conferem a mesma chance de sofrerem mutação.

1.2.1 Modelos evolutivos

Um modelo, em geral, é uma simplificação intencional de uma situação complexa (Hartl, 2000). Em genética populacional, fatores como tamanho populacional, distribuição geográfica dos indivíduos, mutação, seleção natural devem ser levados em consideração. Embora tenhamos interesse em entender e relacionar esses fatores, eles se interagem de uma forma tão complexa, que se torna difícil avaliá-los simultaneamente.

Um modelo muito utilizado em genética populacional é o modelo matemático, o qual é um conjunto de hipóteses que especifica relações matemáticas entre medidas e quantidades ditas como parâmetros que caracterizam uma população. Modelos matemáticos são mais simples do que a situação real para a qual eles são construídos. Muitas características do sistema real são propositadamente deixadas de fora do modelo, porque se incluirmos todos os aspectos do sistema no modelo, poderemos deixá-lo muito complexo. Em condições ideais, um modelo matemático deve incluir todas as características essenciais ao sistema e excluir aquelas que não são essenciais.

Dois modelos são muito utilizados em teoria evolutiva, o **modelo de sítios infinitos** e **modelo de Wright-Fisher**. Kimura (1969) formulou o modelo de sítios infinitos, no qual se considera que o número total de sítios em uma seqüência de DNA é muito grande e a taxa de mutação por sítio é muito pequena. Kimura (1968) sugeriu que parte da variação observada em nível molecular é seletivamente neutra. Fisher (1930) e Wright (1931) desenvolveram o modelo de Wright-Fisher, com as seguintes suposições:

1. Acasalamentos aleatórios;

2. Tamanho populacional grande e constante;

3. Não há migração entre populações;

4. Não há mutação genética;

5. Não há efeito de seleção natural e

6. Presença de neutralidade, ou seja, não existe diferenças seletivas entre os alelos.

As suposições do modelo de sítios infinitos não são razoáveis para o o desenvolvimento de medidas de distância em *loci* de microsatélites, pois estamos analisando o que acontece *locus* a *locus*. No entanto, algumas suposições do modelo de Wright-Fisher serão utilizadas, que correspondem àquelas do equilíbrio de **Hardy-Weinberg**:

1. Acasalamentos aleatórios;

2. Tamanho populacional grande;

3. Não há migração entre populações;

4. Não há mutação genética e

5. Não há efeito de seleção natural.

A suposição 4 será deixada de lado, pois supomos que ocorre mutação genética e que esta é seletivamente neutra.

1.2.2 Medidas de distância genética dentre uma população

Quando falamos em genética de populações uma questão importante surge, a existência de estrutura populacional. Uma população pode ser definida como sendo um grupo de indivíduos que vivem juntos no tempo e no espaço. Quando existem vários segmentos em uma mesma população isso indica que existe uma estrutura populacional. Para esse propósito, em genética populacional considera-se que dentro da população os indivíduos

estão aleatoriamente distribuídos e, então, cada membro tem igual chance de acasalar com o outro. Esse conceito é de uma população ideal, freqüentemente chamado de acasalamento aleatório ou população panmítica.

As medidas utilizadas em genética populacional para encontrar variação genética dentre população são:

1. *Proporção de Loci polimórficos.* Segundo Chakraborty et al. (1991), uma definição mais geral para polimorfismo é a probabilidade do alelo mais comum menor ou igual a um valor fracionário escolhido arbitrariamente ($q < 1$), geralmente $q = 0.99$ ou $q = 0.995$ é tomado como critério de polimorfismo. A proporção de *loci* satisfazendo esse critério é chamada de proporção de *loci* polimórficos (P) e fornece uma medida de variação genética na população. Se todos indivíduos são idênticos, em particular homozigotos num mesmo alelo em um determinado *locus*, este é chamado de monomórfico, sendo assim pouca variação é encontrada. Um problema dessa medida é que o valor q não está totalmente independente do tamanho amostral. Por exemplo, numa amostra de n indivíduos diplóides, n pequeno, qualquer alelo com probabilidade P $(0 < P < 1)$ tem a probabilidade de $(1 - P)^{2n}$ de não ser representado na amostra.

2. *Diversidade Gênica.* É uma medida alternativa de variação genética e é definida pela probabilidade de dois genes de um mesmo *locus*, escolhidos aleatoriamente a partir de uma população, serem dissimilares geneticamente. Se $p_1, p_2, \ldots, p_k$ representam a verdadeira proporção de k alelos segregantes, ou seja, k alelos que são diferentes num mesmo *locus*, a diversidade gênica no *locus* é definida por $h = 1 - \sum_{i=1}^{k} p_i^2$. Em uma população com acasalamento aleatório, isso é equivalente à probabilidade de indivíduos heterozigotos. Originalmente, essa medida foi sugerida por Gini (1912) e reinventada por Simpson (1949) para estudos ecológicos. Nei (1972) definiu $J = \sum_{i=1}^{k} p_i^2$ como identidade gênica e em certas situações tem implicação biológica interessante. Por exemplo, em uma população com acasalamento aleatório essa medida é equiva-

lente à homozigosidade na população. A quantidade h é sempre não negativa. O valor mínimo (zero) é atingido quando existe um perfeito monomorfismo ($p_i = 1$ para um único i e zero para os outros $k-1$). O valor máximo está ligado ao número de alelos segregantes na população, $(k-1)/k$. A diversidade gênica máxima é atingida quando todos os alelos segregantes têm probabilidade igual de ocorrência na população. O valor máximo se aproxima de 1 quando $k \to \infty$.

3. *Índice de informação de Shannon.* É uma outra medida de diversidade muito utilizada no contexto de estudos evolucionários e ecológicos. Esta é definida por $h_s = -\sum_{i=1}^{k} p_i lnp_i$.

4. *Número de alelos.* Embora a diversidade gênica e índice de informação de Shannon tenham recebido o máximo de atenção para medir variação genética dentre populações, uma característica perturbadora dessas medidas é que suas magnitudes não são sensíveis ao número de alelos presentes em baixa proporção na população. Por exemplo, mesmo se a diversidade alélica é muito grande num mesmo *locus* na população, ou seja, $k \to \infty$, se uma ou algumas delas são predominantes e as outras raras, a diversidade gênica pelo índice de Shannon será pequena. Conseqüentemente, uma medida alternativa de variação genética pode ser simplesmente o número de diferentes tipos alélicos no *locus*, desconsiderando suas proporções. Um fator ruim para essa medida é que desconsidera a proporção com que ocorrem os vários alelos na população. Além disso, o número esperado de alelos distintos na amostra depende do tamanho amostral. Com a descoberta de polimorfismo em microsatélites (VNTR-*variable number of tandem repeat*), essa medida tornou-se uma estatística útil e informativa na variação genética da população.

5. *Medidas de diversidade unificadas.* É uma tentativa de unificar as medidas de diversidade sob um único método matemático. Rao (1982a), considerou um único *locus* com k alelos diferentes, ocorrendo com proporções $p_1, p_2, \ldots, p_k$ na população. Denotou $h(\mathbf{p}) = h(p_1, p_2, \ldots, p_k)$ a medida de diversidade e assumiu que esta satisfaz

os seguintes postulados:

i) $h(\mathbf{p})$ é simétrica com respeito aos componentes do vetor $\mathbf{p}$ e tem valor máximo no vetor $\mathbf{e} = (1/k, 1/k, \ldots, 1/k)$, ou seja, quando todo os k alelos têm a mesma proporção de ocorrência na população.

ii) $h(\mathbf{p})$ admite derivadas parciais até segunda ordem para as $k-1$ componentes independentes e $h''(\mathbf{p})$ é a matriz de derivadas de segunda ordem, $h''(\mathbf{p}) = [h''_{ij}(\mathbf{p})]$ para $i, j = 1, 2, \ldots, k-1$ com $h''_{ij}(\mathbf{p}) = \dfrac{\partial^2 h(\mathbf{p})}{\partial p_i \partial p_j}$ contínua e não nula em $\mathbf{p} = \mathbf{e}$.

iii) $h(\mathbf{p}+\mathbf{e}) = 1/2[h(\mathbf{p}) + h(\mathbf{e})] = c[h(\mathbf{e}) - h(\mathbf{p})]$ em que c é constante. Então $h(\mathbf{p})$ tem que ser da forma

$$h(\mathbf{p}) = a[1 - \sum_{i=1}^{k} p_i^2] + b,$$

em que $a > 0$ e b são constantes.

Rao (1982b) e Rao & Boudreau (1984) indicaram medidas de diversidades. Entre elas estão:

- α-ordem entropia de *Havrda e Charavat*, definida por

$$h_\alpha(\mathbf{p}) = [1 - \sum p_i^\alpha][2^{\alpha-1} - 1] \qquad \text{para} \quad \alpha > 0 \quad e\alpha \neq 1;$$

- entropia de *Shannon* pareada, definida por

$$h_{\mathsf{p}}(\mathbf{p}) = -\sum p_i ln p_i - \sum (1-p_i) ln(1-p_i);$$

- α graus de entropia de *Renyi*, definida por

$$h_{\mathsf{r}}(\mathbf{p}) = (1-\alpha)^{-1} ln(\sum p_i^\alpha) \qquad \text{para} \quad 0 < \alpha < 1;$$

- função γ-entropia, definida por

$$h_\gamma(\mathbf{p}) = [1 - (\sum p_i^{1/\gamma})^\gamma]/[1 - 2^{\gamma-1}] \quad \text{para} \quad \gamma > 0 \quad \text{e} \quad \gamma \neq 1.$$

Todas essas medidas satisfazem duas condições:

C1: $h(\mathbf{p}) = 0$ se somente se todas componentes do vetor $\mathbf{p}$ são zeros exceto uma, isto é, $p_i = 1$ para um i.

C2: $h[\lambda\mathbf{p} + (1-\lambda)\mathbf{q}] \geq \lambda h(\mathbf{p}) + (1-\lambda)h(\mathbf{q})$, com igualdade se somente se $\mathbf{p} = \mathbf{q}$ (propriedade de concavidade).

1.2.3 Medidas de distância genética entre populações

Supõe-se que dois indivíduos são escolhidos aleatoriamente, um de cada população. O conceito de distância entre duas populações está ligado à magnitude de diversidade entre elas, corrigido pela variação dentro da população.

Seja D_{ij} a distância entre a população i e j. Essa distância deve satisfazer as seguintes propriedades matemáticas:

i) $\mathsf{D}_{ij} \geq 0$ com igualdade se e somente se $i = j$;

ii) desigualdade triangular, isto é, para 3 populações i, j e k devemos ter $\mathsf{D}_{ij} \leq \mathsf{D}_{ik} + \mathsf{D}_{jk}$.

As duas condições implicam que as populações podem ser consideradas como pontos distintos no espaço euclidiano e suas distâncias medidas como a distância geométrica entre os pontos.

Os índices de distância genética podem ser classificados em 2 classes:

i) aqueles que têm o objetivo de classificar a população;

ii) aqueles que têm o objetivo de estudar a evolução.

Nei (1987) classificou no 1^o grupo distâncias como a Métrica de *Manhattan* (Sneath & Sokal, 1973), distância de *Mahalanobis* (Mahalanobis, 1936), distância de *Bhattacharyya* (Bhattacharyya, 1946) entre outras, que são muito utilizadas em análise discriminante. Ele também agrupou na segunda categoria as distâncias baseadas em índices de Wright F_{ST}, Morton's Kinship (Morton, 1975) e sua própria distância (Nei, 1972).

Seja a população α e suas respectivas freqüências alélicas $(p_{\alpha 1}, p_{\alpha 2}, \ldots, p_{\alpha k})$ de k alelos no mesmo *locus*. Se $\mathbf{V}$ é a matriz de variância e covariância entre os alelos de um mesmo

locus, então ela é uma matriz quadrada simétrica de posto $k-1$ tal que o elemento (i,j), v_{ij}, é definido por

$$v_{ij} = \begin{cases} p_{\alpha i}(1-p_{\alpha i}), & \text{para } i{=}j{=}1,\ 2,\ \ldots,\ k-1; \\ -p_{\alpha i}p_{\alpha j}, & \text{para } i \neq j{=}1,\ 2,\ \ldots,\ k-1. \end{cases}$$

Então a inversa da matriz $\mathbf{V}^{-1} = [(v^{ij})]$ é dada por

$$v^{ij} = \begin{cases} 1/p_{\alpha i} + 1/p_{\alpha k}, & \text{para } i{=}j{=}1,\ 2,\ \ldots,\ k-1; \\ 1/p_{\alpha k}, & \text{para } i \neq j{=}1,\ 2,\ \ldots,\ k-1. \end{cases} \tag{1.2.1}$$

Para duas populações α e β, se definirmos as suas distâncias em termos da distância de *Mahalanobis*, calculadas nos $k-1$ primeiros termos do vetor de freqüências alélicas, temos

$$D_M^2 = \sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(p_{\alpha i} - p_{\beta i})\mathbf{S}^{-1}(p_{\alpha j} - p_{\beta j}); \tag{1.2.2}$$

em que $\mathbf{S}$ é a matriz de variância e covariância baseada nas freqüências alélicas das duas populações definida por $\frac{1}{2}(p_{\alpha i} + p_{\beta i})$ para $i = 1,\ 2,\ \ldots,\ k$. Com simplificações algébricas em (1.2.2) e utilizando (1.2.1), temos

$$D_M^2 = \sum_{i=1}^{k}(p_{\alpha i} - p_{\beta i})^2/[2(p_{\alpha i} + p_{\beta i})].$$

Quando duas populações, α e β, não são muito dissimilares nas freqüências alélicas, o índice D_M^2 é equivalente à distância de *Bhattacharyya*, ϖ^2, em que $cos\varpi = \sum_{i=1}^{k}\sqrt{p_{\alpha i}p_{\beta i}}$. Isso é verdade, porque

$$\begin{aligned} cos\varpi &= \sum_{i=1}^{k}\sqrt{p_{\alpha i}p_{\beta i}} = \frac{1}{2}\sum[(p_{\alpha i} + p_{\beta i})^2 - (p_{\alpha i} - p_{\beta i})^2]^{1/2} \\ &= \frac{1}{2}\sum_{i=1}^{k}(p_{\alpha i} + p_{\beta i})\left[1 - \frac{(p_{\alpha i} - p_{\beta i})^2}{(p_{\alpha i} + p_{\beta i})^2}\right]^{1/2} \approx 1 - \frac{1}{4}\sum_{i=1}^{k}\frac{(p_{\alpha i} - p_{\beta i})^2}{(p_{\alpha i} + p_{\beta i})^2} \end{aligned}$$

obtido por expansão em séries da raiz quadrada.

Assim, quando duas populações são geneticamente próximas, as distâncias de *Mahalanobis* (1936) e *Bhattacharyya* (1946) são iguais.

Na análise genética de múltiplas populações (mais de duas), a matriz de variância e covariância, **S**, tem que ser definida sob todas as populações (baseado na hipótese de que a variância genética dentre populações é homogênea). As distâncias entre populações dependem de quais "outras" populações são incluídas no estudo. Na distância genética de Nei, um peso igual é dado para todas as freqüências alélicas entre populações da forma que $D_m = \frac{1}{2}\sum_{i=1}^{k}(p_{\alpha i} - p_{\beta i})^2$, e esta é a distância genética mínima entre as populações α e β. A distância de Nei padrão pode ser definida por

$$D_s = -ln\left(\frac{J_{\alpha\beta}}{\sqrt{J_{\alpha\alpha}J_{\beta\beta}}}\right),$$

em que $J_{\alpha\beta} = \sum_{i=1}^{k} p_{\alpha i}p_{\beta i}$ é a probabilidade de dois genes escolhidos 1 de cada população α e β serem idênticos e $J_{\alpha\alpha} = \sum_{i=1}^{k} p_{\alpha i}^2$, $J_{\beta\beta} = \sum_{i=1}^{k} p_{\beta i}^2$ são medidas de variação dentro das populações α e β, respectivamente.

Segundo Chakraborty et al. (1991), em estudos empíricos mostrou-se que em populações ou organismos mais próximos, as medidas de distância são equivalentes.

1.3 Teoria de Processo de Coalescência

Quando uma coleção de seqüências de DNA homólogos é comparada para um mesmo *locus*, a similaridade entre diferentes seqüências tipicamente contém informações sobre a história evolucionária destas seqüências. Sob ampla variedade de circunstâncias, seqüências de dados fornecem informações sobre quais seqüências estão intimamente relacionadas entre si, e sobre como o mais recente ancestral comum ocorreu no passado. Se as seqüências são obtidas de diferentes espécies, então a informação é freqüentemente arranjada em forma de árvore filogenética, que pode representar a relação evolutiva de uma espécie da qual as seqüências foram amostradas. Se as seqüências são de diferentes indivíduos de uma mesma espécie, a informação é genealógica e, neste caso, árvores de genes podem ser feitas.

Árvores de genes mostram quais seqüências amostradas estão mais intimamente relacionadas entre si e talvez o tempo em que o mais recente ancestral comum de diferentes seqüências ocorreu. Na ausência de recombinação genética, cada seqüência tem um único ancestral em gerações anteriores. Uma árvore de genes hipotética, isto é, a genealogia de 5 seqüências amostradas, é apresentada na Figura 1.2.

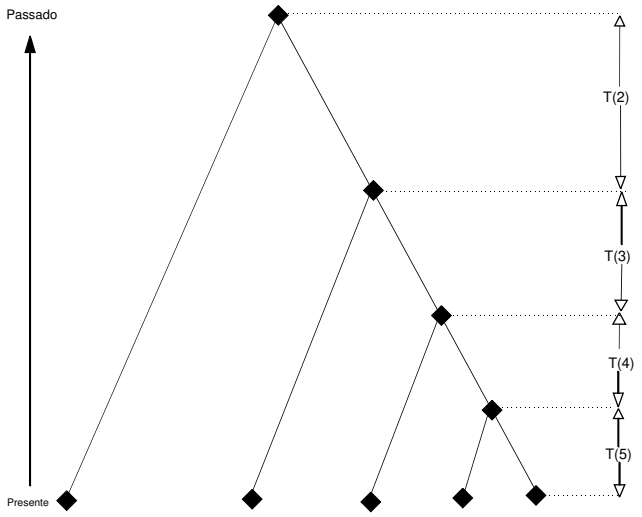

Figura 1.2: Um exemplo de genealogia de uma amostra de 5 alelos em um mesmo *locus*, mostrando o intervalo de tempo entre os eventos de coalescência. Nesta Figura, os intervalos $T(i)$, são mostrados com comprimentos proporcional aos seus valores esperados dados pela equação (1.3.2).

1.3.1 Propriedades de genealogia

As propriedades de genealogia serão apresentadas considerando o modelo de mutação seletivamente neutro definido na Seção 1.2.1. Algumas propriedades serão avaliadas em amostras relativamente pequenas. As propriedades estatísticas de genealogia dependem de fatores como tamanho populacional, estrutura geográfica e presença ou ausência de recombinação genética. Essas propriedades devem depender, também, das condições demográficas pois, a atual genealogia vai depender de quem tem descendência ou não, de

quem migra e para onde e quais descendentes nascem com mutações seletivamente importantes. As mutações seletivamente neutras não devem afetar a genealogia. Isso se deve ao fato de que as mutações seletivamente neutras não afetam o número de descendentes ou a tendência de migração de um indivíduo nascido com essas mutações. É claro que, as propriedades estatísticas sobre o processo genealógico dependem fortemente do processo de mutação. Por exemplo, se a taxa de mutação é muito baixa, todas as seqüências na amostra podem ser idênticas e podemos, assim, não ter informação sobre a genealogia da amostra.

Considere a variável aletória $N(t)$ representando o número de mutações em um mesmo *locus*. A média do número de mutações, μ, é assumida como constante, independente do genótipo, tamanho populacional e do tempo. Assume-se que as mutações ocorrem independentemente dos indivíduos e das gerações. Seja T, a variável aletatória, representando o tempo desde o mais recente ancestral comum de duas amostras de seqüências homólogas, em um mesmo *locus*. Dado t, seja $N(t)$, o número de mutações ocorridas em duas seqüências de descendentes no intervalo $(0, t)$ desde o mais recente ancestral comum. Assume-se que as mutações ocorrem de acordo com um processo de Poisson com uma taxa μ. Assim, $N(t) \mid T = t \sim$ Poisson com média $2\mu t$. A média e variância marginais de $N(t)$ são determinadas pelo momento de T assumindo um processo de mutação neutro com taxa de mutação constante. A distribuição exata de T depende do modelo biológico a ser considerado.

Para enfatizar esse ponto, considere uma população em que no tempo 0 esta é completamente homozigoto no *locus* e somente mutação seletivamente neutra ocorre. Após t gerações de evolução, é examinada a seqüência no *locus* em um único indivíduo selecionado aleatoriamente. Sob o processo de mutação descrito anteriormente, o número de mutações que ocorreu para distinguir o indivíduo amostrado dos indivíduos na população no tempo 0 é somente o número de mutações que ocorreu ao longo do tempo t.

Seja T_{tot} a soma do comprimento dos ramos numa amostra genealógica. Como vimos, $N(t)$ é o número de mutações na genealogia dado $T_{tot} = t$ tem distribuição Poisson com

média $2\mu t$. Os dois primeiros momentos marginais de $N(t)$ podem facilmente serem obtidos, utilizando o seguinte fato

$$\begin{aligned} E(N(t)) &= E(E(N(t) \mid T_{tot})) = 2\mu E(T_{tot}) \qquad \text{e,} \\ Var(N(t)) &= E(Var(N(t) \mid T_{tot})) + Var(E(N(t) \mid T_{tot})) \\ &= 2\mu E(T_{tot}) + 4\mu^2 Var(T_{tot}). \end{aligned} \tag{1.3.1}$$

Hudson (1990) propõe a utilização do modelo de Wright-Fisher, cujas suposições estão na Seção 1.2. Considerando a versão haplóide, n indivíduos haplóides de uma geração descendente são obtidos pela amostragem com reposição n vezes, da geração do pai e da mãe (genitores). Na versão seletivamente neutra, todos os indivíduos têm chances iguais de serem genitores de cada um dos n descendentes haplóides. A descrição detalhada deste modelo está contida em Ewens (1979).

Para um estudo mais detalhado do processo de coalescência, nas Seções 1.3.2 e 1.3.3 serão introduzidos conceitos mais específicos sobre este.

1.3.2 Processo de coalescência sem mutação e sem recombinação genética

Considere uma espécie haplóide de uma população, sem recombinação, sem estrutura geográfica e sem mutação. Suponha que queremos examinar propriedades genealógicas de uma amostra aleatória de n indivíduos haplóides a partir dessa população. Vamos considerar a população a partir da qual a amostra é retirada, de geração 0. A população ancestral t gerações atrás será referida por geração t. A propriedade básica de uma amostra retirada dessa população é baseada na probabilidade $P(n)$, que é a probabilidade de que os n indivíduos haplóides amostrados tenham diferentes ancestrais na geração precedente.

Assim, considere uma primeira amostra de 2 indivíduos haplóides. A probabilidade de que o segundo indivíduo amostrado tenha o mesmo ancestral que o primeiro é $1/N_e$, lembrando que N_e representa o tamanho populacional efetivo. Assim, $P(2) = 1-1/N_e$. Se

3 indivíduos são amostrados, a probabilidade de que os 3 tenham diferentes ancestrais na geração anterior é a probabilidade de que os dois primeiros sejam ancestrais distintos vezes a probabilidade de que o ancestral do terceiro indivíduo seja distinto dos dois primeiros. A probabilidade de que o terceiro tenha ancestrais distintos dos dois primeiros, dado que os dois primeiros têm dois ancestrais diferentes é $(N_e - 2)/N_e = 1 - 2/N_e$. Assim, temos $P(3) = \left(1 - \frac{1}{N_e}\right)\left(1 - \frac{2}{N_e}\right)$.

Em geral, a probabilidade de que n indivíduos haplóides amostrados tenham n ancestrais distintos na geração anterior é

$$\begin{aligned} P(n) &= \prod_{i=1}^{n-1}\left(1 - \frac{i}{N_e}\right) = 1 - \frac{1}{N_e}(\underbrace{1 + 2 + \ldots + n - 1}_{\frac{(n-1)(n-1+1)}{2}}) + \mathrm{O}(N_e^{-2}) \\ &\approx 1 - \frac{\binom{n}{2}}{N_e}. \end{aligned}$$

Qual seria a probabilidade de que eles tenham n diferentes ancestrais uma geração mais cedo? É também $P(n)$. Isso significa que a probabilidade de que n indivíduos amostrados tenham n ancestrais distintos em cada uma das t gerações precedentes e que na geração $t+1$ dois ou mais indivíduos amostrados tenham ancestral comum é

$$(1 - P(n))P(n)^t \approx \frac{\binom{n}{2}}{N_e}\left(1 - \frac{\binom{n}{2}}{N_e}\right)^t \approx \frac{\binom{n}{2}}{N_e} exp\left[-\frac{\binom{n}{2}}{N_e}t\right].$$

A aproximação da distribuição geométrica pela distribuição exponencial é obtida, para $|x| < 1$, como será demonstrado.

Demonstração:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots \quad \Rightarrow \quad \int\left(\frac{1}{1+x}\right)dx = ln(1+x).$$

Assim, temos

$$\left(1-\frac{\binom{n}{2}}{N_e}\right)^t = exp\left[(t)ln\left(1-\frac{\binom{n}{2}}{N_e}\right)\right].$$

Se $x=-\binom{n}{2}/N_e$, em que seu valor absoluto é menor que 1, podemos fazer a aproximação que se segue $ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\ldots\approx x$. Essa aproximação pode ser feita, considerando o fato que o tamanho populacional efetivo, $N_e\to\infty$. Então, o erro que cometemos por essa aproximação é pequeno, na ordem de $\left(-\binom{n}{2}/N_e\right)^2\to 0$, para $N_e\to\infty$. Sendo assim, usando esse fato,

$$exp\left[(t)ln\left(1-\frac{\binom{n}{2}}{N_e}\right)\right] \approx exp\left[\left(-\frac{\binom{n}{2}}{N_e}\right)t\right].$$

Desta forma, a distribuição aproximada é dada por,

$$(1-P(n))P(n)^t \approx \frac{\binom{n}{2}}{N_e}exp\left[-\frac{\binom{n}{2}}{N_e}t\right] = \frac{n(n-1)}{2N_e}exp\left[-\frac{n(n-1)}{2N_e}t\right].$$

□

Em palavras, quando considero uma espécie haplóide de uma população sem mutação, sem recombinação genética, sem migração e para $N_e\to\infty$, o tempo até a primeira ocorrência de um ancestral comum tem distribuição geométrica e pode ser aproximado por uma distribuição exponencial com média $2N_e/n(n-1)$. Para $N_e\to\infty$ e $n\to 0$,

a probabilidade de que mais de dois indivíduos, na amostra, tenham ancestral comum numa única geração é muito pequena e será ignorada. Assim, com alta probabilidade, nossa amostra consiste de t gerações na qual n linhagens distintas existem e então na geração $t+1$ um único par de linhagem coalesce no mais recente ancestral comum de dois indivíduos amostrados. Cada um dos $n(n-1)/2$ possíveis pares de linhagem tem chances iguais para formar os pares coalescentes. Note que nas gerações precedentes, se um par coalesce, vai existir $n-1$ ancestrais ou linhagens a seguir. A probabilidade de que em cada geração todos os indivíduos tenham ancestrais distintos na geração precedente é $P(n-1)$. Então, o tempo para a próxima coalescência tem distribuição aproximadamente exponencial com média $2N_e/(n-1)(n-2)$. Nessa coalescência, cada um dos $(n-1)(n-2)/2$ possíveis pares de linhagens tem chances iguais de coalescerem. Podemos continuar dessa forma até que todas as linhagens tenham coalescido em uma única linhagem, ou seja, tenham um único ancestral comum de uma amostra inteira de n indivíduos. Um exemplo de genealogia da amostra de 5 alelos está mostrado na Figura 1.2. Assim, o processo estocástico que gera a genealogia é referido como processo de coalescência e pode ser resumido brevemente. Desta forma, considere o tempo $T(j)$, durante o qual existem j linhagens distintas com distribuição aproximadamente exponencial e se o tempo é medido em unidades de N_e gerações, a média de $T(j)$ é

$$E[T(j)] = \frac{N_e}{\binom{j}{2}}. \tag{1.3.2}$$

As duas linhagens que coalescem no nó na genealogia, dita geração $t+1$, são duas linhagens escolhidas aleatoriamente a partir da geração t. Note que os intervalos entre as coalescências, $T(j)$'s, são estatisticamente independentes entre si. Para genealogias de amostras pequenas, é importante notar que as partes mais altas da genealogia são idênticas em propriedades estatísticas. Por exemplo, a parte da genealogia de tamanho amostral n é distribuída exatamente como a parte da genealogia de tamanho $n-1$.

Considerando uma população de indivíduos diplóides, para um tamanho populacional

efetivo grande, $N_e \to \infty$, sob a hipótese do modelo de Wright-Fisher com acasalamento aleatório, sem recombinação e sem mutação, os resultados são os mesmos, exceto que N_e é substituído por $2N_e$ e n é substituído por $2n$, pois são n indivíduos diplóides. A genealogia, neste caso, deve ser pensada como a genealogia de *loci* específicos no qual nenhuma recombinação ocorre.

1.3.3 Processo de coalescência com mutação

Dada as propriedades de genealogia descritas anteriormente, assume-se uma taxa constante de mutação seletivamente neutra, na qual cada gameta descendente difere do seu ancestral por uma média μ de mutações. Além disso, assume-se o modelo de sítios infinitos. Sob esse modelo, os *loci* são compostos de muitos sítios, por isso, não mais do que uma mutação ocorre em qualquer sítio na genealogia da nossa amostra. Para nosso propósito o modelo de sítios infinitos e alelos infinitos, no qual assume-se que a cada mutação produz-se um novo alelo, são essencialmente os mesmos.

A primeira propriedade a ser considerada diz respeito à distribuição do número de mutações que ocorre nos ramos da genealogia na amostra. Sob o modelo de sítios infinitos, o número de mutações é idêntico ao número de sítios de nucleotídeos que podem ser polimórficos na amostra. O número de sítios polimórficos na amostra, denotado por S, é freqüentemente chamado de números de sítios segregantes na amostra, ou seja, é a soma de todos os sítios onde encontramos diferença entre os nucleotídeos nas seqüências. Defina S_j o número de sítios segregantes na geração $j = 2, \ldots, n$. Então, $S = S_2 + S_3 + \ldots + S_n$. Temos que o número de sítios segregantes depende exclusivamente da taxa de mutação do gene. Logo, dado $T_{tot} = t$, assume-se que a distribuição de S é Poisson com parâmetro μt. Na Tabela 1.1, apresentam-se 5 seqüências com 16 sítios polimórficos e um não polimórfico correspondendo à coluna 17.

Assim, $E[S \mid T_{tot}] = \mu T_{tot} \Longrightarrow E[S] = \mu E[T_{tot}]$. Segue, à partir da definição de $T(j)$, que a soma dos comprimentos dos ramos da genealogia é $T_{tot} = \sum_{i=2}^{n} iT(i)$. Então, pela

Tabela 1.1: Sítios segregantes.

Seqüências	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
a	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A	C
b	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T	C
c	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A	C
d	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T	C
e	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A	C

equação (1.3.2) e usando o fato de ser um indivíduo diplóide ($2N_e$), segue-se que

$$E(S) = E[E(S \mid T_{tot})] = \mu \sum_{i=2}^{n} iE(T(i)) = \frac{\theta}{2} \sum_{i=2}^{n} i \frac{1}{\begin{pmatrix} i \\ 2 \end{pmatrix}} = \theta \sum_{i=1}^{n-1} \frac{1}{i}, \tag{1.3.3}$$

em que $\theta = 4N_e\mu$, que é interpretado como sendo o número esperado de mutações por sítio por geração. A variância do tempo total é também obtida facilmente e utilizando as equações (1.3.1) e (1.3.3), obtém-se

$$Var(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \quad .$$

De fato, qualquer momento de S pode ser expresso em termos do momento de $T(i)$. Com isso, podemos obter os momentos de S. No entanto, considere a probabilidade de $S = 0$, para uma amostra de tamanho 2. Para que dois alelos amostrados sejam idênticos, sob o modelo de sítios infinitos (ou modelo de alelos infinitos), nenhuma mutação deve ocorrer nas linhagens que descendem dos seus mais recentes ancestrais comuns. Dado t, o número de gerações até o mais recente ancestral comum, a probabilidade de que nenhuma mutação tenha ocorrido nos descendentes na amostra dos alelos é $e^{-2\mu t}$. Isso se deve à hipótese da distribuição do número de mutações ser Poisson. Então, se calcularmos a esperança de $e^{-2\mu T}$, considerando a distribuição de T, que é exponencial com média $2N_e$ no modelo diplóide (para uma amostra de tamanho $n = 2$), temos que

$$E(e^{-2\mu T}) \quad = \quad \int_0^\infty \frac{e^{-t/2N_e}}{2N_e} e^{-2\mu t} dt = \frac{1}{2N_e} \int_0^\infty e^{\left\{-\left(\frac{1}{2N_e}+2\mu\right)t\right\}} dt$$

$$= \frac{1}{2N_e}\left(\frac{1}{2\mu + \frac{1}{2N_e}}\right) = \frac{1}{1+\theta} \quad .$$

Utilizando esse fato, podemos derivar a distribuição do número de mutações que ocorreram desde o mais recente ancestral comum em uma amostra de tamanho 2. A probabilidade $P_2(j)$ de que j mutações ocorrem nas linhagens desde o mais recente ancestral comum, é a probabilidade de que os j primeiros eventos sejam mutações e $(j+1)$-ésimo evento seja um ancestral comum. Assim, temos

$$P_2(j) = \left(\frac{\theta}{\theta+1}\right)^j \frac{1}{1+\theta} \quad .$$

Utilizando um argumento similar, podemos obter a probabilidade $Q_n(j)$, em que j mutações ocorrem no tempo no qual existem n linhagens ancestrais. Para se ter j mutações durante esse tempo, os primeiros j eventos, durante o tempo que existe n linhagens, devem ser mutações, e o $(j+1)$-ésimo evento deve ser um ancestral comum. Então, essa probabilidade é

$$Q_n(j) = \left(\frac{n\mu}{n\mu + \frac{\binom{n}{2}}{2N_e}}\right)^j \frac{\frac{\binom{n}{2}}{2N_e}}{n\mu + \frac{\binom{n}{2}}{2N_e}} = \left(\frac{\theta}{\theta+n-1}\right)^j \frac{n-1}{\theta+n-1},$$

em que $\theta = 4N_e\mu$. Segue que $P_n(j)$, a probabilidade de que j mutações ocorreram numa amostra de tamanho n, pode ser escrita por

$$P_n(j) = \sum_{i=0}^{j} P_{n-1}(j-i)Q_n(i).$$

Capítulo 2

Modelo de mutação "*stepwise*"

2.1 Introdução

O modelo de mutação "*stepwise*" foi introduzido por Ohta & Kimura (1973) e Wehrnahn (1975). Tanto Otha & Kimura (1973) como Wehrnahn (1975) sugerem que existem regularidades na distribuição das freqüências alélicas dentre *locus*. Bulmer (1971) notou que em muitos *loci* existia um alelo comum com mobilidade intermediária e outros menos comuns distribuídos com mobilidade simétrica sobre o alelo comum. Bulmer (1971) sugeriu que essas observações eram consistentes com o modelo de mutação em que os alelos poderiam aumentar ou diminuir em uma ou duas unidades de repetição. Sob essa hipótese, os alelos com a mesma mobilidade não necessariamente seriam idênticos por descendência, mas idênticos por estado. Define-se idêntico por descendência, alelos que são cópias idênticas de um alelo derivado de uma replicação de um ancestral comum. Idêntico por estado é quando os alelos são idênticos em termos da composição e função de DNA, mas sem considerar a ancestralidade (Andrade & Pinheiro, 2002).

O modelo de mutação "*stepwise*" supõe que há uma população de indivíduos diplóides de tamanho efetivo N_e, que equivale à $2N_e$ gametas, e considera um único *locus* em cada alelo distinguido por um inteiro i, $i = 0, \pm1, \pm2, \ldots$, que indica o número de unidades repetidas no *locus*, ou a carga (estado). Supõe-se que todos os alelos nesse *locus* são

seletivamente equivalentes e há o acasalamento aleatório na população e que, a cada geração, cada alelo pode sofrer mutação para outra classe alélica. No caso mais simples supõe-se que i pode aumentar ou diminuir de uma unidade de repetição com probabilidade $\beta/2$ por geração e por *locus*, e continuar no estado i com probabilidade $\alpha = (1-\beta)$. Esse modelo é conhecido por "modelo de um passo" ou "*one-step model*". No modelo de dois passos o número de unidades repetidas do estado i pode aumentar ou diminuir de uma ou duas unidades com probabilidade $\beta/2$ e $\gamma/2$, respectivamente, e continuar no estado i com probabilidade $\alpha = (1-\beta-\gamma)$.

2.2 Distribuição das freqüências alélicas

Seja $p_i(t)$ a proporção do alelo A_i na população no tempo t (alelo A com um número inteiro de unidades de repetição i ou de carga i), ou seja, é a freqüência alélica de i. Uma forma de indexar as freqüências alélicas observadas é $\ldots, \hat{p}_{-2}(t), \hat{p}_{-1}(t), \hat{p}_0(t), \hat{p}_{+1}(t), \hat{p}_{+2}(t), \ldots$, de acordo com a sua mobilidade. Podemos fazer o produto cruzado das freqüências alélicas observadas, definido por

$$\mathcal{C}_j = \sum \hat{p}_i(t)\hat{p}_{i+j}(t), \quad i = \ldots, -2, -1, 0, 1, 2, \ldots \quad .$$

Como $j = \ldots, -2, -1, 0, 1, 2, \ldots$ e $\mathcal{C}_j = \mathcal{C}_{-j}$, considera-se as quantidades $\{2\mathcal{C}_j; j = 1, 2, \ldots\}$ correspondentes às proporções esperadas de heterozigotos cujos alelos diferem de j unidades de carga sob a hipótese de acasalamento aleatório. Para uma população de tamanho N_e

$$\sum_{i=-\infty}^{\infty} p_i(t) = 1. \tag{2.2.1}$$

Os alelos são considerados seletivamente neutros e afetados por mutação com uma taxa μ por *locus* por geração. Assim, se temos o modelo de mutação de um passo, $\mu = \beta$ e se de dois passos $\mu = \beta + \gamma$. Dado t, Moran (1975) mostrou que a seqüência $\{\hat{p}_i(t)\}$, $i = 0, \pm 1, \pm 2, \ldots$ não têm distribuição limite sob o modelo de mutação de um passo.

Seja $S_i(t)$ a variável aleatória representando a freqüência de gametas na população com alelo A_i, alelo de tamanho (estado) i no tempo t e $p_i(t) = S_i(t)/2N_e$, em que $2N_e$ é o tamanho populacional efetivo. Moran (1975) considerou o modelo de mutação de um passo em que cada geração de gametas do tipo i é suposta ser capaz de mutar para os estados $(i-1)$ ou $(i+1)$ com probabilidade $\beta/2$ cada e o restante continuar no estado i com probabilidade $(1-\beta)$. Seja o vetor de freqüências $\mathbf{S}(t+1)$ no tempo $(t+1)$, com distribuição multinomial com total $2N_e$ e a probabilidade do estado i igual à $\pi_i(t+1)$. Considere $W(t+1)$ a variável aleatória representando o tamanho do alelo no tempo $t+1$. A probabilidade do alelo ter tamanho i no tempo t, $\pi_i(t+1)$, é definida por $\pi_i(t+1) = P(W(t+1) = i)$, ou seja, é igual a probabilidade do tamanho alélico ser igual a i no tempo $t+1$. Pelo Teorema da probabilidade total essa probabilidade é igual à

$$\begin{aligned} \pi_i(t+1) &= \sum_{j=i-1}^{i+1} P(W(t+1) = i, W(t) = j) \\ &= \sum_{j=i-1}^{i+1} P(W(t+1) = i \mid W(t) = j) P(W(t) = j). \end{aligned}$$

A probabilidade condicional $P(W(t+1) = i \mid W(t) = j)$ é a probabilidade de transição de um estado para o outro, e a probabilidade $P(W(t) = i)$ pode ser estimada por $\hat{p}_i(t) = n_i(t)/2N_e$, que é a proporção do estado i na população no tempo t, que pode ser conhecida ou observada, assim $\hat{\pi}_i(t+1)$ é obtido por

$$\begin{aligned} \hat{\pi}_i(t+1) &= \frac{1}{2N_e}\left\{(1-\beta)n_i(t) + \frac{\beta}{2}n_{i-1}(t) + \frac{\beta}{2}n_{i+1}(t)\right\} \\ &= (1-\beta)\hat{p}_i(t) + \frac{\beta}{2}\hat{p}_{i-1}(t) + \frac{\beta}{2}\hat{p}_{i+1}(t), \end{aligned} \tag{2.2.2}$$

que equivale na prática à

$\pi_i(t+1)$ = (a probabilidade de continuar no estado i multiplicada pela proporção do estado i na população no tempo t) + (a probabilidade de aumentar de uma unidade à partir do estado $i-1$ multiplicada pela proporção do estado $i-1$ na população no tempo t) + (a probabilidade de diminuir de uma unidade o estado i multiplicada pela proporção do estado $i+1$ na população no tempo t).

A cada geração, o conjunto de quantidades $n_i(t)$ forma a distribuição empírica das freqüências alélicas dos estados. Não importa o tamanho de N_e, se $t \to \infty$, a distribuição empírica de $S_i(t)$ não converge para uma distribuição limite e, consequentemente, as quantidades $\hat{p}_i(t)$ também não, o que vai ser ilustrado a seguir.

Suponha $n_i(0)$ fixo no tempo 0 e defina

$$\begin{aligned} M_1(t) &= (2N_e)^{-1}\sum_i i n_i(t) = \sum_i i\hat{p}_i(t) \\ M_1^2(t) &= (2N_e)^{-2}\sum_{i,j} ij n_i(t)n_j(t) = \sum_{i,j} ij\hat{p}_i(t)\hat{p}_j(t) \\ M_2(t) &= (2N_e)^{-1}\sum_i i^2 n_i(t) = \sum_i i^2\hat{p}_i(t), \end{aligned} \tag{2.2.3}$$

em que $M_1(t)$ e $M_2(t)$ são os primeiro e segundo momentos amostrais da distribuição das freqüências $S_i(t)$ na geração t, respectivamente. Considere $E(\cdot)$ a esperança condicional no t-ésimo tempo dada o tempo $t-1$ e E_1 denota a esperança condicional do estado da população dado t=0 e E_0 a esperança da situação estacionária, isto é, o limite da esperança de E_1, se essa existir.

Para encontrar o valor esperado de $M_1(t)$ devemos considerar a variável aleatória $S_i(t) \sim Binomial(2N_e, \pi_i(t))$ e probabilidade $\pi_i(t)$ definida anteriormente. Então

$$\begin{aligned} E[M_1(t)] &= (2N_e)^{-1}\sum_i iE[S_i(t)] = \sum_i i\pi_i(t) \\ &= (2N_e)^{-1}\sum_i i\left\{(1-\beta)n_i(t-1) + \frac{\beta}{2}n_{i-1}(t-1) + \frac{\beta}{2}n_i + 1(t-1)\right\} \\ &= M_1(t-1). \end{aligned}$$

Então $E_1[M_1(t)] = M_1(0)$. Temos também que

$$\begin{aligned} E\left[M_1^2(t)\right] &= (2N_e)^{-2}\sum_{i,j} ijE[S_i(t)S_j(t)] \\ &= (1-(2N_e)^{-1})M_1^2(t-1) + (2N_e)^{-1}M_2(t-1) + \beta(2N_e)^{-1}, \end{aligned}$$

cuja demonstração se encontra em Apêndice A número 1.

Se $M_2(t-1) \geq M_1^2(t-1)$, então $E[M_1^2(t)] \geq M_1^2(t-1) + \beta(2N_e)^{-1}$ e quando $t \to \infty$, $M_1^2(t) \to \infty$ e assim, $E_1[M_1^2(t)] \to \infty$. E, desta forma, o segundo momento da distribuição empírica das freqüências alélicas condicionada ao tempo $t = 0$ não existe.

Então as quantidades $S_i(t)$ não convergem ou não se aproximam de uma distribuição limite e, conseqüentemente, $\hat{p}_i(t)$ também não.

2.3 Momento das freqüências alélicas

Como $\hat{p}_i(t)$ não tem distribuição limite, podemos trabalhar com o momento das freqüências alélicas. Desta forma, considerando a seqüência $\{\hat{p}_i(t)\}$ na população, o momento das freqüências alélicas $\{C_j\}$ é definido por

$$C_j = C_{-j} = E(\mathcal{C}_j) = E\left(\sum_{i=-\infty}^{\infty} \hat{p}_i(t)\hat{p}_{i+j}(t)\right) \quad \text{e} \tag{2.3.1}$$

$$C_0 + 2\sum_{j=1}^{\infty} C_j = 1, \qquad 0 < C_j < 1, \tag{2.3.2}$$

em que a soma é sobre todos os estados alélicos na população em um *locus*. O valor C_j é a probabilidade de que dois genes, aleatoriamente amostrados da população, se diferenciem de exatamente j repetições. Note que C_0 é o valor esperado da soma dos quadrados das freqüências alélicas. Assim, ele nos dá a média de homozigotos sob acasalamento aleatório. A correlação entre freqüências dos alelos que estão j passos separados pode ser dada por C_j/C_0.

Ohta & Kimura (1973) consideraram o modelo de mutação de um passo para obter um conjunto de equações, dada a taxa de mutação por gene e por geração μ, resolvendo equações de geratriz de momentos. Essas equações tomam a seguinte forma

$$\frac{d}{dt}E[f(\cdot)] = E\{L[f(\cdot)]\}, \tag{2.3.3}$$

em que L é o operador diferencial da equação "*backward*" de Kolmogorov definido por

$$L(f(p_i(t))) = \frac{V_{\delta p_i(t)}}{2} f''(p_i(t)) + M_{\delta p_i(t)} f'(p_i(t)), \tag{2.3.4}$$

$f(\cdot)$ é uma função contínua arbitrária da seqüência $p_i(t)$ e $\delta p_i(t)$ é o incremento na freqüência alélica p_i durante um curto período de tempo entre t e $t+\delta t$. Então,

$$p_i(t+\delta t) = p_i(t) + \delta p_i(t) \quad \text{e} \quad \delta p_i(t) = -\mu p_i(t) + \frac{\mu}{2}[p_{i-1}(t) + p_{i+1}(t)] + \xi_i,$$

em que ξ_i representa a mudança devido a amostra aleatória dos gametas, ou seja, o erro aleatório, e $E(\xi_i) = 0$, $Var(\xi_i) = p_i(t)(1-p_i(t))/2N_e$ e $\text{cov}(\xi_i, \xi_j) = -p_i(t)p_j(t)/2N_e$.

Considere o valor esperado

$$E\{f[\hat{p}_i(t+\delta t)]\} = E_\phi E_\delta\{f[\hat{p}_i(t) + \delta\hat{p}_i(t)]\}, \tag{2.3.5}$$

em que E_δ é o operador esperança com respeito à mudança $\delta\hat{p}_i(t)$ e E_ϕ é o operador esperança com respeito à distribuição da freqüência alélica $\hat{p}_i(t)$ no tempo t.

Expandindo em série de Taylor o lado direito da equação (2.3.5) em termos de $\delta\hat{p}_i(t)$ e desconsiderando os termos de terceira ordem e ordem superior, têm-se

$$\begin{aligned} E\{f[\hat{p}_i(t+\delta t)]\} &= E_\phi E_\delta \left\{ f[\hat{p}_i(t)] + [\delta\hat{p}_i(t)]f'[\hat{p}_i(t)] + \frac{[\delta\hat{p}_i(t)]^2}{2!} f''(\hat{p}_i(t)) \right\} \\ &= E_\phi \left\{ f[\hat{p}_i(t)] + E_\delta[\delta\hat{p}_i(t)]f'[\hat{p}_i(t)] + \frac{E_\delta[\delta\hat{p}_i(t)]^2}{2!} f''(\hat{p}_i(t)) \right\} \end{aligned}$$

e

$$\frac{E\{f[\hat{p}_i(t+\delta t)]\} - E_\phi\{f(\hat{p}_i(t))\}}{\delta t} = E_\phi \left\{ \frac{E_\delta[\delta\hat{p}_i(t)]}{\delta t} f'(\hat{p}_i(t)) + \frac{1}{2}\frac{E_\delta[\delta\hat{p}_i(t)]^2}{\delta t} f''(\hat{p}_i(t)) \right\}.$$

Considerando $E_\phi = E$, no limite de $\delta t \to 0$, obtém-se

$$\frac{d}{dt} E\{f(\hat{p}_i(t))\} = E\left\{ \frac{V_{\delta\hat{p}_i(t)}}{2} f''(\hat{p}_i(t)) + M_{\delta\hat{p}_i(t)} f'(\hat{p}_i(t)) \right\},$$

em que $M_{\delta\hat{p}_i(t)}$ e $V_{\delta\hat{p}_i(t)}$ são, respectivamente, a média e a variância do incremento da freqüência alélica e são aproximações para

$$\lim_{\delta t \to 0} \left\{ \frac{E_\delta[\delta\hat{p}_i(t)]}{\delta t} \right\} \quad \text{e} \quad \lim_{\delta t \to 0} \left\{ \frac{E_\delta[\delta\hat{p}_i(t)]^2}{\delta t} \right\},$$

respectivamente.

A extensão da equação (2.3.4) para o caso multivariado, considera $f = f(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$ para n estados alélicos. Da mesma forma, utiliza-se a expansão em Taylor de ordem 2 que, para duas variáveis, é dada por

$$\begin{aligned} f(x+h, y+k) &= f(x,y) + h\frac{\partial f(x,y)}{\partial x} + h\frac{\partial f(x,y)}{\partial y} + \frac{1}{2}\left[h^2\frac{\partial^2 f(x,y)}{\partial x^2} + 2hk\frac{\partial^2 f(x,y)}{\partial x \partial y} + \right. \\ &+ \left. k^2\frac{\partial^2 f(x,y)}{\partial y^2}\right] \end{aligned}$$

e para o caso multivariado, é a extensão dessa equação.

Aplicando o valor esperado na expansão em Taylor para o caso multivariado, da mesma maneira que anteriormente e aplicando o limite ($\delta t \to 0$), obtém-se o seguinte operador diferencial L

$$L(f(\cdot)) = \frac{1}{2}\sum_{i=1}^{n} V_{\delta\hat{p}_i(t)}\frac{\partial^2 f(\cdot)}{\partial \hat{p}_i^2} + \sum_{i>j} W^*_{\delta\hat{p}_i(t)\delta\hat{p}_j(t)}\frac{\partial^2 f(\cdot)}{\partial \hat{p}_i \partial \hat{p}_j} + \sum_{i=1}^{n} M_{\delta\hat{p}_i(t)}\frac{\partial f(\cdot)}{\partial \hat{p}_i}, \qquad (2.3.6)$$

em que $W^*_{\delta\hat{p}_i(t)\delta\hat{p}_j(t)} = \mathrm{cov}[\delta\hat{p}_i(t), \delta\hat{p}_j(t)]$.

Supondo que os alelos são seletivamente neutros, a média (M), a variância (V) e a covariância (W^*) da freqüência $\delta\hat{p}_i(t)$ do gene por geração para população de tamanho N_e ($2N_e$ gametas) são, respectivamente

$$\begin{aligned} E[\delta\hat{p}_i(t)] &= M_{\delta\hat{p}_i(t)} = \frac{\mu}{2}[p_{i-1}(t) + p_{i+1}(t)] - \mu p_i(t), \\ Var[\delta\hat{p}_i(t)] &= V_{\delta\hat{p}_i(t)} = \frac{p_i(t)(1-p_i(t))}{2N_e}, \\ \mathrm{cov}(\delta\hat{p}_i(t), \delta\hat{p}_j(t)) &= W^*_{\delta\hat{p}_i(t)\delta\hat{p}_j(t)} = -\frac{p_i(t)p_j(t)}{2N_e} \quad (i \neq j) \end{aligned}$$

e, seja $f = \hat{p}_i^2(t)$, nas equações (2.3.3) e (2.3.4) então

$$\begin{aligned} \frac{d}{dt}E[\hat{p}_i^2(t)] &= E\left\{\mu\left[\hat{p}_{i-1}(t) + \hat{p}_{i+1}(t)\right]\hat{p}_i(t) - 2\mu\hat{p}_i^2(t) + \frac{\hat{p}_i(t)(1-\hat{p}_i(t))}{2N_e}\right\} \quad \text{ou} \\ \frac{d}{dt^*}E[\hat{p}_i^2(t)] &= E\left\{2N_e\mu\left[\hat{p}_{i-1}(t) + \hat{p}_{i+1}(t)\right]\hat{p}_i(t) - \left[4N_e\mu + 1\right]\hat{p}_i^2(t) + \hat{p}_i(t)\right\}, \end{aligned}$$

em que $t^* = t/2N_e$ é a medida de tempo na unidade de $2N_e$ gerações. Somando essa equação para todo i, e levando em consideração as relações (2.2.1) e (2.3.1) e que $E(\cdot)$ é

um operador linear, obtém-se

$$\begin{aligned}\frac{d}{dt^*}C_0 &= -(1+4N_e\mu)C_0 + 4N_e\mu C_1 + 1 \\ &= -(1+\theta)C_0 + \theta C_1 + 1,\end{aligned} \tag{2.3.7}$$

em que, pela Seção 1.3.3, sabemos que $\theta = 4N_e\mu$ é o número esperado de mutações por sítio por geração. No contexto de microsatélites, esse parâmetro será interpretado como sendo o número esperado de mutações por *locus* por geração, para o modelo de mutação de um passo.

Similarmente, considerando $f = \hat{p}_i(t)\hat{p}_{i+j}(t)$, substituindo na equação (2.3.3) e (2.3.6) obtém-se

$$\begin{aligned}\frac{d}{dt}E[\hat{p}_i(t)\hat{p}_{i+j}(t)] &= E\Big\{\frac{\mu}{2}[\hat{p}_{i-1}(t)+\hat{p}_{i+1}(t)-2\hat{p}_i(t)]\hat{p}_{i+j}(t)+ \\ &+ \frac{\mu}{2}[\hat{p}_{i+j-1}(t)+\hat{p}_{i+j+1}(t)-2\hat{p}_{i+j}(t)]\hat{p}_i(t) - \frac{\hat{p}_i(t)\hat{p}_j(t)}{2N_e}\Big\}\end{aligned}$$

ou

$$\begin{aligned}\frac{d}{dt^*}E[\hat{p}_i(t)\hat{p}_{i+j}(t)] &= E\{N_e\mu\left[\hat{p}_{i-1}(t)+\hat{p}_{i+1}(t)-2\hat{p}_i(t)\right]\hat{p}_{i+j}(t)+ \\ &+ N_e\mu[\hat{p}_{i+j-1}(t)+\hat{p}_{i+j+1}(t)-2\hat{p}_{i+j}(t)]\hat{p}_i(t) - \hat{p}_i(t)\hat{p}_j(t)\}.\end{aligned}$$

Assim, somando para todo i, e com as mesmas relações utilizadas anteriormente, tem-se que

$$\begin{aligned}\frac{d}{dt^*}C_j &= 2N_e\mu C_{j-1} - (1+4N_e\mu)C_j + 2N_e\mu C_{j+1}, \quad (j \geq 1) \\ \frac{d}{dt^*}C_j &= \frac{\theta}{2}C_{j-1} - (1+\theta)C_j + \frac{\theta}{2}C_{j+1}, \quad (j \geq 1).\end{aligned} \tag{2.3.8}$$

A população atinge o estado de equilíbrio quando as freqüências alélicas permanecem contantes com o passar do tempo. No estado de equilíbrio, tem-se que $dC_j/dt = 0$ para todo j. Igualando as equações (2.3.7) e (2.3.8) à zero, obtém-se as seguintes soluções para C_j

$$\begin{aligned}C_j &= a_1\lambda_1^j, \quad \lambda_1 = \frac{1+\theta-\sqrt{1+2\theta}}{\theta} \quad \text{e} \\ a_1 &= \frac{1}{\sqrt{1+2\theta}}.\end{aligned} \tag{2.3.9}$$

Brown et al. (1975) mostraram, para o modelo de mutação de dois passos, que o valor esperado de C_j tem valores no equilíbrio na forma

$$C_j = a_1\lambda_1^j + a_2\lambda_2^j \qquad j = 0, 1, 2, ... \tag{2.3.10}$$

em que a_m e λ_m, para $m = 1, 2$, são constantes que dependem de $4N_e\beta$ e $4N_e\gamma$ com a condição de que $|\lambda_m| < 1$. Considerando o mesmo raciocínio de Ohta & Kimura (1973) para o modelo de dois passos, o valor esperado de $\delta\hat{p}_i(t)$ é dado por

$$\begin{aligned} E[\delta\hat{p}_i(t)] &= M_{\delta\hat{p}_i(t)} \\ &= \frac{\beta}{2}\left[p_{i-1}(t) + p_{i+1}(t)\right] + \frac{\gamma}{2}\left[p_{i-2}(t) + p_{i+2}(t)\right] - (\beta+\gamma)p_i(t). \end{aligned} \tag{2.3.11}$$

Assim, o operador diferencial L para $f = \hat{p}_i^2(t)$ é

$$\begin{aligned} \frac{d}{dt}E[\hat{p}_i^2(t)] &= E\left\{2E[\delta\hat{p}_i(t)]\hat{p}_i(t) + \frac{\hat{p}_i(t)(1-\hat{p}_i(t))}{2N_e}\right\} \quad \text{ou} \\ \frac{d}{dt^*}E\left[\hat{p}_i^2(t)\right] &= E\{2N_e(\beta+\gamma)[\hat{p}_{i-2}(t) + \hat{p}_{i-1}(t) + \hat{p}_{i+1}(t) + \hat{p}_{i+2}(t)]\hat{p}_i(t) \\ &- [4N_e(\beta+\gamma)+1]\hat{p}_i^2(t) + \hat{p}_i(t)\} \end{aligned}$$

em que $t^* = t/2N_e$. Somando essa equação para todo i, com as mesmas condições do modelo de um passo, obtém-se

$$\begin{aligned} \frac{d}{dt^*}C_0 &= -(1 + 4N_e\gamma + 4N_e\beta)C_0 + 4N_e\beta C_1 + 4N_e\gamma C_2 + 1 \\ &= -(1 + 2u + 2v)C_0 + 2uC_1 + 2vC_2 + 1, \end{aligned} \tag{2.3.12}$$

em que $u = 2N_e\beta$ e $2u$ é o número esperado de mutações de uma unidade de repetição por *locus* por geração e $v = 2N_e\gamma$ e $2v$ é o número esperado de mutações de duas unidades de repetição por *locus* por geração. Com isso, $\theta = 4N_e\mu = 2u + 2v$, para o modelo de mutação de dois passos. Como no equilíbrio ($dC_0/dt = 0$) então por (2.3.12), a relação fica $C_0(1 + 2u + 2v) = 1 + 2uC_1 + 2vC_2$.

Da mesma forma, para obter C_j para o modelo de dois passos considera-se o operador diferencial L definido na equação (2.3.6) para $f = \hat{p}_i(t)\hat{p}_{i+j}(t)$, o valor esperado definido

em (2.3.11) e, somando para todo i, a seguinte relação é válida

$$\begin{aligned}\frac{d}{dt^*}C_j &= 2N_e\gamma C_{j-2} + 2N_e\beta C_{j-1} - (1 + 4N_e\beta + 4N_e\gamma)C_j + 2N_e\beta C_{j+1} + 2N_e\gamma C_{j+2} \\ &= vC_{j-2} + uC_{j-1} - (1 + 2u + 2v)C_j + uC_{j+1} + vC_{j+2}.\end{aligned}$$

Da mesma forma, no equilíbrio iguala-se essa equação à zero e obtém-se a seguinte relação de recorrência entre os C_j's

$$(1 + 2u + 2v)C_j = vC_{j-2} + uC_{j-1} + uC_{j+1} + vC_{j+2} \quad j = 1, 2, 3, \ldots.$$

Então, no modelo de mutação de dois passos, as seguintes relações são válidas

$$\begin{aligned}(1 + 2u + 2v)C_0 &= 1 + 2uC_1 + 2vC_2 \\ (1 + 2u + 2v)C_1 &= vC_1 + uC_0 + uC_2 + vC_3 \\ (1 + 2u + 2v)C_j &= vC_{j-2} + uC_{j-1} + uC_{j+1} + vC_{j+2} \quad j = 2, 3, ... \end{aligned} \tag{2.3.13}$$

Devido à relação (2.3.13), segundo Li (1955) é possível escrever

$$C_j = a_1\lambda_1^j + a_2\lambda_2^j + a_3\lambda_3^j + a_4\lambda_4^j.$$

Para obter λ_m ($m = 1, 2, 3$ e 4) substituímos $C_j = a\lambda^j$ na relação (2.3.13) e temos

$$[1 + 2u + 2v] = u\left[\lambda + \frac{1}{\lambda}\right] + v\left[\lambda + \frac{1}{\lambda}\right]^2 - 2v,$$

cuja demonstração se encontra em Apêndice A número 2.

Substituindo $k = \lambda + \frac{1}{\lambda}$ e resolvendo a equação

$$vk^2 + uk - [1 + 4v + 2u] = 0,$$

têm-se

$$k = \frac{u \pm \sqrt{u^2 + 4v(1 + 4v + 2u)}}{2v} \quad v \neq 0.$$

De $k = \lambda + \frac{1}{\lambda}$ têm-se as seguintes equações

$$\lambda^2 - k\lambda + 1 = 0 \quad \text{e} \quad \lambda = \frac{k \pm \sqrt{k^2 - 4}}{2}.$$

Definindo $S = \sqrt{u^2 + 4v(1 + 4v + 2u)}$ as quatro possíveis raízes, são

$$\begin{aligned}
\lambda_1 &= \frac{S - u - \sqrt{(S-u)^2 - (4v)^2}}{4v} \\
\lambda_2 &= \frac{-S - u - \sqrt{(-S-u)^2 - (4v)^2}}{4v} \\
\lambda_3 &= \frac{S - u + \sqrt{(S-u)^2 - (4v)^2}}{4v} \\
\lambda_4 &= \frac{-S - u + \sqrt{(-S-u)^2 - (4v)^2}}{4v}.
\end{aligned}$$

Pode ser mostrado que $\lambda_1 = \lambda_3^{-1}$ e que $\lambda_2 = \lambda_4^{-1}$ (ver Apêndice A número 3).

Desta forma como $0 < C_j < 1$ para todo inteiro j e $\lambda_4 < -1 < \lambda_2 < 0 < \lambda_1 < 1 < \lambda_3$, segue que $a_3 = a_4 = 0$. Os valores λ_3 e λ_4 são apropriados para as quantidades simétricas C_{-j}, que numericamente é igual à C_j.

Então, o valor esperado C_j pode ser encontrado pela relação (2.3.10) , em que os parâmetros λ_1 e λ_2 devem ser estimados a partir dos dados. Da relação (2.3.2), segue que

$$a_1 \frac{1 + \lambda_1}{1 - \lambda_1} + a_2 \frac{1 + \lambda_2}{1 - \lambda_2} = 1, \tag{2.3.14}$$

enquanto que a relação de recorrência (2.3.13) nos dá

$$a_1 \frac{1 - \lambda_1^2}{\lambda_1} + a_2 \frac{1 - \lambda_2^2}{\lambda_2} = 0, \tag{2.3.15}$$

ver demonstração no Apêndice A número 4.

Utilizando as relações (2.3.14) e (2.3.15), podemos encontrar uma solução para a_1 e a_2

$$\begin{aligned}
a_1 &= \frac{\lambda_1(1 - \lambda_1)(1 - \lambda_2)^2}{(1 + \lambda_1)(\lambda_1 - \lambda_2)(1 - \lambda_1\lambda_2)} \\
a_2 &= \frac{-\lambda_2(1 - \lambda_2)(1 - \lambda_1)^2}{(1 + \lambda_2)(\lambda_1 - \lambda_2)(1 - \lambda_1\lambda_2)}.
\end{aligned}$$

Utilizando esses valores para a_1 e a_2, o C_j pode ser escrito como

$$C_j = \frac{(1 - \lambda_1)(1 - \lambda_2)}{(1 + \lambda_1)(1 + \lambda_2)(\lambda_1 - \lambda_2)(1 - \lambda_1\lambda_2)} \left[(1 - \lambda_2^2)\lambda_1^{j+1} - (1 - \lambda_1^2)\lambda_2^{j+1}\right].$$

Uma vez que os λ's foram estimados, os parâmetros de mutação, $u = 2N_e\beta$ e $v = 2N_e\gamma$, podem ser dados por

$$\begin{aligned} u &= \frac{(\lambda_1 + \lambda_2)(1 + \lambda_1\lambda_2)}{(1-\lambda_1)^2(1-\lambda_2)^2} \\ v &= \frac{-\lambda_1\lambda_2}{(1-\lambda_1)^2(1-\lambda_2)^2}. \end{aligned}$$

A magnitude relativa do peso da mutação de um e dois passos β/γ ou u/v, segue diretamente de λ_1 e λ_2, sem o conhecimento do tamanho populacional N_e.

2.4 Momentos dos tamanhos alélicos

Considere uma população de indivíduos diplóides de tamanho efetivo N_e. Seja a variável aleatória $Y_i(t)$ representando o tamanho alélicos, ou o número de repetições da i-ésima cópia ($i = 1, 2 \ldots 2n$) no tempo t e que tenha distribuição com média $\eta = \sum ip_i(t)$ e variância σ^2. Brown et al. (1975) encontraram os momentos centrais da variável aletória $Y_i(t)$, considerando o modelo de mutação de 2 passos,

$$\begin{aligned} \mu_1 &= E(Y_i(t) - \eta) = 0; \\ \mu_2 &= E[Y_i(t) - \eta]^2 = \sigma^2 \approx 2N_e(\beta + 4\gamma); \\ \mu_3 &= E[Y_i(t) - \eta]^3 = 0 \\ \mu_4 &= E[Y_i(t) - \eta]^4 \approx 3[2N_e(\beta + 4\gamma)]^2 + 2N_e(\beta + 16\gamma)/4. \end{aligned}$$

Assim, todos os momentos centrais ímpares são zeros.

Brown et al. (1975) derivaram essas fórmulas utilizando um mesmo método que será exemplificado utilizando μ_4. Assim, considere que o incremento na quantidade $K = \sum p_i(t)(i-\eta)^4$ numa geração devido ao modelo de mutação "*stepwise*" de dois passos é

$$\begin{aligned} \Delta K &= \sum_i \left[(1-\beta-\gamma)p_i(t) + \frac{\beta}{2}(p_{i+1}(t) + p_{i-1}(t)) + \frac{\gamma}{2}(p_{i+2}(t) + p_{i-2}(t))\right](i-\eta)^4 + \\ &- \sum_i (i-\eta)^4 p_i(t) \end{aligned}$$

$$
\begin{aligned}
&= -\gamma\sum_i (i-\eta)^4 p_i(t) - \beta\sum_i (i-\eta)^4 p_i(t) + \frac{\beta}{2}\sum_i (i-\eta)^4 (p_{i+1}(t)+p_{i-1}(t)) + \\
&+ \frac{\gamma}{2}\sum_i (i-\eta)^4 (p_{i+2}(t)+p_{i-2}(t)) \\
&= -\gamma\sum_i (i-\eta)^4 p_i(t) - \beta\sum_i (i-\eta)^4 p_i(t) + \frac{\beta}{2}\left[\sum_i (i+1-\eta)^4 p_{i+1}(t) + \right. \\
&+ \left. 6\sum_i (i+1-\eta)^2 p_{i+1}(t) + 2 + \sum_i (i-1-\eta)^4 p_{i-1}(t) + 6\sum_i (i-1-\eta)^2 p_{i-1}(t)\right] + \\
&+ \frac{\gamma}{2}\left[\sum_i (i+2-\eta)^4 p_{i+2}(t) + 24\sum_i (i+2-\eta)^2 p_{i+2}(t) + 32 + \sum_i (i-2-\eta)^4 p_{i-2}(t) + \right. \\
&+ \left. 24\sum_i (i-2-\eta)^2 p_{i-2}(t)\right] \\
&= -\gamma K - \beta K + \beta K + 6\beta V + \beta + \gamma K + 24\gamma V + 16\gamma,
\end{aligned}
$$

em que

$$
\begin{aligned}
V &= \sum_i (i-\eta)^2 p_i(t) = \sum_i (i+1-\eta)^2 p_{i+1}(t) = \sum_i (i-1-\eta)^2 p_{i-1}(t) \quad \text{e} \\
K &= \sum_i (i-\eta)^4 p_i(t) = \sum_i (i+1-\eta)^4 p_{i+1}(t) = \sum_i (i-1-\eta)^4 p_{i-1}(t).
\end{aligned}
$$

Logo, $\Delta K = 6(\beta+4\gamma)V + (\beta+16\gamma)$.

Existe, além dessa fonte de variação em K um outro efeito que gera mudanças em $p_i(t)$, é ξ_i conforme definido por Ohta & Kimura (1973), representa a mudança devido a amostra aleatória dos gametas, o que gera mudanças em μ_4. Assim, considere

$$
\Delta K_\xi = \sum_i (p_i(t)+\xi_i)(i-\eta')^4 - \sum_i p_i(t)(i-\eta)^4 = \sum_i (p_i(t)+\xi_i)(i-\eta')^4 - K,
$$

em que η' é a média devido essa variação em $p_i(t)$, ou seja, $\eta' = \sum_i i(p_i(t)+\xi_i)$. Assim,

$$
\begin{aligned}
&\sum_i (p_i(t)+\xi_i)(i-\eta')^4 = \sum_i (p_i(t)+\xi_i)\left[(i-\eta) - \sum_j j\xi_j\right]^4 \\
&= \sum_i (p_i(t)+\xi_i)\left[(i-\eta)^4 - 4(i-\eta)^3\sum_j j\xi_j + 6(i-\eta)^2\left(\sum_j j\xi_j\right)^2 + \right. \\
&\left. -4(i-\eta)\left(\sum_j j\xi_j\right)^3 + \left(\sum_j j\xi_j\right)^4\right]
\end{aligned}
$$

$$= K - 4\sum_i p_i(t)(i-\eta)^3 \sum_j j\xi_j + 6V\left(\sum_j j\xi_j\right)^2 + \left(\sum_j j\xi\right)^4 + \sum_i \xi_i(i-\eta)^4 +$$
$$-4\sum_i \xi_i(i-\eta)^3 \sum_j j\xi_j + 6\sum_i \xi_i(i-\eta)^2\left(\sum_j j\xi_j\right)^2 - 4\sum_i \xi_i(i-\eta)\left(\sum_j \xi_j\right)^3 +$$
$$+\sum_i \xi_i\left(\sum_j j\xi_j\right)^4 .$$

Os momentos de ξ_i são dados conforme anteriormente, ou seja, $E_\xi(\xi_i) = 0$, $E_\xi(\xi_i\xi_j) = -p_ip_j/2N_e$ (para $i \neq j$) e $E_\xi(\xi_i^2) = p_i(1-p_i)/2N_e$. Os momentos de ordem maior que 2 de ξ_i são da ordem de $2N_e^{-2}$ e então serão desconsiderados do cálculo (Kimura, 1955). Então, temos

$$\begin{aligned}
E_\xi[\Delta K_\xi] &\simeq \frac{3V^2}{N_e} + \frac{\eta}{N_e}\sum_j (j-\eta)p_j + \frac{\eta^2}{2N_e} - \frac{1}{2N_e}\sum_j j^2 p_j^2(t) + \\
&- \frac{1}{2N_e}\sum_j \sum_{j'>j} jj' p_j(t)p_{j'}(t) - \frac{2}{N_e}\left(\sum_j j(j-\eta)^3 p_j(t)(1-p_j(t)) + \right. \\
&- \left. 2\sum_j \sum_{j'} (j-\eta)^3 j' p_j(t)p_{j'}(t)\right) \\
&= \frac{3V^2}{N_e} + \frac{\eta^2}{2N_e} - \frac{1}{2N_e}\sum_j \sum_{j'} jj' p_j(t)p_{j'}(t) - \frac{2}{N_e}\left(\sum_j j(j-\eta)^3 p_j(t) + \right. \\
&- \left. \sum_j j(j-\eta)^3 p_j^2(t) - 2\sum_j \sum_{j'>j} (j-\eta)^3 j' p_j(t)p_{j'}(t)\right) \\
&= \frac{3V^2}{N_e} + \frac{\eta^2}{2N_e} - \frac{\eta^2}{2N_e} - \frac{2}{N_e}\left(\sum_j (j-\eta)^4 p_j(t) + \eta\sum_j (j-\eta)^3 p_j(t) + \right. \\
&- \left. \sum_j \sum_{j'} j'(j-\eta)^3 p_j(t)p_{j'}(t)\right) \\
&= \frac{3V^2}{N_e} - \frac{2}{N_e}\left(\sum_j (j-\eta)^4 p_j(t) + \eta\sum_j (j-\eta)^3 p_j(t) - \eta\sum_j (j-\eta)^3 p_j(t)\right) \\
&= \frac{3V^2}{N_e} - \frac{2K}{N_e}.
\end{aligned}$$

Quando as freqüências alélicas permanecem constantes com o passar do tempo, a

mudança total em K é zero se $\Delta K_\xi + \Delta K = 0$. Então

$$\begin{aligned} K &= \frac{3V^2}{2} + \frac{6N_e}{2}(\beta + 4\gamma)V + \frac{N_e(\beta + 16\gamma)}{2} \\ &= \frac{3[2N_e(\beta + 4\gamma)]^2}{2} + \frac{3[2N_e(\beta + 4\gamma)]^2}{2} + \frac{2N_e(\beta + 16\gamma)}{4} \\ &= 3[2N_e(\beta + 4\gamma)]^2 + \frac{2N_e(\beta + 16\gamma)}{4}. \end{aligned}$$

Wehrhahn (1975) usou função geradora de probabilidades e o modelo de mutação "*stepwise*" para mostrar que para 2 alelos, denotados por 1 e 2, amostrados aleatoriamente de uma população, em que $Y_1(t)$ e $Y_2(t)$ representam seus respectivos tamanhos alélicos no tempo t, a esperança de $(Y_1(t) - Y_2(t))^2$ é $4N_e(\beta + 4\gamma)$.

Primeiramente, supondo o modelo de mutação "*stepwise*" de um passo, de forma que os alelos podem mutar de um passo para direita com probabilidade β_1 e um passo para esquerda com probabilidade β_{-1}, por *locus* por geração. Considere dois alelos 1 e 2, cujos os tamanhos alélicos são representados por $Y_1(t)$ e $Y_2(t)$, respectivamente. Seja $P_k^*(t)$ a probabilidade de que o alelo 1 ocupe o estado de k passos à direita do alelo 2 no tempo t. Assim, $P_k^*(t) = P^*[Y_1(t) - Y_2(t) = k \mid T = t]$. A diferença entre os alelos 1 e 2 é k no tempo t, então no tempo $t-1$ a diferença deve ser $k-1$, k ou $k+1$. Se for $k-1$, ocorreu a mutação do alelo 2 para direita (com probabilidade β_1) ou a mutação de 1 para esquerda (com probabilidade β_{-1}). Se a diferença entre 1 e 2 for $k+1$, a mutação do alelo 2 para esquerda ou a mutação do alelo 1 para direita deve ter ocorrido (com probabilidade β_{-1} e β_1, respectivamente). Então

$$\begin{aligned} P_k^*(t) &= (1 - 2\beta_{-1} - 2\beta_1)P_k^*(t-1) + (\beta_1 + \beta_{-1})\, P_{k-1}^*(t-1) + \\ &+ (\beta_{-1} + \beta_1)\, P_{k+1}^*(t-1). \end{aligned} \tag{2.4.1}$$

Seja $\beta_{-1} = \beta/2 = \beta_1$. Então a equação (2.4.1) pode ser reescrita na forma

$$P_k^*(t) = (1 - 2\beta)P_k^*(t-1) + \beta P_{k-1}^*(t-1) + \beta P_{k+1}^*(t-1).$$

A função geratriz de probabilidades é obtida por $H(z;t) = E\left[z^{Y_1(t)-Y_2(t)}\right] = \sum_{k=-\infty}^{\infty} z^k P_k^*(t)$.

Logo,

$$\begin{aligned} H(z;t) &= \sum_{k=-\infty}^{\infty} z^k \left[(1-2\beta)P_k^*(t-1) + \beta P_{k-1}^*(t-1) + \beta P_{k+1}^*(t-1)\right] \\ &= (1-2\beta)\sum_k P_k^*(t-1)z^k + \beta z \sum_{k-1} P_{k-1}^*(t-1)z^{k-1} + \beta z^{-1}\sum_{k+1} P_{k+1}^*(t-1)z^{k+1} \\ &= H(z;t-1)\left[1-2\beta+\beta z+\beta z^{-1}\right] = \left(1-2\beta+\beta z+\beta z^{-1}\right)^t. \end{aligned}$$

Para z perto de 1 temos a seguinte aproximação $\ln(1+z) \approx z$. Utilizando esse fato, a função geratriz de probabilidades para o tempo contínuo para $z \approx 1$ é

$$H(z;t) \approx exp\left[t\left(1-2\beta+\beta z+\beta z^{-1}\right)\right]. \tag{2.4.2}$$

A equação (2.4.2) pode ser generalizada para o caso em que tanto a mutação de um passo ocorre como também a de dois passos. Seja γ_{-2} e γ_2 as probabilidades de um alelo mutar 2 passos para esquerda e direita por geração, respectivamente. Considere $\gamma_{-2} = \gamma/2 = \gamma_2$. Então, generalizando (2.4.2), obtém-se

$$H(z;t) \approx exp\left\{t\left[1-2(\beta+\gamma)+\gamma z^{-2}+\beta z^{-1}+\beta z+\gamma z^2\right]\right\}. \tag{2.4.3}$$

Denote $F'(t)$ a probabilidade de dois alelos numa população finita serem obtidos por replicação de um gene no tempo t. Se todos os genes são idênticos em estado há $t_\circ$ gerações, então a probabilidade de aleatoriamente escolher dois alelos k passos à frente dos outros é

$$P_k(t_\circ) = P[Y_1(t_\circ) - Y_2(t_\circ) = k] = \int_0^{t_\circ} F'(t)P_k^*(t)dt + [1-F(t_\circ)]P_k^*(t_\circ).$$

Conseqüentemente, a função geratriz de probabilidades para população finita, cujos membros eram inicialmente idênticos em estado é

$$\begin{aligned} G(Z;t_\circ) &= \sum_{k=-\infty}^{\infty} P_k(t_\circ)Z^k \\ &= \sum_{k=-\infty}^{\infty} \int_0^{t_\circ} F'(t)P_k^* dt + [1-F(t_\circ)]P_k^*(t_\circ)Z^k \end{aligned}$$

$$
\begin{aligned}
&= \int_0^{t_\circ} F'(t) \left\{ \sum_{k=-\infty}^{\infty} P_k^* Z^k \right\} dt + [1 - F(t_\circ)] H(Z; t_\circ) \\
&= \int_0^{t_\circ} F'(t) H(Z; t) dt + [1 - F(t_\circ)] H(Z; t_\circ). \qquad (2.4.4)
\end{aligned}
$$

Wehrhahn (1975) considerou $F(t)$ como tendo distribuição exponencial com parâmetro $1/2N_e$, ou seja, $F(t) = 1 - e^{-t/2N_e}$ e então

$$F'(t) = \frac{e^{-t/2N_e}}{2N_e}. \qquad (2.4.5)$$

Neste ponto é conveniente considerar

$$
\begin{aligned}
a(Z) &= -1/2N_e - 2(\beta + \gamma) + \gamma Z^{-2} + \beta Z^{-1} + \beta Z + \gamma Z^2 \quad \text{e} \\
b(Z) &= -2(\beta + \gamma) + \gamma Z^{-2} + \beta Z^{-1} + \beta Z + \gamma Z^2.
\end{aligned}
$$

Substituindo (2.4.3), (2.4.4) e (2.4.5) na função geratriz

$$
\begin{aligned}
G(Z; t_\circ) &= \int_0^{t_\circ} (1/2N_e) exp[ta(Z)]dt + [1 - F(t_\circ)] exp[b(Z)t_\circ] \\
&= \frac{e^{a(Z)t_\circ}}{2N_e a(Z)} - \frac{1}{2N_e a(Z)} + e^{a(Z)t_\circ}.
\end{aligned}
$$

Por propridade da função geratriz de probabilidades, temos que

$$E[(Y_1(t_\circ) - Y_2(t_\circ))^m] = \left. \frac{\partial^m G(Z; t_\circ)}{\partial Z^m} \right|_{Z=1}.$$

Assim,

$$
\begin{aligned}
\frac{\partial G(Z; t_\circ)}{\partial Z} &= \frac{1}{2N_e} \left[\frac{a'(Z)}{a^2(Z)} + \frac{e^{a(Z)t_\circ} a'(Z)(a(Z)t_\circ - 1)}{a^2(Z)} \right] + e^{a(Z)t_\circ} t_\circ a'(Z) \\
\frac{\partial^2 G(Z; t_\circ)}{\partial Z^2} &= \frac{1}{2N_e} \left[\frac{a''(Z)a(Z) - 2a'(Z)}{a^3(Z)} + \right. \\
&+ \frac{\left[e^{a(Z)t_\circ} t_\circ (a'(Z))^2 (a(Z)t_\circ - 1) + e^{a(Z)t_\circ} a''(Z)(a(Z)t_\circ - 1) \right] a(Z)}{a^3(Z)} + \\
&+ \left. \frac{\left[e^{a(Z)t_\circ} a'(Z) \right] t_\circ a(Z) - 2e^{a(Z)t_\circ} (a'(Z))^2 (a(Z)t_\circ - 1)}{a^3(Z)} \right] + \\
&+ e^{a(Z)t_\circ} (a'(Z))^2 (t_\circ)^2 + e^{a(Z)t_\circ} t_\circ a''(Z).
\end{aligned}
$$

Derivando $a(Z)$ com respeito a Z, temos,

$$\begin{aligned} a'(Z) &= -2\gamma Z^{-3} - \beta Z^{-2} + \beta + 2\gamma Z \\ a''(Z) &= 6\gamma Z^{-4} + 2\beta Z^{-3} + 2\gamma \end{aligned}$$

e calculando o valor em $Z = 1$, temos

$$a(1) = -1/2N_e \quad a'(1) = 0 \quad \text{e} \quad a''(1) = 8\gamma + 2\beta. \tag{2.4.6}$$

Com isso,

$$\begin{aligned} \left.\frac{\partial G(Z;t_\circ)}{\partial Z}\right|_{Z=1} &= 0 \\ \left.\frac{\partial^2 G(Z;t_\circ)}{\partial Z^2}\right|_{Z=1} &= \frac{1}{2N_e}\left[\frac{a''(1)a(1)}{a^3(1)} + \frac{\left[e^{a(1)t_\circ}a''(1)(a(1)t_\circ - 1)\right]a(1)}{a^3(1)}\right] + e^{a(1)t_\circ}t_\circ a''(1), \end{aligned}$$

substituindo pelos valores das derivadas de $a(Z)$ calculadas em $Z = 1$, conforme (2.4.6), temos

$$\begin{aligned} \left.\frac{\partial^2 G(Z;t_\circ)}{\partial Z^2}\right|_{Z=1} &= \frac{1}{2N_e}\left[4N_e^2(8\gamma+2\beta) - 4N_e^2 e^{-t_\circ/2N_e}(8\gamma+2\beta)\left(1+\frac{t_\circ}{2N_e}\right)\right] + \\ &+ e^{-t_\circ/2N_e}t_\circ(8\gamma+2\beta) \\ &= 2N_e(8\gamma+2\beta)\left[1 - e^{-t_\circ/2N_e}\left(1+\frac{t_\circ}{2N_e}\right)\right] + e^{-t_\circ/2N_e}t_\circ(8\gamma+2\beta) \\ &= (8\gamma+2\beta)\left[2N_e - 2N_e e^{-t_\circ/2N_e}\left(1+\frac{t_\circ}{2N_e}\right) + t_\circ e^{-t_\circ/2N_e}\right] \\ &= 4N_e(\beta+4\gamma)\left[1 - e^{-t_\circ/2N_e}\right]. \end{aligned}$$

Então, a esperança da diferença dos tamanhos alélicos ao quadrado para população finita é

$$E[(Y_1(t_\circ) - Y_2(t_\circ))^2] = \left.\frac{\partial^2 G(Z;t_\circ)}{\partial Z^2}\right|_{Z=1} \approx 4N_e(\beta+4\gamma)(1 - e^{-t_\circ/2N_e}).$$

Quando $t_\circ \to \infty$, o valor esperado aproxima-se de um valor estável de $E[(Y_1(t_\circ) - Y_2(t_\circ))^2] \approx 4N_e(\beta+4\gamma)$. A taxa de aproximação para esse valor é independente da taxa de mutação.

Para o modelo de mutação "*stepwise*" de um passo $E[(Y_1(t_\circ) - Y_2(t_\circ))^2] \approx 4N_e\beta = \theta$. No Capítulo 3, vamos mostrar de outra maneira esse resultado, utilizando teoria de coalescência para *loci* de microsatélites.

Podemos ver que

$$\begin{aligned} E[(Y_1(t_\circ) - Y_2(t_\circ))^2] &= E[(Y_1(t_\circ) - \eta + \eta - Y_2(t_\circ))^2] \\ &= E[(Y_1(t_\circ) - \eta)^2] + E[(Y_2(t_\circ) - \eta)^2] + 2E[(Y_1(t_\circ) - \eta)(Y_2(t_\circ) - \eta)] \\ &= \sigma_1^2 + \sigma_2^2 \;=\; 2\sigma^2 \;\approx\; 4N_e(\beta + 4\gamma). \end{aligned}$$

Note que $E[(Y_1(t_\circ) - \eta)(Y_2(t_\circ) - \eta)] = 0$, pois assume-se que os alelos são independentes, assim a variância esperada do número de repetições por alelo, é aproximadamente $\sigma^2 \approx 2N_e(\beta + 4\gamma)$.

No Capítulo 3 mostraremos que a aproximação é válida para o modelo de mutação "*stepwise*" de um passo. Observe que esse parâmetro é somente a variância na mudança no número de repetições em uma geração $(\beta + 4\gamma)$ multiplicada pelo número esperado de gerações $(2N_e)$. Se $\gamma = 0$, tem-se o modelo de um passo e então $\hat{\sigma}^2$ é um estimador de $2N_e\beta = \theta/2$. Considerando a aproximação obtida por Brown et al. (1975), σ^2 é aproximadamente $u + 4v$ para o modelo de mutação de 2 passos.

Capítulo 3

Distância genética para *loci* de microsatélites baseada nos desvios quadráticos

3.1 Introdução

O processo mutacional em *loci* de microsatélites pode ser modelado pelo modelo de mutação "*stepwise*", que foi proposto por Ohta & Kimura (1973), descrito no Capítulo 2. Nos *loci* de microsatélites, o novo alelo gerado de uma mutação depende do comprimento do alelo que sofreu mutação. Segundo Valdes et al. (1993), os estudos em humanos mostraram que essa dependência é de uma a duas unidades de repetição. Neste Capítulo consideraremos o modelo de mutação "*stepwise*" de um passo, para estudar medidas de distância genética em *loci* de microsatélites.

O processo mutacional "*stepwise*" não apaga informação sobre o estado ancestral. Segundo Pritchard & Feldman (1996), a teoria geral de genética populacional não é imediatamente aplicável a dados de microsatélites, devido a esse modelo de mutação proposto.

Geralmente, o processo de coalescência é aplicado, principalmente, para modelos de sítios infinitos, que supõe que a cada mutação surge um novo alelo, que não é encontrado

na população, resultando num número infinito de alelos. As suposições deste modelo não são razoáveis para *loci* de microsatélites, pois temos variação diferente em cada *locus* e, desta forma, avaliamos o que acontece *locus* a *locus*.

Neste Capítulo, faremos uma extensão da teoria de processo de coalescência à teoria aplicada a dados de microsatélites, ou seja, ao modelo de mutação "*stepwise*".

Com a descoberta de polimorfismo em microsatélites (VNTR-*variable number of tandem repeat*), medidas de distância genética foram sugeridas por diversos autores para a análise de variação genética com base no número de repetições. Introduziremos medidas de distância baseada no processo de coalescência, definidas por Goldstein et al. (1995) e Slatkin (1995), levando em conta o modelo de mutação "*stepwise*".

Uma área de aplicação da análise de distância genética é para comparação entre populações. Sugeriremos um teste de homogeneidade, para avaliar se há diferenças entre os tamanhos de repetições entre populações, ou seja, verificaremos se há homogeneidade entre populações em termos de microsatélites.

3.2 Modelo de coalescência para comparação de duas sequências

Primeiramente, considere uma única população com acasalamento aleatório composta de indivíduos diplóides de tamanho efetivo N_e, o que equivale à $2N_e$ cromossomos. Assume-se o modelo de mutação "*stepwise*" de um passo, pelo qual pode-se mutar em uma unidade de repetição com probabilidade $\mu/2$, por geração e por *locus*. Além disso, assume-se que a média μ de mutações é constante por *locus* e por geração, independente do genótipo e tamanho populacional. Assume-se que as mutações ocorrem independentemente em diferentes indivíduos e diferentes gerações.

A diferença em número de repetições entre duas seqüências, que será denotada por $\Delta(t)$, depende de três variáveis aleatórias, denotadas por T, $N(t)$ e $Y_{igl}(t)$. De forma análoga ao Capítulo 2, a variável aleatória $Y_{igl}(t)$ representa o número de unidades de

repetições da i-ésima cópia, g-ésima população no l-ésimo *locus* no tempo t.

Podemos definir, T como sendo a variável aleatória que representa o tempo desde o mais recente ancestral comum destas duas seqüências. Com isso,

- $T \sim exp(1/2N_e)$ ($1/2N_e$ é a probabilidade de que 2 seqüências tenham um ancestral comum num mesmo *locus* na geração anterior numa população com acasalamento aleatório).

Seja $N(t)$ o número de mutações que ocorreram em duas seqüências no tempo t desde o mais recente ancestral comum. Como no processo de coalescência, assume-se que as mutações ocorrem de acordo com um processo de Poisson com média $2\mu t$, em que μ é a taxa de mutação por *locus* e por geração e $2t$ é duas vezes o tempo de coalescência entre essas duas seqüências. Então,

- $N(t) \mid T = t \sim$ Poisson $(2\mu t)$.

Definição 3.1. *1. Seja $\Delta_{ii'l}(t) = Y_{il}(t) - Y_{i'l}(t)$ a diferença em número de repetições entre duas seqüências, i e i' em uma única população sem estrutura populacional, ou seja, supõe-se que só existe essa população, no locus l no tempo t.*

2. Seja $\Delta_{ii'ggl}(t) = Y_{igl}(t) - Y_{i'gl}(t)$ a diferença em número de repetições entre duas seqüências, i e i', numa mesma população g no locus l no tempo t.

3. Seja $\Delta_{ii'gg'l}(t) = Y_{igl}(t) - Y_{i'g'l}(t)$ a diferença em número de repetições entre duas seqüências, i e i', nas populações g e g', respectivamente, no locus l no tempo t.

Dado $N(t) = \alpha$, as diferenças em número de repetições entre duas sequências são representadas como um passeio aleatório simétrico em 0 com α passos, com igual probabilidade de ir para direita ou para esquerda em cada passo.

Assim,

- $\Delta_{ii'l}(t) \mid T = t, N(t) = \alpha$, é um passeio aleatório simétrico em 0 com α passos e

- $\Delta_{ii'gg'l}(t) \mid T = t, N(t) = \alpha$, para $g, g' = 1, \ldots, G$ é um passeio aleatório simétrico em 0 com α passos.

Proposição 3.1. *(James (2004))(**Princípio da substituição para distribuição condicional**.) Sejam X e Y variáveis aleatórias em (Ω,$\mathbb{A}$,P), $\varphi(x, y)$ uma função. Se a distribuição condicional de X dado Y é*

$$P(X \in B \mid Y = y), \ B \in \mathcal{B}, \ y \in \mathbb{R},$$

então a distribuição condicional para $\varphi(X, Y)$ dado Y é

$$P(\varphi(X, Y) \in B \mid Y = y) = P(\varphi(X, y) \in B \mid Y = y), \ y \in \mathbb{R}.$$

Essa proposição é importante pois estamos trabalhando com distribuições condicionais. Vamos admitir verdadeiro esse princípio para os futuros cálculos.

Definição 3.2. *Sejam $Z_1, Z_2, \ldots$ variáveis aleatórias independentes com distribuição desconhecida e comum f, com média 0 e variância ϖ_i^2, para $i = 1, 2, \ldots, .$ Seja X_0 uma variável aleatória, representando um valor inteiro independente de Z_i's e $X_\alpha = X_0 + Z_1 + Z_2 + \ldots + Z_\alpha$. A seqüência X_α, $\alpha \geq 1$ é um passeio aleatório de α passos.*

Teorema 3.1. *(Révész (1990)) Considere um passeio aleatório simétrico em 0, que move um passo para direita com probabilidade 1/2 e um passo para esquerda com probabilidade 1/2, durante uma unidade de tempo. A distribuição exata desse passeio aleatório com 2α passos é*

$$P[X_{2\alpha} = 2k] = \begin{pmatrix} 2\alpha \\ \alpha - k \end{pmatrix} 2^{-2\alpha}, \tag{3.2.1}$$

em que $k = -\alpha, -\alpha + 1, \ldots, \alpha$; $\alpha = 1, 2, \ldots,$ e para $2\alpha + 1$ é

$$P[X_{2\alpha+1} = 2k + 1] = \begin{pmatrix} 2\alpha + 1 \\ \alpha - k \end{pmatrix} 2^{-2\alpha-1}, \tag{3.2.2}$$

em que $k = -\alpha - 1, -\alpha, \ldots, \alpha$; $\alpha = 1, 2, \ldots$. *De (3.2.1) e (3.2.2), temos*

$$P[X_\alpha = k] = \begin{cases} \binom{\alpha}{\frac{\alpha-k}{2}} 2^{-\alpha}, & \text{se } k \equiv \alpha \text{ (mod 2);} \\ \\ 0\ , & \text{caso contrário,} \end{cases}$$

em que $k = -\alpha, -\alpha + 1, \ldots, \alpha$; $\alpha = 1, 2, \ldots$. *Então, para algum* $\alpha = 1, 2, \ldots, j \in \mathbb{R}$ *temos,*

$$\mathrm{E}[X_\alpha] = 0, \quad \mathrm{E}[X_\alpha^2] = \alpha \quad e \quad \mathrm{E}[exp(tX_\alpha)] = \left(\frac{e^t + e^{-t}}{2}\right)^\alpha,$$

em que $\mathrm{E}[exp(tX_\alpha)] = M_{X_\alpha}(t)$ *é a função geratriz de momentos de* X_α.

A demonstração deste teorema se encontra no Apêndice A número 5.

Considere as variáveis aleatórias $Y_{1l}(t)$ e $Y_{2l}(t)$ representando o número de unidades de repetições da cópia 1 e 2, respectivamente, no l-ésimo *locus* no tempo t. Logo, pela Definição 3.1, $\Delta_{12l}(t) = Y_{1l}(t) - Y_{2l}(t)$. Definido Z_i's como as variáveis aleatórias independentes e identicamente distribuídas, com média 0 e variância ϖ_i^2, independente do número de repetições, com $X_0 = 0$, temos que $\Delta_{12l}(t) = \sum_{i=1}^{\alpha(t)} Z_i$. Considerando o modelo de mutação "*stepwise*" de um passo, temos que $\varpi_i^2 = 1$.

Pelo Teorema 3.1, a função geratriz de momentos para Δ num passeio aleatório simétrico de α passos é $M_\Delta(x) = E(e^{x\Delta}) = \left(\frac{e^x + e^{-x}}{2}\right)^\alpha$. Assim, podemos calcular o valor esperado de Δ^m diferenciando a expressão acima m vezes com respeito a x e avaliando em $x = 0$. Desta forma, temos que

$$\begin{aligned} E[\Delta_{12l}(t) \mid N(t) = \alpha, T = t] &= 0, \\ E[\Delta_{12l}^2(t) \mid N(t) = \alpha, T = t] &= \alpha, \\ E[\Delta_{12l}^3(t) \mid N(t) = \alpha, T = t] &= 0 \quad \text{e} \\ E[\Delta_{12l}^4(t) \mid N(t) = \alpha, T = t] &= 3\alpha^2 - 2\alpha, \end{aligned} \tag{3.2.3}$$

mais detalhes ver Apêndice A número 6.

Para obtermos os momentos de $\Delta_{12l}(t)$ precisamos do primeiro e segundo momentos de $N(t)$, que tem distribuição de Poisson com média $2\mu t$ e de T, que tem distribuição exponencial com parâmetro $1/2N_e$, conforme foi descrito no início desta Seção. Admite-se que a taxa de mutação é igual para toda população. Assim,

$$\begin{aligned}
E[\Delta_{12l}(t)] &= \int_0^\infty \sum_{\alpha=0}^{\infty} \sum_{\delta=-\alpha}^{\alpha} \delta P[\Delta_{12l}(t)=\delta \mid N(t)=\alpha, T=t] P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty \sum_{\alpha=0}^{\infty} E[\Delta_{12l}(t) \mid N(t)=\alpha, T=t] P[N(t)=\alpha \mid T=t] f_T(t)dt = 0, \\
E[\Delta_{12l}^2(t)] &= \int_0^\infty \sum_{\alpha=0}^{\infty} \sum_{\delta=-\alpha}^{\alpha} \delta^2 P[\Delta_{12l}(t)=\delta \mid N(t)=\alpha, T=t] P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty \sum_{\alpha=0}^{\infty} E[\Delta_{12l}^2(t) \mid N(t)=\alpha, T=t] P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty \sum_{\alpha=0}^{\infty} \alpha P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty E[N(t) \mid T=t] f_T(t)dt \;=\; \int_0^\infty 2\mu t f_T(t)dt \;=\; 2\mu E[T] \;=\; 4\mu N_e.
\end{aligned}$$

Da mesma forma que o primeiro momento,

$$E[\Delta_{12l}^3(t)] \;=\; 0.$$

Além disso, o quarto momento de $\Delta_{12l}(t)$ é

$$\begin{aligned}
E[\Delta_{12l}^4(t)] &= \int_0^\infty \sum_{\alpha=0}^{\infty} \sum_{\delta=-\alpha}^{\alpha} \delta^4 P[\Delta_{12l}(t)=\delta \mid N(t)=\alpha, T=t] P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty \sum_{\alpha=0}^{\infty} E[\Delta_{12l}^4(t) \mid N(t)=\alpha, T=t] P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty \sum_{\alpha=0}^{\infty} (3\alpha^2 - 2\alpha) P[N(t)=\alpha \mid T=t] f_T(t)dt \\
&= \int_0^\infty (3E[N(t)^2 \mid T=t] - 2E[N(t) \mid T=t]) f_T(t)dt \\
&= \int_0^\infty (3(2\mu t + 4\mu^2 t^2) - 4\mu t) f_T(t)dt \\
&= \int_0^\infty (12\mu^2 t^2 + 2\mu t) f_T(t)dt \;=\; 12\mu^2 E[T^2] + 2\mu E[T] \;=\; 96\mu^2 N_e^2 + 4\mu N_e,
\end{aligned}$$

em que

$$E[N(t) \mid T = t] = 2\mu t;$$
$$Var[N(t) \mid T = t] = 2\mu t = E[N(t)^2 \mid T = t] + (E[N(t) \mid T = t])^2;$$
$$E[T] = 2N_e \quad \text{e} \quad E[T^2] = 8N_e^2.$$

Sabemos que $\theta = 4N_e\mu$, como definido no Capítulo 2. Logo

$$E[\Delta^2_{12l}(t)] = \theta \quad \text{e} \quad E[\Delta^4_{12l}(t)] \ = \ 6\theta^2 + \theta. \tag{3.2.4}$$

Com isso, $Var[\Delta^2_{12l}(t)] = E[\Delta^4_{12l}(t)] - \left(E[\Delta^2_{12l}(t)]\right)^2 \ = \ 5\theta^2 + \theta$.

Usando o mesmo argumento para uma única população, considere que temos G populações independentes e que não ocorra migração entre estas populações. Cada população é constituída de N_e indivíduos diploídes e com acasalamento aleatório. Supõe-se que não haja coalescência entre as populações e que $T_{\tau_{ggl}}$ é o tempo desde o mais recente ancestral comum de duas seqüências somado ao tempo de uma possível divergência populacional, isto é, $T_{\tau_{ggl}} = T + \tau_{ggl}$, em que τ_{ggl} é o parâmetro que denota o tempo que essa população pode sofrer divergência e T representa o tempo desde o mais recente ancestral comum destas duas seqüências e tem distribuição exponencial com média $2N_e$ (mais detalhes ver Figura 3.1). Considere as variáveis aleatórias $Y_{1gl}(t)$ e $Y_{2gl}(t)$ representando o número de unidades de repetições da cópia 1 e 2, respectivamente, na g-ésima população, no l-ésimo *locus* no tempo t. Logo, pela Definição 3.1, $\Delta_{12ggl}(t) = Y_{1gl}(t) - Y_{2gl}(t)$. O número de mutações num intervalo, $N(t)$, tem distribuição Poisson, com média $2\mu t$, em que μ refere-se a taxa de mutação.

Temos que

$$E[T_{\tau_{ggl}}] = 2N_e + \tau_{ggl} \quad \text{e} \quad E[T^2_{\tau_{ggl}}] = 8N_e^2 + 4N_e\tau_{ggl} + \tau^2_{ggl}.$$

Logo,

$$E[\Delta^2_{12ggl}(t)] \ = \ 2\mu E[T_{\tau_{ggl}}] \ = \ 2\mu(2N_e + \tau_{ggl}) \ = \ \theta + 2\mu\tau_{ggl};$$
$$E[\Delta^4_{12ggl}(t)] \ = \ 12\mu^2 E[T^2_{\tau_{ggl}}] + 2\mu E[T_{\tau_{ggl}}]$$

$$
\begin{aligned}
&= 12\mu^2(8N_e^2 + 4N_e\tau_{ggl} + \tau_{ggl}^2) + 2\mu(2N_e + \tau_{ggl}) \\
&= 6\theta^2 + 12\theta\mu\tau_{ggl} + 12\mu^2\tau_{ggl}^2 + \theta + 2\mu\tau_{ggl},
\end{aligned}
$$

em que $\theta = 4N_e\mu$. Definimos $\rho_{ggl} = 2\mu\tau_{ggl}$ como o desvio no número esperado de mutações no *locus* l na população g provocado pelo tempo τ_{ggl}. Note que, θ é o número esperado de mutações por *locus* por geração, ρ_{ggl} é um desvio nesse número quando temos variações dentro da população e $\rho_{ggl} \geq 0$. Assim, temos

$$
\begin{aligned}
E[\Delta^2_{12ggl}(t)] &= \theta + \rho_{ggl}; \\
E[\Delta^4_{12ggl}(t)] &= 6\theta^2 + 6\theta\rho_{ggl} + 3\rho^2_{ggl} + \theta + \rho_{ggl} = 3(\theta + \rho_{ggl})^2 + 3\theta^2 + \theta + \rho_{ggl}; \\
Var[\Delta^2_{12ggl}(t)] &= 2(\theta + \rho_{ggl})^2 + 3\theta^2 + \theta + \rho_{ggl}.
\end{aligned}
$$

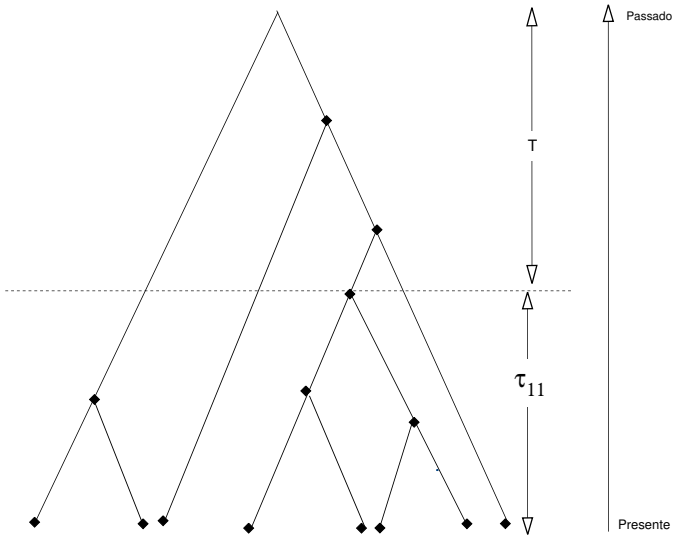

Figura 3.1: Desenho esquemático representando os tempos.

A Figura 3.1 representa um desenho esquemático dos tempos que são considerados para uma população. O tempo τ_{11} refere-se a um tempo fixo em que se esperaria a divergência populacional mas esta não ocorre.

Agora, considere duas populações que divergiram de uma única população ancestral, sem a ocorrência de migração depois da divergência. Cada população é composta de indivíduos diplóides com tamanho efetivo N_e e com acasalamento aleatório. A população

ancestral também tem tamanho efetivo N_e. O tempo de coalescência entre as duas populações é computado antes da divergência. Suponha que o tempo de coalescência entre duas populações é $\tau_{gg'l}$, em que $\tau_{gg'l}$ é o parâmetro que denota o tempo de coalescência entre as populações g e g', devido ao *locus* l. Supõe-se que a divisão das duas populações ocorreu devido à divergência física, ou seja, separação das populações por motivo de mudança comportamental ou geográfica, que possam ser geradas ou gerar mudanças em fatores genéticos associados a mutações em *loci* de microsatélites. Assim, o tempo de coalescência para comparação entre duas populações num mesmo *locus* l, $T_{\tau_{gg'l}}$, é uma variável aleatória com distribuição exponencial, $T_{\tau_{gg'l}} = T + \tau_{gg'l}$, em que T representa o tempo desde o mais recente ancestral comum de duas seqüências na população ancestral e tem distribuição exponencial com média $2N_e$. Da mesma forma, o número de mutações num intervalo, $N(t)$, tem distribuição Poisson, com média $2\mu t$, em que μ refere-se a taxa de mutação. Considere as variáveis aleatórias $Y_{1gl}(t)$ e $Y_{2g'l}(t)$ representando o número de unidades de repetições da cópia 1 na população g e da cópia 2 na população g', respectivamente, no l-ésimo *locus* no tempo t. Conforme a Definição 3.1, $\Delta_{12gg'l}(t) = Y_{1gl}(t) - Y_{2g'l}(t)$ tem a mesma distribuição de $\Delta_{12ggl}(t)$.

Assim, utilizando os momentos da relação (3.2.3) e os momentos de $T_{\tau_{gg'l}}$,

$$E[T_{\tau_{gg'l}}] = 2N_e + \tau_{gg'l} \quad \text{e} \quad E[T^2_{\tau_{gg'l}}] = 8N_e^2 + 4N_e\tau_{gg'l} + \tau^2_{gg'l}.$$

temos

$$\begin{aligned} E[\Delta^2_{12gg'l}(t)] &= 2\mu E[T_{\tau_{gg'l}}] = 2\mu(2N_e + \tau_{gg'l}) = \theta + 2\mu\tau_{gg'l}; \\ E[\Delta^4_{12gg'l}(t)] &= 12\mu^2 E[T^2_{\tau_{gg'l}}] + 2\mu E[T_{\tau_{gg'l}}] \\ &= 12\mu^2(8N_e^2 + 4N_e\tau_{gg'l} + \tau^2_{gg'l}) + 2\mu(2N_e + \tau_{gg'l}). \end{aligned}$$

Com $\theta = 4N_e\mu$. Definimos $\rho_{gg'l} = 2\mu\tau_{gg'l}$ como o desvio no número esperado de mutações no *locus* l por geração devido a divergência entre duas populações g e g', considerando o modelo de mutação "*stepwise*" de um passo. Note que, θ é o número esperado de mutações por *locus* por geração da população ancestral, $\rho_{gg'l}$ é um desvio nesse número

quando temos duas populações e $\rho_{gg'l} \geq 0$. Temos que

$$\begin{aligned} E[\Delta^2_{12gg'l}(t)] &= \theta + \rho_{gg'l}; \\ E[\Delta^4_{12gg'l}(t)] &= 6\theta^2 + 6\theta\rho_{gg'l} + 3\rho^2_{gg'l} + \theta + \rho_{gg'l}; \\ Var[\Delta^2_{12gg'l}(t)] &= 2(\theta + \rho_{gg'l})^2 + 3\theta^2 + \theta + \rho_{gg'l}. \end{aligned}$$

A Figura 3.2 representa um desenho esquemático da divisão de uma população ancestral em duas populações. O tempo τ_{12} refere-se ao tempo desde o mais recente ancestral comum para cada população 1 e 2 e $T + \tau_{12}$ é o tempo desde o mais recente ancestral comum na população ancestral.

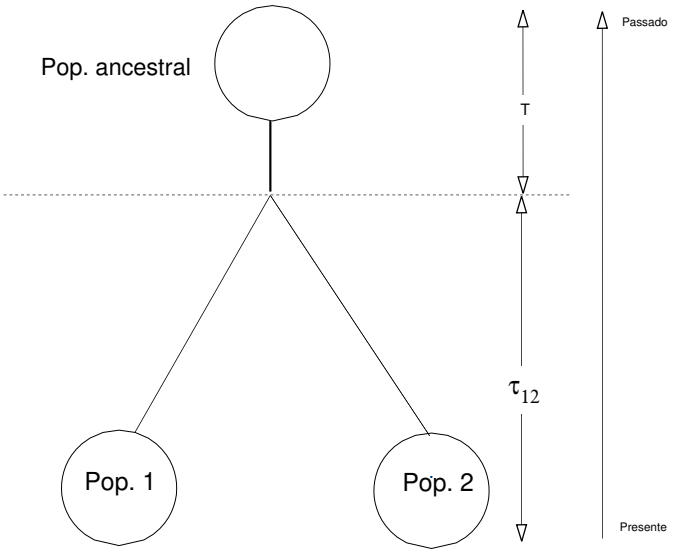

Figura 3.2: Desenho esquemático representando a divisão populacional.

Uma quantidade adicional de interesse é a probabilidade $\pi^\star$ de que dois cromossomos amostrados aleatoriamente de uma população não tenham o mesmo número de repetições num mesmo *locus* l. Essa quantidade é análoga a heterozigosidade de um gene. Claramente, temos que $\pi^\star = P[\Delta_{12l}(t) \neq 0]$. Observe que podemos ter $\Delta_{12l}(t) = 0$ quando α, o número de mutações, é par. Para que α seja par, temos que ter o mesmo número de passos (i) para direita e para esquerda. A probabilidade de ter i passos para direita e

para esquerda com $\alpha = 2i$ mutações, num passeio aleatório é

$$P[\Delta_{12l}(t) = 0 \mid N(t) = 2i, T = t] = \begin{pmatrix} 2i \\ i \end{pmatrix} \left(\frac{1}{2}\right)^{2i}.$$

A probabilidade de ter exatamente $2i$ mutações numa Poisson é

$$P[N(t) = 2i \mid T = t] = \frac{(2\mu t)^{2i} e^{-2\mu t}}{(2i)!}.$$

Então,

$$\begin{aligned}
\pi^\star &= P[\Delta_{12l}(t) \neq 0] = 1 - P[\Delta_{12l}(t) = 0] \\
&= 1 - \int_0^\infty \sum_{i=0}^{\infty} P[\Delta_{12l}(t) = 0 \mid N(t) = 2i, T = t] P[N(t) = 2i \mid T = t] f_T(t) dt \\
&= 1 - \int_0^\infty \sum_{i=0}^{\infty} \begin{pmatrix} 2i \\ i \end{pmatrix} \left(\frac{1}{2}\right)^{2i} \frac{(2\mu t)^{2i} e^{-2\mu t}}{(2i)!} \frac{1}{2N_e} e^{-t/2N_e} dt \\
&= 1 - \int_0^\infty \frac{1}{2N_e} e^{-\left(\frac{1}{2N_e} + 2\mu\right)t} \sum_{i=0}^{\infty} \frac{\left(\frac{(2\mu t)^2}{4}\right)^i}{(i!)^2} dt = 1 - \int_0^\infty \frac{1}{2N_e} e^{-\left(\frac{1}{2N_e} + 2\mu\right)t} I_0(2\mu t) dt,
\end{aligned}$$

em que $I_0(z)$ é a função de Bessel modificada definida por

$$I_0(z) = \sum_{i=0}^{\infty} \frac{\left(\frac{(z)^2}{4}\right)^i}{(i!)^2}. \tag{3.2.5}$$

Definição 3.3. *A função de Bessel modificada é obtida pela solução em w da equação diferencial*

$$z^2 \frac{d^2 w}{dz^2} + z \frac{dw}{dz} - (z^2 + \nu^2) w = 0,$$

em que $z = x + iy$, x e y são reais, o que faz de z uma variável complexa e ν um número real (Abramowitz & Stegun, 1972).

Segundo Abramowitz & Stegun (1972), $I_\nu(z)$ é solução dessa equação diferencial, em particular, $I_0(z)$ é solução para $\nu = 0$, com a restrição de que $I_0(z)$ seja limitada quando $z \to 0$. No nosso caso, z é um número real e I_0 é igual a 1 quando $z = 0$.

Se $s = 2\mu + 1/2N_e$. Então,

$$M_{I_0}(s) \;=\; \int_0^\infty e^{-st} I_0(2\mu t)dt$$

é a transformada de Laplace ou função geratriz de momentos da função de Bessel modificada. Temos que a função geratriz de momentos da função de Bessel modificada, na forma, $I_0(a\sqrt{t^2-k^2})u(t-k)$, em que

$$u(t) = \begin{cases} 0 & \text{se } t < 0; \\ \frac{1}{2}, & \text{se t=0;} \\ 1, & t > 0; \end{cases}$$

é dada por

$$M_{I_0}(s) = \frac{e^{-k\sqrt{s^2-a^2}}}{\sqrt{s^2-a^2}} \qquad (k \geq 0). \tag{3.2.6}$$

No nosso caso, $k = 0$, $a = 2\mu$ e $u(t) = 1$, pois $t > 0$. Com isso,

$$M_{I_0}(s) = \frac{1}{\sqrt{s^2-(2\mu)^2}}.$$

Substituindo s por $2\mu + 1/2N_e$, temos,

$$\begin{aligned} \pi^\star &= 1 - \frac{1}{2N_e} M_{I_0}(2\mu + 1/2N_e) = 1 - \frac{1}{2N_e} \frac{1}{\sqrt{(2\mu + 1/2N_e)^2 - (2\mu)^2}} \\ &= 1 - \frac{1}{\sqrt{(\theta+1)^2 - (\theta)^2}} = 1 - \frac{1}{\sqrt{2\theta+1}} \;=\; 1 - a_1, \quad a_1 \text{ de (2.3.9).} \end{aligned}$$

Na Figura 3.3, podemos ver que essa probabilidade tem crescimento, aproximadamente logarítmico para diferentes taxas de mutações, variando de 10^{-5} a 10^{-2}, com tamanhos populacionais efetivos de 5000 e 10000. Além disso, quanto maior a taxa de mutação maior a heterozigosidade e quanto menor a taxa de mutação, mais próxima ela fica de 0.

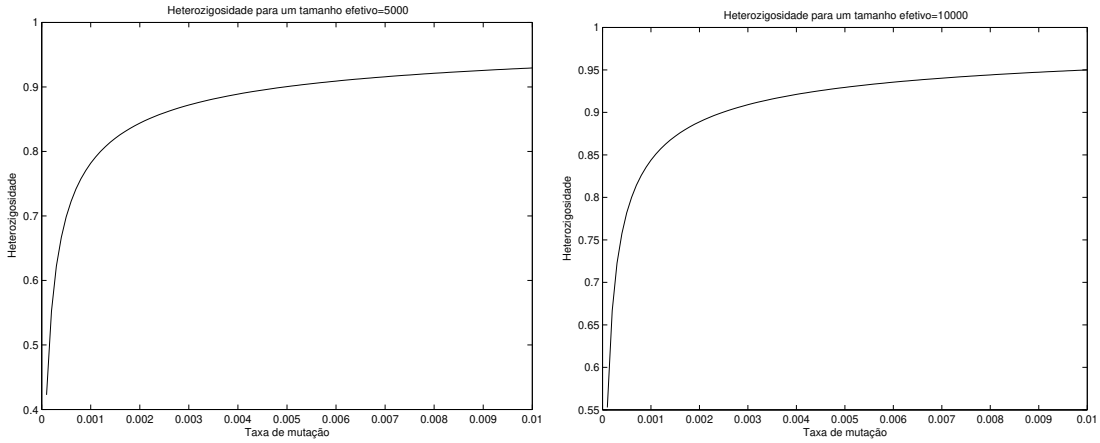

Figura 3.3: Heterozigosidade para diferentes taxas de mutação com tamanhos populacionais efetivos de 5000 e 10000, respectivamente.

3.3 Medidas de distância dentre e entre populações

Suponha G populações independentes, cada uma com tamanho efetivo $2N_e$. Considere que essas populações satisfaçam as suposições da teoria de coalescência introduzida na Seção 3.2. Assim, seja uma amostra de n indivíduos de cada população e $Y_{igl}(t)$ o número de repetições da i-ésima cópia ($i = 1, 2, \ldots, 2n$) na g-ésima população ($g = 1, 2, \ldots, G$) no l-ésimo *locus* ($l = 1, 2, \ldots, L$) no tempo t. Assume-se que as cópias sejam independentes umas das outras. Slatkin (1995) propôs como medida de distância dentre população a soma de quadrado das diferenças no número de repetições para a população g no l-ésimo *locus* no tempo t, isto é,

$$D_{ggl}(t) = \sum_{i=1}^{2n} \sum_{i'>i} \Delta^2_{ii'ggl}(t),$$

em que $\Delta_{ii'ggl}(t) = (Y_{igl}(t) - Y_{i'gl}(t))$. Então, define-se a distância média para a população g por

$$\bar{D}_{ggl}(t) = \frac{D_{ggl}(t)}{\begin{pmatrix} 2n \\ 2 \end{pmatrix}}.$$

A soma de quadrados dentre populações pode ser definida por $S_{Wl}(t) = \sum_{g=1}^{G} D_{ggl}(t)$ e o quadrado médio dentre populações é dado por

$$QM_{Wl}(t) = \frac{S_{Wl}(t)}{G\begin{pmatrix} 2n \\ 2 \end{pmatrix}} = \frac{1}{G}\sum_{g=1}^{G} \bar{D}_{ggl}(t).$$

Para estimar a soma de quadrados das diferenças entre os pares de cópias em diferentes populações, Slatkin (1995) propôs como medida de distância entre população,

$$D_{gg'l}(t) = \sum_{i=1}^{2n}\sum_{i'=1}^{2n} \Delta^2_{ii'gg'l}(t),$$

em que $\Delta_{ii'gg'l}(t) = (Y_{igl}(t) - Y_{i'g'l}(t))$. Então, a distância média entre as populações g e g' no l-ésimo *locus* no tempo t é

$$\bar{D}_{gg'l}(t) = \frac{D_{gg'l}(t)}{(2n)^2}.$$

A soma de quadrados entre populações, denotada por $S_{Bl}(t) = \sum_{g<g'} D_{gg'l}(t)$ e o quadrado médio entre populações é dado por

$$QM_{Bl}(t) = \frac{S_{Bl}(t)}{(2n)^2\begin{pmatrix} G \\ 2 \end{pmatrix}}.$$

Essas duas medidas são equivalentes àquelas introduzidas por Goldstein et al. (1995).

A soma de quadrados total é $S_{Totl}(t) = S_{Wl}(t) + S_{Bl}(t)$ e o quadrado médio total é

$$QM_{Totl}(t) = \frac{S_{Totl}(t)}{\begin{pmatrix} 2nG \\ 2 \end{pmatrix}}.$$

Note que, ignorando os grupos, temos que

$$S_{Totl}(t) = \sum_{1\le i<i'\le 2nG} \Delta^2_{ii'l}(t), \quad \text{com} \quad \begin{pmatrix} 2nG \\ 2 \end{pmatrix} \quad \text{comparações,} \qquad (3.3.1)$$

em que $\Delta_{ii'l}(t) = (Y_{il}(t) - Y_{i'l}(t))$, pois

$$\begin{aligned}
S_{Totl}(t) &= \left[\sum_{g=1}^{G}\sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t) - Y_{i'gl}(t))^2 + \sum_{g=1}^{G}\sum_{g'>g}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}(Y_{igl}(t) - Y_{i'g'l}(t))^2\right] \\
&= \left[\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}(Y_{igl}(t) - Y_{i'gl}(t))^2 + \frac{1}{2}\sum_{g=1}^{G}\sum_{g'\neq g}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}(Y_{igl}(t) - Y_{i'g'l}(t))^2\right] \\
&= \frac{1}{2}\left[\sum_{g=1}^{G}\sum_{g'=1}^{G}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}(Y_{igl}(t) - Y_{i'g'l}(t))^2\right] \quad \text{ignorando os grupos,} \\
&= \left[\sum_{i=1}^{2nG}\sum_{i'>i}(Y_{il}(t) - Y_{i'l}(t))^2\right],
\end{aligned}$$

então

$$S_{Totl}(t) = \sum_{g=1}^{G}\binom{2n}{2}\bar{D}_{ggl}(t) + \sum_{g=1}^{G}\sum_{g'>g}(2n)^2\bar{D}_{gg'l}(t) \quad \text{e}$$

$$QM_{Totl}(t) = \frac{2n-1}{2nG-1}QM_{Wl}(t) + \frac{2n(G-1)}{2nG-1}QM_{Bl}(t). \tag{3.3.2}$$

Pode-se mostrar que $QM_{Wl}(t)$ é 2 vezes a média das variâncias amostrais do número de repetições dentro de cada população no tempo t, ou seja,

$$QM_{Wl}(t) = \frac{2}{G}\sum_{g=1}^{G}S^2_{gl}(t),$$

em que $S^2_{gl}(t) = \dfrac{1}{2n-1}\left[\sum_{i=1}^{2n}Y^2_{igl}(t) - 2n\bar{Y}^2_{gl}(t)\right]$ e $\bar{Y}_{gl}(t) = \dfrac{1}{2n}\sum_{i=1}^{2n}Y_{igl}(t)$, ou seja, $S^2_{gl}(t)$ e $\bar{Y}_{gl}(t)$ são os estimadores da média e da variância do número de repetições na g-ésima população, no l-ésimo *locus*, no tempo t, respectivamente (demonstração em Apêndice A número 7).

Da mesma forma, pode-se mostrar que $QM_{Totl}(t)$ é 2 vezes a variância amostral do número de repetições na coleção de populações. Portanto, $QM_{Totl}(t) = 2S^2_l(t)$, em que

$$S^2_l(t) = \frac{1}{2nG-1}\left[\sum_{g=1}^{G}\sum_{i=1}^{2n}Y^2_{igl}(t) - 2ng\bar{Y}^2_l(t)\right] \quad \text{e} \quad \bar{Y}_l(t) = \frac{1}{2nG}\sum_{g=1}^{G}\sum_{i=1}^{2n}Y_{igl}(t),$$

ou seja, $\bar{Y}_l(t)$ e $S_l^2(t)$ são os estimadores da média e variância do número de repetições no l-ésimo *locus* e no tempo t, respectivamente (demonstração no Apêndice A número 8.

Para encontrar o valor esperado e a variância de $QM_{Wl}(t)$, teremos que utilizar a teoria de coalescência da Seção 3.2. Assim, temos que

$$E[QM_{Wl}(t)] = \frac{1}{G}\sum_{g=1}^{G}\frac{2}{2n(2n-1)}\sum_{i=1}^{2n}\sum_{i'>i}E[\Delta^2_{ii'ggl}(t)].$$

Foi visto, na Seção 3.2 que $E[\Delta^2_{ii'ggl}(t)] = \theta + \rho_{ggl}$. Utilizando as mesmas suposições feitas na Seção anterior temos que

$$E[QM_{Wl}(t)] = \theta + \frac{1}{G}\sum_{g=1}^{G}\rho_{ggl}.$$

Definindo $\bar{\rho}_{Wl} = \sum_{g=1}^{G}\rho_{ggl}/G$, como o desvio médio no número esperado de mutações dentre G populações, então, $QM_{Wl}(t)$ é um estimador não viesado para $\theta + \bar{\rho}_{Wl}$. A variância de $QM_{Wl}(t)$ é

$$\begin{aligned} Var[QM_{Wl}(t)] &= Var\left[\frac{1}{G}\sum_{g=1}^{G}\frac{2}{2n(2n-1)}\sum_{i=1}^{2n}\sum_{i'>i}\Delta^2_{ii'ggl}(t)\right] \\ &= \frac{1}{G^2}\sum_{g=1}^{G}\frac{4}{[2n(2n-1)]^2}\sum_{i=1}^{2n}\sum_{i'>i}Var[\Delta^2_{ii'ggl}(t)], \end{aligned}$$

assumindo que no tempo t as populações são independentes e as cópias também o são. Foi visto que $Var[\Delta^2_{ii'ggl}(t)] = 2(\theta+\rho_{ggl})^2 + 3\theta^2 + \theta + \rho_{ggl}$. Com isso,

$$\begin{aligned} Var[QM_{Wl}(t)] &= \frac{1}{G^2}\sum_{g=1}^{G}\frac{4}{[2n(2n-1)]^2}\sum_{i=1}^{2n}\sum_{i'>i}\left[2(\theta+\rho_{ggl})^2 + 3\theta^2 + \theta + \rho_{ggl}\right] \\ &= \frac{4}{2nG^2(2n-1)}\sum_{g=1}^{G}(\theta+\rho_{ggl})^2 + \frac{6\theta^2}{2nG(2n-1)} + \\ &+ \frac{2}{2nG^2(2n-1)}\sum_{g=1}^{G}(\theta+\rho_{ggl}). \end{aligned}$$

Para calcular o valor esperado de $QM_{Bl}(t)$, considere que o tempo de coalescência entre a população g e g' seja $\tau_{gg'l}$ gerações no *locus* l. Então, defina a variável aleatória

$T_{\tau_{gg'l}} = T + \tau_{gg'l}$, em que T tem distribuição exponencial com média $2N_e$. Assim, temos que

$$\begin{aligned} E(QM_{Bl}(t)) &= \frac{1}{G(G-1)} \sum_{g<g'} \frac{2}{(2n)^2} \sum_{i=1}^{2n} \sum_{i'=1}^{2n} E(\Delta^2_{ii'gg'l}(t)) \\ &= \frac{1}{G(G-1)} \sum_{g<g'} \frac{2}{(2n)^2} \sum_{i=1}^{2n} \sum_{i'=1}^{2n} (\theta + 2\mu\tau_{gg'l}) \\ &= \theta + \frac{2}{G(G-1)} \sum_{g<g'} \rho_{gg'l} \\ &= \theta + \bar{\rho}_{Bl}, \end{aligned}$$

ou seja, é um estimador não viesado para $\theta + \bar{\rho}_{Bl}$, em que $\bar{\rho}_{Bl}$ é o desvio médio no número esperado de mutações no *locus* l por gerações em G populações, com $\rho_{gg'l} \geq 0$. Ou seja, $\bar{\rho}_{Bl}$ é o desvio médio em θ no *locus* l, provocado pela divergência populacional.

A variância é dada por

$$\begin{aligned} Var(QM_{Bl}(t)) &= \frac{1}{[G(G-1)]^2} \sum_{g<g'} \frac{4}{(2n)^4} \sum_{i=1}^{2n} \sum_{i'=1}^{2n} Var(\Delta^2_{ii'gg'l}(t)) \\ &= \frac{4}{[2nG(G-1)]^2} \sum_{g<g'} \left[2(\theta + \rho_{gg'l})^2 + 3\theta^2 + \theta + \rho_{gg'l}\right] \\ &= \frac{8}{[2nG(G-1)]^2} \sum_{g<g'} (\theta + \rho_{gg'l})^2 + \frac{6\theta^2}{(2n)^2 G(G-1)} + \\ &+ \frac{4}{[2nG(G-1)]^2} \sum_{g<g'} (\theta + \rho_{gg'l}), \end{aligned}$$

assumindo que $\Delta_{ii'gg'l}(t)$'s sejam não correlacionados.

Podemos calcular o valor esperado de $QM_{Totl}(t)$, utilizando a relação (3.3.2). Desta forma,

$$E[QM_{Totl}(t)] = \theta + \frac{2n-1}{2nG-1}\bar{\rho}_{Wl} + \frac{2n(G-1)}{2nG-1}\bar{\rho}_{Bl}.$$

Podemos definir $QM_{Wl}(t)$ e $QM_{Bl}(t)$ para diferentes tamanhos amostrais entre as populações, ou seja, $2n_g$ para $g = 1, \ldots, G$. Assim,

$$D_{ggl}(t) = \sum_{i=1}^{2n_g} \sum_{i'>i} \Delta^2_{ii'ggl}(t).$$

Então a distância média para a população g no l-ésimo *locus* no tempo t é

$$\bar{D}_{ggl}(t) = \frac{D_{ggl}(t)}{\begin{pmatrix} 2n_g \\ 2 \end{pmatrix}}.$$

A soma de quadrados dentre populações é $S_{Wl}(t) = \sum_{g=1}^{G} D_{ggl}(t)$ e o quadrado médio dentre populações é

$$QM_{Wl}(t) = \frac{1}{G}\sum_{g=1}^{G} \bar{D}_{ggl}(t).$$

Da mesma forma podemos definir $QM_{Bl}(t)$,

$$D_{gg'l}(t) \quad = \quad \sum_{i=1}^{2n_g}\sum_{i'=1}^{2n_{g'}} \Delta^2_{ii'gg'l}(t).$$

Então, a distância média entre as populações g e g' no l-ésimo *locus* no tempo t é

$$\bar{D}_{gg'l}(t) = \frac{D_{gg'l}(t)}{2n_g 2n_{g'}}.$$

A soma de quadrados entre populações é $S_{Bl}(t) = \sum_{g<g'} D_{gg'l}(t)$ e o quadrado médio entre populações é

$$QM_{Bl}(t) = \frac{1}{\begin{pmatrix} G \\ 2 \end{pmatrix}} \sum_{g'>g} \bar{D}_{gg'l}(t).$$

Da mesma forma, para amostras não balanceadas $(n_1, n_2, \ldots, n_G)$, temos

$$QM_{Totl}(t) \quad = \quad \begin{pmatrix} \sum_{g=1}^{G} 2n_g \\ 2 \end{pmatrix}^{-1} \left(\sum_{g=1}^{G} \begin{pmatrix} 2n_g \\ 2 \end{pmatrix} \bar{D}_{ggl}(t) + \sum_{g=1}^{G}\sum_{g'>g} 2n_g 2n_{g'} \bar{D}_{gg'l}(t) \right).$$

Note que

$$E[QM_{Wl}(t)] = \theta + \bar{\rho}_{Wl} \qquad \text{e} \qquad E[QM_{Bl}(t)] = \theta + \bar{\rho}_{Bl}.$$

Além disso,

$$E[QM_{Totl}(t)] = \theta + \left(\begin{array}{c} \sum_{g=1}^{G} 2n_g \\ 2 \end{array} \right)^{-1} \left(\sum_{g=1}^{G} \left(\begin{array}{c} 2n_g \\ 2 \end{array} \right) \rho_{ggl} + \sum_{g=1}^{G} \sum_{g'>g} 2n_g 2n_{g'} \rho_{gg'l} \right).$$

Na prática, ver aplicação Capítulo 5, para cada população podemos ter tamanhos amostrais muito distintos. A população com tamanho amostral pequeno é menos informativa que as outras. Desta forma, define-se pesos para cada população, w_g $g = 1, \ldots, G$, para diminuir a influência desta população na estatística $QM_{Wl}(t)$. Assim, seja n_g o tamanho amostral da população g, sendo assim, temos $2n_g$ cópias. Define-se

$$w_g = \frac{2n_g}{\sum_{g=1}^{G} 2n_g} \quad \text{e} \quad QMP_{Wl}(t) = \sum_{g=1}^{G} w_g \bar{D}_{ggl}(t), \tag{3.3.3}$$

em que $QMP_{Wl}(t)$ é o quadrado médio ponderado. Note que $\sum_g w_g = 1$. Desta forma,

$$E[QMP_{Wl}(t)] = \sum_{g=1}^{G} w_g(\theta + \rho_{ggl}) = \theta + \bar{\rho}_{PWl}.$$

Para estudar a distribuição de $QM_{Wl}(t)$ e $QM_{Bl}(t)$ introduzimos a definição de Estatística U, suas propriedades e teoria assintótica.

3.4 Estatística U

Nesta Seção vamos considerar uma classe de estatísticas que foi introduzida por Hoeffding (1948).

Primeiramente, considere a estatística U de uma amostra. Sejam $X_1, X_2, \ldots$, variáveis aleatórias independentes com distribuição F. Considere uma função paramétrica $\theta = \theta(F)$ para a qual existe um estimador não viesado, $h(X_1, \ldots, X_m)$. Ou seja, $\theta(F)$ pode ser representado por

$$\theta(F) = E_F[h(X_1, \ldots, X_m)] = \int \ldots \int h(x_1, \ldots, x_m) dF(x_1) \ldots dF(x_m),$$

em que a função $h(x_1, \ldots, x_m)$ chamaremos de kernel. Sem perda de generalidade, podemos assumir que h é simétrico. Para o caso que não seja, devemos substituir por um kernel simétrico

$$\frac{1}{m!}\sum_{\pi} h(x_{i_1}, \ldots, x_{i_m}),$$

em que $\sum_{\pi}$ denota a soma sobre todas as $m!$ permutações $\{i_1, \ldots, i_m\}$ de $\{1, \ldots, m\}$.

Definição 3.4. *(Estatística U). Para algum kernel h, a estatística U correspondente a estimar θ baseado na amostra $X_1, \ldots, X_n$ de tamanho $n \geq m$ é obtida fazendo a média do kernel h, simetricamente, sobre todas as observações. Assim,*

$$U_n = U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_{c} h(X_{i_1}, \ldots, X_{i_m}),$$

em que $\sum_c$ denota a soma sobre as combinações de m elementos distintos $\{i_1, \ldots, i_m\}$ à partir de $\{1, \ldots, n\}$. Claramente U_n é um estimador não viesado para θ.

Exemplos.

(i) (Estatística U de grau 1) $\theta(F)$=média de $F = \mu(F) = \int x dF(x)$. Para o kernel $h(x) = x$, a estatística U correspondente é

$$U(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$

que é a média amostral.

(ii) (Estatística U de grau 2) $\theta(F) = \mu^2(F) = \left[\int x dF(x)\right]^2$. Para o kernel $h(x_1, x_2) = x_1 x_2$, a estatística U correspondente é

$$U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} X_i X_j.$$

(iii) (Estatística U de grau 2) $\theta(F)$=variância de $F = \sigma^2(F) = \int (x - \mu)^2 dF(x)$. Para o kernel

$$h(x_1, x_2) = \frac{x_1^2 + x_2^2 - 2x_1 x_2}{2} = \frac{1}{2}(x_1 - x_2)^2,$$

a estatística U correspondente é

$$\begin{aligned} U(X_1,\ldots,X_n) &= \frac{2}{n(n-1)}\sum_{1\le i<j\le n} h(X_i,X_j) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = S^2, \end{aligned}$$

que é a variância amostral.

(iv) (Estatística U de grau 1) $\theta(F) = F(t_0) = \int_{-\infty}^{t_0} dF(x) = P_F(X_1 \le t_0)$. Para o kernel $h(x) = I(x \le t_0)$, a estatística U correspondente é

$$U(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le t_0) = F_n(t_0),$$

em que F_n denota a função distribuição amostral.

(v) (Estatística U de grau 2) $\theta(F) = E_F|X_1 - X_2|$. Para o kernel $h(x_1,x_2) = |x_1 - x_2|$, a estatística U correspondente é

$$U(X_1,\ldots,X_n) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} |X_i - X_j|,$$

estatística conhecida por "diferença média de Gini".

A extensão para o caso de vários grupos foi obtida por Lehmann (1951) e Dwass (1956).

Definição 3.5. *(Estatística U generalizada) Considere k grupos independentes com observações independentes* $\left\{X_1^{(1)}, X_2^{(1)}, \ldots\right\}, \ldots, \left\{X_1^{(k)}, X_2^{(k)}, \ldots\right\}$, *obtidas das distribuições* $F^{(1)},\ldots,F^{(k)}$, *respectivamente. Seja a função paramétrica* $\theta = \theta\left(F^{(1)},\ldots,F^{(k)}\right)$, *da qual existe um estimador não viesado. Ou seja,*

$$\theta = \mathrm{E}\left[h\left(X_1^{(1)},\ldots,X_m^{(1)};\ldots;X_1^{(k)},\ldots,X_m^{(k)}\right)\right],$$

em que, sem perda de generalidade h é simétrico dentro de cada um dos k grupos de argumentos. Para esse kernel h, assumindo que $n_1 \ge m_1,\ldots,n_k \ge m_k$, *a estatística U*

para estimar θ é definida por

$$U^{(\boldsymbol{m})} = \frac{1}{\prod_{j=1}^{k} \begin{pmatrix} n_j \\ m_j \end{pmatrix}} \sum_c h\left(X_{i_1 1}^{(1)}, \ldots, X_{i_1 m_1}^{(1)}; \ldots; X_{i_k 2}^{(k)}, \ldots, X_{i_k m_k}^{(k)}\right) \quad ,$$

em que $\boldsymbol{m} = (m_1, m_2, \ldots, m_k)$ e $\{i_{j1}, \ldots, i_{jm_j}\}$ denota um conjunto de m_j elementos distintos do conjunto $\{1, 2, \ldots, n_j\}$, $1 \leq j \leq k$ e $\sum_c$ denota a soma sobre todas combinações.

Exemplo. (Estatística U generalizada de grau (1,1)) Estatística de Wilcoxon para 2 grupos. Seja $\{X_1, \ldots, X_{n1}\}$ e $\{Y_1, \ldots, Y_{n2}\}$ observações independentes de uma distribuição F e G, respectivamente. Então, o estimador não viesado de

$$\theta(F, G) = \int F dG = P(X \leq Y) \quad \text{é} \quad U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_i \leq Y_j).$$

3.4.1 Propriedades da Estatística U

A estatística U pode ser representada como o resultado "de condicionar" o kernel nas estatísticas de ordem. Isto é, para o kernel $h(x_1, \ldots, x_m)$ e uma amostra $X_1, \ldots, X_n$, para $n \geq m$, a estatística U correspondente pode ser representada por

$$U_n = E\left[h(X_1, \ldots, X_m) \mid \mathbf{X}_{(\cdot)}\right],$$

em que $\mathbf{X}_{(\cdot)} = \left(X_{(1)}, \ldots, X_{(n)}\right)$, o vetor ordenado das variáveis aleatórias. Uma implicação dessa representação é que alguma estatística $S = S\left(X_1, \ldots, X_n\right)$ para um estimador não viesado de $\theta = \theta(F)$ pode melhorar a estatística U condicionando no vetor de estatística de ordem.

Teorema 3.2. *Seja $S = S(X_1, \ldots, X_n)$ um estimador não viesado de $\theta(F)$ baseado na amostra $X_1, \ldots, X_n$ de distribuição F. A estatística U correspondente é também não viesada e $Var_F(U) \leq Var_F(S)$ com igualdade se somente se $P_F(U = S) = 1$.*

A prova deste Teorema se encontra em Apêndice A número 9.

A estatística de ordem é suficiente e completa para alguma família $\mathfrak{F}$ de distribuições, contendo F, a estatística U é um resultado do condicionamento numa estatística suficiente. O resultado precedente é um caso especial do Teorema de Rao-Blackwell (Casella & Berger, 2002).

Variância da estatística U e Propriedades assintóticas

Considere um Kernel simétrico $h(x_1, \ldots, x_n)$ satisfazendo $E[h^2(X_1, \ldots, X_m)] < \infty$. Chamaremos $h_m = h$ e, para $1 \leq c \leq m-1$,

$$h_c(x_1, \ldots, x_c) = E[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m)],$$

e $\tilde{h} = h - \theta$, $\tilde{h}_c = h_c - \theta$ $(1 \leq c \leq m)$, em que $\theta = \theta(F) = E_F[h(X_1, \ldots, X_m)]$, e que, para $1 \leq c \leq m-1$,

$$h_c(x_1, \ldots, x_c) = E[h_{c+1}(x_1, \ldots, x_c, X_{c+1})].$$

Note que

$$E_F[\tilde{h}_c(X_1, \ldots, X_c)] = 0, \quad 1 \leq c \leq m.$$

Defina $\zeta_0 = 0$ e, para $1 \leq c \leq m$,

$$\zeta_c = Var_F[h_c(X_1, \ldots, X_c)] = E_F[\tilde{h}_c^2(X_1, \ldots, X_c)].$$

Então, temos $0 = \zeta_0 \leq \zeta_1 \leq \ldots \leq \zeta_m = Var_F(h) < \infty$.

Exemplo A. $\theta(F) = \sigma^2(F)$. Escrevendo $\mu = \mu(F)$, $\sigma^2 = \sigma^2(F)$ e $\mu_4 = \mu_4(F)$ (o quarto momento de F), temos

$$\begin{aligned}
h(x_1, x_2) &= \frac{1}{2}(x_1^2 + x_2^2 - 2x_1x_2) = \frac{1}{2}(x_1 - x_2)^2, \\
\tilde{h}(x_1, x_2) &= h(x_1, x_2) - \sigma^2, \\
h_1(x) &= \frac{1}{2}(x^2 + \sigma^2 + \mu^2 - 2x\mu), \\
\tilde{h}_1(x) &= \frac{1}{2}(x^2 - \sigma^2 + \mu^2 - 2x\mu) = \frac{1}{2}[(x-\mu)^2 - \sigma^2],
\end{aligned}$$

$$\begin{aligned}
E[h^2] &= \frac{1}{4}E\{[(X_1-\mu)-(X_2-\mu)]^4\} \\
&= \frac{1}{4}\sum_{j=0}^{4}(-1)^{4-j}E[(X_1-\mu)^j]E[(X_2-\mu)^{4-j}] \\
&= \frac{1}{4}(2\mu_4+6\sigma^4), \quad \text{lembrando que } E[X_i-\mu]=0 \\
\zeta_2 &= E[h^2]-\sigma^4 = \frac{1}{2}(\mu_4+\sigma^4) \\
\zeta_1 &= E[\tilde{h}_1^2] = \frac{1}{4}Var_F[(X_1-\mu)^2] = \frac{1}{4}(\mu_4-\sigma^4).
\end{aligned}$$

Considere dois conjuntos $\{a_1,\ldots,a_m\}$ e $\{b_1,\ldots,b_m\}$ de m distintos elementos de $\{1,\ldots,m\}$ e seja c o número de inteiros comum para os dois conjuntos. Segue, por simetria de $\tilde{h}$ e por independência de $X_1,\ldots,X_n$ que

$$E_F\left[\tilde{h}(X_{a_1},\ldots,X_{a_m})\tilde{h}(X_{b_1},\ldots,X_{b_m})\right] = \zeta_c.$$

Note, também que o número de escolhas distintas para que os dois conjuntos tenham exatamente c elementos em comum é $\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}$.

Com essas suposições, podemos obter a variância da estatística U, escrevendo

$$U_n-\theta = \binom{n}{m}^{-1}\sum_{c}\tilde{h}(X_{i_1},\ldots,X_{i_m}),$$

temos

$$\begin{aligned}
Var_F(U_n) &= E_F[(U_n-\theta)^2] \\
&= \binom{n}{m}^{-2}\sum_{c}\sum_{(c)}E_F[\tilde{h}(X_{a_1},\ldots,X_{a_m})\tilde{h}(X_{b_1},\ldots,X_{b_m})] \\
&= \binom{n}{m}^{-2}\sum_{c=0}^{n}\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}\zeta_c,
\end{aligned}$$

em que o número de termos em $\sum_{(c)}$ é igual ao número de escolhas para que os dois conjuntos tenham exatamente c elementos em comum. Assim, temos o seguinte lema.

Lema 3.1. *A variância de U_n é dada por*

$$Var_F(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{n} \binom{m}{c} \binom{n-m}{m-c} \zeta_c$$

e satisfaz

(i) $\frac{m^2}{n}\zeta_1 \leq Var_F(U_n) \leq \frac{m}{n}\zeta_m$;
(ii) $(n+1)Var_F(U_{n+1}) \leq nVar_F(U_n)$;
(iii) $Var_F(U_n) = \frac{m^2\zeta_1}{n} + \mathrm{O}(n^{-2})$.

Exemplo B. (continuação do exemplo A).

$$\begin{aligned} Var_F(S^2) &= \binom{n}{2}^{-1} [2(n-2)\zeta_1 + \zeta_2] = \frac{4\zeta_1}{n} + \frac{2\zeta_2}{n(n-1)} - \frac{4\zeta_1}{n(n-1)} \\ &= \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)} = \frac{\mu_4 - \sigma^4}{n} + \mathrm{O}(n^{-2}). \end{aligned}$$

Então, se $E[h^2] < \infty$ e $\zeta_1 > 0$,

$$n^{1/2}(U_n - \theta) \xrightarrow{\mathcal{D}} N(0, m^2\zeta_1).$$

Teorema 3.3. *(Hoeffding, 1948) Se* $\mathrm{E}[h^2] < \infty$ *e* $\zeta_1 > 0$, *então*

$$n^{1/2}(U_n - \theta) \xrightarrow{\mathcal{D}} N(0, m^2\zeta_1).$$

Isto é U_n *é assintoticamente* $N\left(\theta, \frac{m^2\zeta_1}{n}\right)$.

Exemplo C. (continuação do exemplo B). $\theta(F) = \sigma^2(F)$ e

$$U_n = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Assumindo que é tal que $\sigma^4 < \mu_4 < \infty$ e então $E[h^2] < \infty$ e $\zeta_1 > 0$, obtemos que

$$n^{1/2}(S^2 - \sigma^2) \xrightarrow{\mathcal{D}} N\left(0, \mu_4 - \sigma^4\right),$$

e pelo teorema $S^2 \sim N\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right)$.

Considere a estatística U generalizada, conforme Definição 3.5, faz-se a extensão da teoria assintótica para esse caso. Para isso, considere d_j, tal que $0 \leq d_j \leq m_j$, $1 \leq j \leq k$, seja $\mathbf{d} = (d_1, \ldots, d_k)$ e

$$\Psi_{d_1,\ldots,d_k}\left(x_1^{(j)}, \ldots, x_{d_j}^{(j)}; 1 \leq j \leq k\right) = E\left[h(x_1^{(j)}, \ldots, x_{d_j}^{(j)}, X_{d_j+1}^{(j)}, \ldots, X_{m_j}^{(j)}; 1 \leq j \leq k)\right],$$

então $\Psi_0 = \theta(F)$, pois $h\left(X_1^{(j)}, \ldots, X_m^{(j)}; 1 \leq j \leq k\right)$ é um estimador não viesado para $\theta(F)$ e $\Psi_{\mathbf{m}} = h$, com $\mathbf{m} = (m_1, m_2, \ldots, m_k)$. Então,

$$\zeta_{\mathbf{d}} = E\left[\Psi_{\mathbf{d}}^2(X_1^{(j)}, \ldots, X_{d_j}^{(j)}; 1 \leq j \leq k)\right] - \theta^2(F), \quad \mathbf{0} \leq \mathbf{d} \leq \mathbf{m},$$

com $\zeta_{\mathbf{0}} = 0$. Então, para todo $\mathbf{n} \geq \mathbf{m}$

$$Var\left(U^{(\mathbf{m})}\right) = \sum_{j=1}^{k} n_j^{-1}\sigma_j^2[1 + \mathrm{O}(n_0^{-1})],$$

em que $n_0 = min\{n_1, \ldots, n_k\}$ e $\sigma_j^2 = m_j^2\zeta_{\delta_{j_1},\ldots,\delta_{j_k}}$, $j = 1, \ldots, k$ com $\delta_{\alpha\beta} = 1$ ou 0 se $\alpha = \beta$ ou não.

Então, se $E[h^2] < \infty$,

$$\gamma_{n_1,\ldots,n_k}^{-1}(U^{(\mathbf{m})} - \theta) \xrightarrow{\mathcal{D}} N(0,1),$$

quando $n_0 = min\{n_1, \ldots, n_k\} \to \infty$, em que

$$\gamma_{n_1,\ldots,n_k}^2 = \sum_{j=1}^{k} \frac{m_j^2\zeta_{\delta_{j_1},\ldots,\delta_{j_k}}}{n_j}.$$

Com isso, a estatística U generalizada tem distribuição assintoticamente normal com média θ e variância $\gamma_{n_1,\ldots,n_k}^2$.

Exemplo. A estatística U para 2 grupos de grau (m_1, m_2), o kernel é definido por

$$h(X_1, \ldots, X_{m_1}; Y_1, \ldots, Y_{m_2}) \quad \text{e}$$

$$U^{(m_1,m_2)} = \begin{pmatrix} n_1 \\ m_1 \end{pmatrix}^{-1} \begin{pmatrix} n_2 \\ m_2 \end{pmatrix}^{-1} \sum_{(\mathbf{n},\ \mathbf{m})} h\left(X_{i_1 1}^{(1)}, \ldots, X_{i_1 m_1}^{(1)}; Y_{i_2 1}^{(2)}, \ldots, Y_{i_2 m_2}^{(2)}\right),$$

em que $\sum_{(\mathbf{n},\mathbf{m})}$ estende sobre todos $1 \leq i_{j1} < \ldots < i_{jm_j} \leq n_j$, $j = 1, 2$. A estatística $U^{(m_1,m_2)}$ é um estimador não viesado de $\theta(F^{(1)}, F^{(2)})$.

$$\Psi_{d_1 d_2}(x_1, \ldots, x_{d_1}, y_1, \ldots, y_{d_2}) = E[h(x_1, \ldots, x_{d_1}, X_{d_1+1}, \ldots, X_{m_1}; y_1, \ldots, y_{d_2}, Y_{d_2+1}, \ldots, Y_{m_2}],$$

$$\zeta_{d_1 d_2} = E\left[\Psi^2_{d_1 d_2}(X_1, \ldots, X_{d_1}; Y_1, \ldots, Y_{d_2})\right] - \theta^2(F^{(1)}, F^{(2)}),$$

para $d_1 = 0, \ldots, m_1$, $d_2 = 0, \ldots, m_2$. ($\zeta_{00} = 0$). Então, se $E[h^2] < \infty$,

$$\gamma^{-1}_{n_1,n_2}\left(U^{(m_1,m_2)} - \theta\left(F^{(1)}, F^{(2)}\right)\right) \xrightarrow{\mathcal{D}} N(0,1),$$

em que $\gamma^2_{n_1,n_2} = \dfrac{m_1^2}{n_1}\zeta_{10} + \dfrac{m_2^2}{n_2}\zeta_{01}$.

A covariância de duas estatísticas U para uma amostra

Considere um conjunto de g estatísticas U,

$$U_\gamma = \begin{pmatrix} n \\ m_{(\gamma)} \end{pmatrix}^{-1} \sum_c h^{(\gamma)}(X_{\alpha_1}, \ldots, X_{\alpha_{m_{(\gamma)}}}), \qquad \gamma = 1, \ldots, g,$$

em que cada U_γ é função da mesma amostra independente, identicamente distribuída de tamanho n do vetor $X_1, \ldots, X_n$. Assume-se que a função $h^{(\gamma)}$ é simétrica nos $m_{(\gamma)}$ argumentos $\gamma = 1, \ldots, g$. Sejam

$$\begin{aligned}
E[U_\gamma] &= E[h^{(\gamma)}(X_1, \ldots, X_{m_{(\gamma)}})] = \theta^{(\gamma)}, \qquad \gamma = 1, \ldots, g; \\
\phi^{(\gamma)}(x_1, \ldots, x_{m_{(\gamma)}}) &= h^{(\gamma)}(x_1, \ldots, x_{m_{(\gamma)}}) - \theta^{(\gamma)}; \\
\phi^{(\gamma)}_c(x_1, \ldots, x_c) &= E[\phi^{(\gamma)}(x_1, \ldots, x_c, X_{c+1}, \ldots, X_{m_{(\gamma)}})], \quad c = 1, \ldots, m_{(\gamma)}; \\
\xi^{(\gamma,\upsilon)}_c &= E[\phi^{(\gamma)}_c(X_1, \ldots, X_c)\phi^{(\upsilon)}_c(X_1, \ldots, X_c)], \quad \gamma, \upsilon = 1, \ldots, g.
\end{aligned}$$

Em particular, se $\gamma = \upsilon$, então escrevemos,

$$\xi_c = \xi^{(\gamma,\gamma)}_c = E[\phi^{(\gamma)}_c(X_1, \ldots, X_c)]^2.$$

Seja,

$$\sigma(U_\gamma, U_\upsilon) = E[(U_\gamma - \theta_{(\gamma)})(U_\upsilon - \theta_{(\upsilon)})]$$

a covariância entre U_γ e U_υ.

Se $m_{(\gamma)} \leq m_{(\upsilon)}$, da mesma forma que para a variância, encontra-se que,

$$\sigma(U_\gamma, U_\upsilon) = \begin{pmatrix} n \\ m_{(\gamma)} \end{pmatrix}^{-1} \sum_{c=1}^{m_{(\gamma)}} \begin{pmatrix} m_{(\upsilon)} \\ c \end{pmatrix} \begin{pmatrix} n - m_{(\upsilon)} \\ m_{(\gamma)} - c \end{pmatrix} \xi_c^{(\gamma,\upsilon)}.$$

Para $\gamma = \upsilon$, $\sigma(U_\gamma, U_\upsilon)$ é a variância de U_γ. Segundo Hoeffding (1948),

$$\lim_{n\to\infty} n\sigma(U_\gamma, U_\upsilon) = m_{(\gamma)} m_{(\upsilon)} \xi_1^{(\gamma,\upsilon)}.$$

Assim, conforme o Lema 3.1 para a variância podemos fazer a seguinte aproximação

$$\sigma(U_\gamma, U_\upsilon) \approx \frac{m_{(\gamma)} m_{(\upsilon)}}{n} \xi_1^{(\gamma,\upsilon)} + \mathrm{O}(n^{-2}).$$

3.5 Teste de homogeneidade

Nosso interesse é encontrar evidências de que existe divergência populacional provocada por mudanças de fatores genéticos associados a mutações em *loci* de microsatélite.

No nosso caso, o teste de homogeneidade se traduziria em testar se existem diferenças entre as variações entre e dentre populações. Desta forma, gostaríamos que $\rho_{gg'l} = \rho_{ggl} = \rho_{g'g'l} = 0$, ou seja, que o desvio no número esperado de mutações provocado pela divergência populacional no *locus* l seja igual ao desvio no número de mutações no *locus* l dentre populações e que estes sejam zero e, assim,

$$E[QM_{Totl}(t)] = \theta.$$

Então, sob $H_0 : \rho_{gg'l} = \rho_{ggl} = \rho_{g'g'l} = \rho = 0$ para todo $g = 1, \ldots, G$ e $g' > g$. Desta forma, sob $H_0 : E(QM_{Totl}(t)) = \theta$, ou seja, há homogeneidade populacional.

Assim, temos que

$$\begin{aligned} H_0 &: \rho_{gg'l} = 0 \quad \text{para todo } g,\ g' = 1, \ldots, G \\ H_1 &: \rho_{gg'l} > 0 \quad \text{para algum } g \neq g'. \end{aligned}$$

A hipótese alternativa implica que existe alguma diferença em pelo menos duas populações.

Para esse propósito, define-se a estatística

$$QM_{(B-W)l}(t) = QM_{Bl}(t) - QM_{Wl}(t), \tag{3.5.1}$$

que é a distância entre a variação média entre e dentre populações.

Utilizando a teoria de estatística U podemos encontrar a distribuição assintótica de $QM_{Wl}(t)$ e de $QM_{Bl}(t)$. Na Seção 3.5.1 encontraremos a distribuição dessas duas estatísticas para amostras de tamanhos iguais (balanceadas) e na Seção 3.5.2 para tamanhos de amostras diferentes.

3.5.1 Distribuição de $QM_{Wl}(t)$ e $QM_{Bl}(t)$ para amostras balanceadas

Primeiramente, encontraremos a distribuição de $QM_{Wl}(t)$. Sabemos que, para o caso de tamanho amostral para cada população igual a n,

$$QM_{Wl}(t) = \frac{1}{G}\sum_{g=1}^{G} \bar{D}_{ggl}(t),$$

em que

$$\bar{D}_{ggl}(t) = \frac{2}{2n(2n-1)}\sum_{i=1}^{2n}\sum_{i'>i} \Delta^2_{ii'ggl}(t)$$

e $\Delta_{ii'ggl}(t) = Y_{igl}(t) - Y_{i'gl}(t)$. Além disso, $E[\Delta^2_{ii'ggl}(t)] = \theta + \rho_{ggl}$. Se $\Delta^2_{ii'ggl}(t)$ é o kernel para o parâmetro $\theta(F)$ e $Y_{1gl}(t), \ldots, Y_{2ngl}(t)$, $(g = 1, 2, \ldots, G)$ são independentes com distribuição comum F, com $E(Y_igl(t)) = \eta_{gl}$ e variância $\sigma^2 = (\theta + \rho_{ggl})/2$ pois,

$$\begin{aligned}
E[\Delta^2_{ii'ggl}(t)] &= E[Y_{igl}(t) - Y_{i'gl}(t)]^2 \\
&= E[Y_{igl}(t) - \eta_{gl} + \eta_{gl} - Y_{i'gl}(t)]^2 \\
&= E[Y_{igl}(t) - \eta_{gl}]^2 + E[\eta_{gl} - Y_{i'gl}(t)]^2 \\
&= 2\sigma^2 = \theta + \rho_{ggl} \qquad \Longrightarrow \sigma^2 = \frac{\theta + \rho_{ggl}}{2},
\end{aligned}$$

como também foi visto no Capítulo 2 (Seção 2.4). Então temos que, $\bar{D}_{ggl}(t)$ é uma estatística U de grau $m = 2$ com kernel igual à $\Delta^2_{ii'ggl}$ e com $E[\bar{D}_{ggl}(t)] = \theta + \rho_{ggl}$, ou seja, é um estimador não viesado para $\theta + \rho_{ggl}$.

Se $\theta + \rho_{ggl} < \infty$, o que é muito razoável pois, o número médio de mutações é da ordem de 10^{-2} a 10^{-5} e espera-se que τ_{ggl} seja pequeno, então $E[\Delta^4_{ii'ggl}(t)] = 3\theta^2 + 3(\theta + \rho_{ggl})^2 + \theta + \rho_{ggl} < \infty$. Aplicando a teoria de estatística U, temos que para $c = 1$

$$\begin{aligned} h_1(x) &= x^2 - 2x\eta_{gl} + \sigma^2 + \eta^2_{gl} \\ &= x^2 - 2x\eta_{gl} + \frac{\theta + \rho_{ggl}}{2} + \eta^2_{gl}, \\ \tilde{h}_1(x) &= h_1 - \theta - \rho_{ggl} \\ &= x^2 - 2x\eta_{gl} - \frac{\theta + \rho_{ggl}}{2} + \eta^2_{gl} \\ &= (x - \eta_{gl})^2 - \frac{\theta + \rho_{ggl}}{2}. \end{aligned}$$

Então,

$$\begin{aligned} \zeta^g_1 &= E\left[\tilde{h}^2_1(Y)\right] = E\left[(Y - \eta_{gl})^4 - (Y - \eta_{gl})^2(\theta + \rho_{ggl}) + \left(\frac{\theta + \rho_{ggl}}{2}\right)^2\right] \\ &= \mu_4 - \frac{(\theta + \rho_{ggl})^2}{2} + \frac{(\theta + \rho_{ggl})^2}{4} = \mu_4 - \frac{(\theta + \rho_{ggl})^2}{4}, \end{aligned}$$

em que μ_4 é o quarto momento central de $Y_{igl}(t)$. Utilizando a teoria de coalescência deste Capítulo, podemos encontrar μ_4. Sabemos que

$$E[\Delta^4_{ii'ggl}(t)] = 3\theta^2 + 3(\theta + \rho_{ggl})^2 + \theta + \rho_{ggl}.$$

Denotamos $\mu_4 = E[Y_{igl}(t) - \eta_{gl}]^4$. Então, podemos reescrever

$$\begin{aligned} E[\Delta^4_{ii'ggl}(t)] &= E[(Y_{igl}(t) - \eta_{gl}) + (\eta_{gl} - Y_{i'gl}(t))]^4 \\ &= E[(Y_{igl}(t) - \eta_{gl})^4 + 4(Y_{igl}(t) - \eta_{gl})^3(Y_{i'gl}(t) - \eta_{gl}) + \\ &+ 6(Y_{igl}(t) - \eta_{gl})^2(Y_{i'gl}(t) - \eta_{gl})^2 + 4(Y_{igl}(t) - \eta_{gl})(Y_{i'gl}(t) - \eta_{gl})^3 + \\ &+ (\eta_{gl} - Y_{i'gl}(t))^4] \\ &= 2\mu_4 + 6E[(Y_{igl}(t) - \eta_{gl})^2]E[(Y_{i'gl}(t) - \eta_{gl})^2]. \end{aligned}$$

Logo, $\mu_4 = [6\theta^2 + 3(\theta + \rho_{ggl})^2 + 2(\rho_{ggl} + \theta)]/4$. Com isso,

$$\begin{aligned}\zeta_1^g &= \frac{6\theta^2 + 3(\theta + \rho_{ggl})^2 + 2(\rho_{ggl} + \theta)}{4} - \frac{(\theta + \rho_{ggl})^2}{4} \\ &= \frac{3\theta^2 + (\theta + \rho_{ggl})^2 + \rho_{ggl} + \theta}{2} > 0, \end{aligned} \tag{3.5.2}$$

satisfazendo as duas condições do Teorema 3.3. Assim,

$$\sqrt{2n}(\bar{D}_{ggl}(t) - \theta - \rho_{ggl}) \xrightarrow{\mathcal{D}} N(0, 4\zeta_1^g).$$

Desta forma, $\bar{D}_{11l}(t), \bar{D}_{22l}(t), \ldots, \bar{D}_{GGl}(t)$ são assintoticamente $N\left(\theta, \frac{4\zeta_1^g}{2n}\right)$, com $\bar{D}_{ggl}(t)$ assintoticamente independente de $\bar{D}_{g'g'l}(t)$ $(g \neq g')$. Utilizando a teoria de estatística U, encontramos a covariância entre $(Y_{igl}(t) - Y_{i'gl}(t))^2$ e $(Y_{ig'l} - Y_{i'g'l})^2$, com $g \neq g'$. Assim,

$$\xi_c^{(g,g')} = E[\phi_c^g(Y_{igl}(t))\phi_c^{g'}(Y_{ig'l}(t))]$$

com

$$\begin{aligned}\phi_1^g(y) &= E[h_1^g - \theta - \rho_{ggl}] \\ &= E[(Y_{igl}(t) - Y_{i'gl}(t))^2 - \theta - \rho_{ggl} \mid Y_{igl} = y] \\ &= (y - \eta_{gl})^2 - \frac{\theta + \rho_{ggl}}{2}.\end{aligned}$$

Da mesma forma

$$\begin{aligned}\phi_1^{g'}(x) &= E[h_1^{g'} - \theta - \rho_{g'g'}] \\ &= (x - \eta_{g'l})^2 - \frac{\theta + \rho_{g'g'}}{2}.\end{aligned}$$

Então

$$\begin{aligned}\xi_1^{g,g'} &= E\left\{\left[(Y_{igl}(t) - \eta_{gl})^2 - \frac{\theta + \rho_{ggl}}{2}\right]\left[(Y_{ig'l}(t) - \eta_{g'l})^2 - \frac{\theta + \rho_{g'g'l}}{2}\right]\right\} \\ &= 0,\end{aligned}$$

pois assume-se que as cópias e as populações são independentes. Logo,

$$\sigma\left(\bar{D}_{ggl}(t), \bar{D}_{g'g'l}(t)\right) = 0.$$

Em situação de normalidade, covariância zero implica em independência. Assim,

$$\sqrt{2nG}\,(QM_{Wl}(t) - \theta - \bar{\rho}_{Wl}) \xrightarrow{\mathcal{D}} N\,(0, 4\zeta_1^*)\,,$$

em que

$$\zeta_1^* = \frac{3\theta^2}{2} + \frac{1}{2G}\sum_{g=1}^{G}(\theta + \rho_{ggl}) + \frac{1}{2G}\sum_{g=1}^{G}(\theta + \rho_{ggl})^2. \tag{3.5.3}$$

Considerando a estatística U generalizada podemos encontrar a distribuição assintótica de

$$QM_{Bl}(t) = \frac{2}{G(G-1)}\sum_{g<g'}\bar{D}_{gg'l}(t),$$

em que

$$\bar{D}_{gg'l}(t) = \frac{1}{(2n)^2}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}\Delta^2_{ii'gg'l}(t)$$

e $\Delta_{ii'gg'l}(t) = (Y_{igl}(t) - Y_{i'g'l}(t))$, com $E\left[\Delta^2_{ii'gg'l}(t)\right] = \theta + \rho_{gg'l} = \theta^*$. Se $\Delta^2_{ii'gg'l}(t)$ é o kernel para o parâmetro $\theta^*(F^{(g)}, F^{(g')})$, para duas amostras independentes $\{Y_{1gl}(t), \ldots, Y_{2ngl}(t)\}$ e $\{Y_{1g'l}(t), \ldots, Y_{2ng'l}(t)\}$ com $g \neq g' = 1, 2, \ldots, G$. Temos

$$\begin{aligned}\bar{D}_{gg'l}(t) &= \frac{1}{\begin{pmatrix} 2n \\ 1 \end{pmatrix}\begin{pmatrix} 2n \\ 1 \end{pmatrix}}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}\Delta^2_{ii'gg'l}(t) \\ &= \frac{1}{(2n)^2}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}\Delta^2_{ii'gg'l}(t)\end{aligned}$$

é uma estatística U de 2 amostras de grau $\mathbf{m} = (1, 1)$ com kernel igual à $\Delta^2_{ii'gg'l}(t)$ e $E\left[\bar{D}_{gg'l}(t)\right] = \theta + \rho_{gg'l} = \theta^*$, ou seja, é não viesado para θ^*. Temos $E\left[\Delta^4_{ii'gg'l}(t)\right] = 3\theta^2 + 3(\theta + \rho_{gg'l})^2 + \theta + \rho_{gg'l} < \infty$, o que é razoável considerando que o número médio de mutações é da ordem de 10^{-2} a 10^{-5}. Assim,

$$\Psi_{10}(x) = E\left[(Y_{igl}(t) - Y_{i'g'l}(t))^2 \mid Y_{igl}(t) = x\right]$$

$$
\begin{aligned}
&= x^2 - 2x\eta_{g'l} + \frac{\theta + \rho_{g'g'l}}{2} + \eta_{g'l}^2 \\
&= (x - \eta_{g'l})^2 + \frac{\theta + \rho_{g'g'l}}{2} \\
\Psi_{01}(y) &= E\left[(Y_{igl}(t) - Y_{i'g'l}(t))^2 \mid Y_{i'g'l}(t) = y\right] \\
&= y^2 - 2y\eta_{gl} + \frac{\theta + \rho_{ggl}}{2} + \eta_{gl}^2 \\
&= (y - \eta_{gl})^2 + \frac{\theta + \rho_{ggl}}{2}.
\end{aligned}
$$

Além disso,

$$
\begin{aligned}
E\left[(Y_{igl}(t) - Y_{i'g'l}(t))^2\right] &= \theta + \rho_{gg'l} \\
&= E\left[(Y_{igl}(t) - \eta_{gl})^2 + (\eta_{gl} - Y_{i'g'l}(t))^2 + \right. \\
&+ \left. 2(\eta_{gl} - Y_{i'g'l}(t))(\eta_{gl} - Y_{igl}(t))\right] \\
&= \frac{\theta + \rho_{ggl}}{2} + E\left[(\eta_{gl} - Y_{i'g'l}(t))^2\right] \\
\Longrightarrow E\left[(\eta_{gl} - Y_{i'g'l}(t))^2\right] &= \frac{\theta - \rho_{ggl}}{2} + \rho_{gg'l} \quad \text{e} \qquad (3.5.4) \\
E\left[(Y_{igl}(t) - Y_{i'g'l}(t))^4\right] &= 3\theta^2 + 3(\theta + \rho_{gg'l})^2 + \theta + \rho_{gg'l} \\
&= E\left(Y_{igl}(t) - \eta_{gl}\right)^4 + 6E\left(Y_{igl}(t) - \eta_{gl}\right)^2 E\left(Y_{i'g'l}(t) - \eta_{gl}\right)^2 + \\
&+ E\left(Y_{i'g'l}(t) - \eta_{gl}\right)^4 \\
&= \frac{6\theta^2 + 3(\theta + \rho_{ggl})^2 + 2(\theta + \rho_{ggl})}{4} + \\
&+ 3(\theta + \rho_{ggl})\left(\frac{\theta - \rho_{ggl}}{2} + \rho_{gg'l}\right) + E\left(Y_{i'g'l}(t) - \eta_{gl}\right)^4 \\
\Longrightarrow E\left[(Y_{i'g'l}(t) - \eta_{gl})^4\right] &= \frac{3\theta^2}{2} + 3(\theta + \rho_{gg'l})^2 - \frac{3(\theta + \rho_{ggl})^2}{4} + \theta + \rho_{gg'l} - \frac{(\theta + \rho_{ggl})}{2} + \\
&- 3(\theta + \rho_{ggl})\left(\frac{\theta - \rho_{ggl}}{2} + \rho_{gg'l}\right) \\
&= \frac{9\theta^2 + 2\theta}{4} + 3\theta\left[\rho_{gg'l} - \frac{\rho_{ggl}}{2}\right] + \rho_{gg'l} - \frac{\rho_{ggl}}{2} + 3\left(\rho_{gg'l} - \frac{\rho_{ggl}}{2}\right)^2 \\
&= \frac{9\theta^2 + 2\theta}{4} + (3\theta + 1)\left[\rho_{gg'l} - \frac{\rho_{ggl}}{2}\right] + 3\left(\rho_{gg'l} - \frac{\rho_{ggl}}{2}\right)^2,
\end{aligned}
$$

resultado obtido considerando que $E[(Y_{igl}(t) - \eta_{gl})] = E[(Y_{igl}(t) - \eta_{gl})]^3 = 0$, visto na Seção 2.4. Então,

$$
\zeta_{01} = E\left[\Psi_{01}^2(Y_{i'g'l}(t))\right] - (\theta^*)^2
$$

$$
\begin{aligned}
&= E\left[(Y_{i'g'l}(t)-\eta_{gl})^4+(\theta+\rho_{ggl})(Y_{i'g'l}(t)-\eta_{gl})^2+\frac{(\theta+\rho_{ggl})^2}{4}\right]-(\theta^*)^2\\
&= \frac{9\theta^2+2\theta}{4}+(3\theta+1)\left[\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right]+3\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)^2+\\
&+ (\theta+\rho_{ggl})\left(\frac{\theta-\rho_{ggl}}{2}+\rho_{gg'l}\right)+\frac{(\theta+\rho_{ggl})^2}{4}-(\theta+\rho_{gg'l})^2\\
&= 2\theta^2+\frac{\theta}{2}+2\theta\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)+\rho_{gg'l}-\frac{\rho_{ggl}}{2}+2\left(\rho^2_{gg'l}-\rho_{ggl}\rho_{gg'l}+\frac{\rho^2_{ggl}}{4}\right)\\
&= 2\theta^2+\frac{\theta}{2}+(2\theta+1)\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)+2\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)^2, \qquad (3.5.5)
\end{aligned}
$$

e analogamente,

$$
\begin{aligned}
\zeta_{10} &= E\left[\Psi^2_{10}(Y_{igl}(t))\right]-(\theta^*)^2\\
&= 2\theta^2+\frac{\theta}{2}+(2\theta+1)\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)+2\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)^2. \qquad (3.5.6)
\end{aligned}
$$

A condição para que ζ_{01} e ζ_{10} sejam maiores que zero é que

$$
\begin{aligned}
\zeta_{01} &= 2\theta^2+\frac{\theta}{2}+(2\theta+1)\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)+2\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)^2>0\\
\zeta_{10} &= 2\theta^2+\frac{\theta}{2}+(2\theta+1)\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)+2\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)^2>0
\end{aligned}
$$

desta forma,

$$
\begin{aligned}
&2\theta^2+\frac{\theta}{2}+2\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)^2>-(2\theta+1)\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right) \quad \text{e}\\
&2\theta^2+\frac{\theta}{2}+2\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)^2>-(2\theta+1)\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right).
\end{aligned}
$$

Se $\rho_{gg'l}>\rho_{ggl}/2$ e $\rho_{gg'l}>\rho_{g'g'l}/2$, a relação de desigualdade é verdadeira sempre pois, $\theta>0$. Assim, vale a extensão do Teorema 3.3 para a estatística U generalizada, e

$$
\sqrt{2n}\left(\bar{D}_{gg'l}(t)-\theta-\rho_{gg'l}\right)\xrightarrow{\mathcal{D}}N\left(0,\zeta_{10}+\zeta_{01}\right).
$$

Agora, se $\rho_{gg'l}<\rho_{ggl}/2$ ou $\rho_{gg'l}<\rho_{g'g'l}/2$, a extensão do Teorema 3.3 só é válida se

$$
\begin{aligned}
&2\theta^2+\frac{\theta}{2}+2\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)^2>(2\theta+1)\left(\frac{\rho_{ggl}}{2}-\rho_{gg'l}\right) \quad \text{e}\\
&2\theta^2+\frac{\theta}{2}+2\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)^2>(2\theta+1)\left(\frac{\rho_{g'g'l}}{2}-\rho_{gg'l}\right).
\end{aligned}
$$

Note que sob H_0 essa condição é verdadeira.

Queremos encontrar a distribuição de

$$QM_{Bl}(t) = \frac{2}{G(G-1)} \sum_{g=1}^{G} \sum_{g'>g} \bar{D}_{gg'l}(t).$$

Temos que

$$Var\left[\sum_{g=1}^{G} \sum_{g'>g} \bar{D}_{gg'l}(t)\right] = \sum_{g=1}^{G} \sum_{g'>g} \sum_{k=1}^{G} \sum_{k'>k} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kk'l}(t)\right).$$

Note que

1. quando $k = g$ e $k' = g'$, temos a variância, ou seja,

$$\sum_{g=1}^{G} \sum_{g'>g} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gg'l}(t)\right) = \sum_{g=1}^{G} \sum_{g'>g} Var\left[\bar{D}_{gg'l}(t)\right];$$

2. $\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kk'l}(t)\right) = 0$ para $g \neq g' \neq k \neq k'$, pois assume-se que as populações são independentes;

3. quando $g = k$, temos

$$\sum_{g=1}^{G} \sum_{g'>g} \sum_{k'>g'} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gk'l}(t)\right) \quad \text{ou} \quad \sum_{g=1}^{G} \sum_{k'>k} \sum_{g'>k'} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gk'l}(t)\right);$$

4. quando $g' = k'$, temos

$$\sum_{g=1}^{G} \sum_{k>g} \sum_{g'>k} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kg'l}(t)\right) \quad \text{ou} \quad \sum_{k=1}^{G} \sum_{g>k} \sum_{g'>k} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kg'l}(t)\right);$$

5. quando $g' = k$, temos

$$\sum_{g=1}^{G} \sum_{g'>g} \sum_{k'>k} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{g'k'l}(t)\right);$$

6. quando $g = k'$, corresponde ao caso em que $g' = k$.

Caso particular é para $G = 2$. Neste caso, temos que

$$\text{cov}(\bar{D}_{12l}(t), \bar{D}_{12l}(t)) = Var(\bar{D}_{12l}(t)).$$

Tomemos os exemplos para $G = 3$ e $G = 5$. Para $G = 3$, temos

$$\begin{aligned}
&\text{cov}\left(\bar{D}_{12l}(t) + \bar{D}_{13l}(t) + \bar{D}_{23l}(t), \bar{D}_{12l}(t) + \bar{D}_{13l}(t) + \bar{D}_{23l}(t)\right) = \\
&= \sum_{g=1}^{2}\sum_{g'>g} Var[\bar{D}_{gg'l}(t)] + 2\text{cov}\left(\bar{D}_{12l}(t), \bar{D}_{13l}(t)\right) + 2\text{cov}\left(\bar{D}_{12l}(t), \bar{D}_{23l}(t)\right) + \\
&+2\text{cov}\left(\bar{D}_{13l}(t), \bar{D}_{23l}(t)\right).
\end{aligned}$$

Para $G = 5$, temos

$$\begin{aligned}
&\text{cov}\left(\sum_{g=1}^{4}\sum_{g'>g}\bar{D}_{gg'l}(t), \sum_{k=1}^{5}\sum_{k'>k}\bar{D}_{kk'l}(t)\right) = \sum_{g=1}^{5}\sum_{g'>g} Var[\bar{D}_{gg'l}(t)] + \\
&+\sum_{g=1}^{3}\sum_{g'>g}\sum_{k>g'}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gkl}(t)\right) + \sum_{g=1}^{3}\sum_{k>g}\sum_{g'>k}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gkl}(t)\right) + \\
&+\sum_{g=1}^{3}\sum_{k>g}\sum_{g'>k}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kg'l}(t)\right) + \sum_{k=1}^{3}\sum_{g>k}\sum_{g'>g}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kg'l}(t)\right) + \\
&+\sum_{g=1}^{3}\sum_{k>g}\sum_{g'>k}\text{cov}\left(\bar{D}_{gkl}(t), \bar{D}_{kg'l}(t)\right) + \sum_{k=1}^{3}\sum_{g>k}\sum_{g'>g}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kgl}(t)\right) = \\
&= \sum_{g=1}^{4}\sum_{g'>g} Var[\bar{D}_{gg'l}(t)] + 2\sum_{g=1}^{3}\sum_{g'>g}\sum_{k>g'}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gkl}(t)\right) + \\
&+2\sum_{g=1}^{3}\sum_{k>g}\sum_{g'>k}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kg'l}(t)\right) + 2\sum_{g=1}^{3}\sum_{k>g}\sum_{g'>k}\text{cov}\left(\bar{D}_{gkl}(t), \bar{D}_{kg'l}(t)\right).
\end{aligned}$$

Para o caso 1, temos

$$\begin{aligned}
\sum_{g=1}^{G}\sum_{g'>g} Var[\bar{D}_{gg'l}(t)] &= \frac{1}{2n}\sum_{g=1}^{G}\sum_{g'>g}\left[4\theta^2 + \theta + 2(2\theta+1)\rho_{gg'l} - (2\theta+1)\left(\frac{\rho_{ggl} + \rho_{g'g'l}}{2}\right) + \right. \\
&+ \left. 2\left(\rho_{gg'l} - \frac{\rho_{ggl}}{2}\right)^2 + 2\left(\rho_{gg'l} - \frac{\rho_{g'g'l}}{2}\right)^2\right].
\end{aligned}$$

Por definição, temos

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

Utilizando essa definição, calculemos a covariância expressa no caso 3, ou seja

$$\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{gk'l}(t)\right) = E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right) - E\left(\bar{D}_{gg'l}(t)\right) E\left(\bar{D}_{gk'l}(t)\right).$$

Temos que $E\left(\bar{D}_{gg'l}(t)\right) = \theta + \rho_{gg'l}$ e $E\left(\bar{D}_{gk'l}(t)\right) = \theta + \rho_{gk'l}$. Falta calcular $E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right)$. Desta forma,

$$E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right) = \frac{1}{16n^4} E\left(\sum_i \sum_{i'} (Y_{igl}(t) - Y_{i'g'l}(t))^2 \sum_i \sum_{i'} (Y_{igl}(t) - Y_{i'k'l}(t))^2\right)$$

$$\sum_i \sum_{i'} (Y_{igl}(t) - Y_{i'g'l}(t))^2 = 2n \sum_i Y_{igl}^2(t) - 2 \sum_i \sum_{i'} Y_{igl}(t) Y_{i'g'l}(t) + 2n \sum_{i'} Y_{i'g'l}^2(t),$$

da mesma forma,

$$\sum_i \sum_{i'} (Y_{igl}(t) - Y_{i'k'l}(t))^2 = 2n \sum_i Y_{igl}^2(t) - 2 \sum_i \sum_{i'} Y_{igl}(t) Y_{i'k'l}(t) + 2n \sum_{i'} Y_{i'k'l}^2(t).$$

Assim, temos que

$$\begin{aligned}
&16n^4 E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right) = \\
&= 4n^2 E\left[\left(\sum_i Y_{igl}^2(t)\right)^2\right] - 4nE\left(\sum_i Y_{igl}^2(t) \sum_i \sum_{i'} Y_{igl} Y_{i'k'l}(t)\right) + \\
&+4n^2 E\left(\sum_i Y_{igl}^2(t)\right) E\left(\sum_{i'} Y_{i'k'l}^2(t)\right) - 4nE\left(\sum_i Y_{igl}^2(t) \sum_i \sum_{i'} Y_{igl} Y_{i'g'l}(t)\right) + \\
&+4E\left(\sum_i \sum_{i'} Y_{igl}(t) Y_{i'g'l}(t) \sum_i \sum_{i'} Y_{igl}(t) Y_{i'k'l}(t)\right) - 4nE\left(\sum_{i'} Y_{i'k'l}^2(t) \sum_i \sum_{i'} Y_{igl} Y_{i'g'l}(t)\right) + \\
&+4n^2 E\left(\sum_{i'} Y_{i'g'l}^2(t)\right) E\left(\sum_i Y_{igl}^2(t)\right) - 4nE\left(\sum_{i'} Y_{i'g'l}^2(t) \sum_i \sum_{i'} Y_{igl} Y_{i'k'l}(t)\right) + \\
&+4n^2 E\left(\sum_{i'} Y_{i'k'l}^2(t)\right) E\left(\sum_{i'} Y_{i'g'l}^2(t)\right),
\end{aligned}$$

com

$$E\left[\left(\sum_i Y_{igl}^2(t)\right)^2\right] = \sum_i E\left(Y_{igl}^4(t)\right) + 2 \sum_i \sum_{i'>i} E\left(Y_{igl}^2(t)\right) E\left(Y_{i'gl}^2(t)\right),$$

$$E\left(\sum_i Y_{igl}^2(t)\sum_i\sum_{i'} Y_{igl}Y_{i'k'l}(t)\right) = E\left(\sum_i Y_{igl}^2\sum_i Y_{igl}(t)\right)E\left(\sum_{i'} Y_{i'k'l}(t)\right)$$
$$= \left[\sum_i E\left(Y_{igl}^3(t)\right) + 2\sum_i\sum_{i'>i} E\left(Y_{igl}^2(t)Y_{i'gl}(t)\right)\right]\sum_{i'} E\left(Y_{i'k'l}(t)\right)$$
$$= \left[\sum_i E\left(Y_{igl}^3(t)\right) + 2\sum_i\sum_{i'>i} E\left(Y_{igl}^2(t)\right)E\left(Y_{i'gl}(t)\right)\right]\sum_{i'} E\left(Y_{i'k'l}(t)\right),$$

$$E\left(\sum_i\sum_{i'} Y_{igl}(t)Y_{i'g'l}(t)\sum_i\sum_{i'} Y_{igl}(t)Y_{i'k'l}(t)\right) =$$
$$= E\left[\left(\sum_i Y_{igl}(t)\right)^2\right]E\left(\sum_{i'} Y_{i'g'l}(t)\right)E\left(\sum_{i'} Y_{i'k'l}(t)\right),$$

$$E\left[\left(\sum_i Y_{igl}(t)\right)^2\right] = \left(\sum_i E\left(Y_{igl}^2(t)\right) + 2\sum_i\sum_{i'>i} E\left(Y_{igl}(t)\right)E\left(Y_{i'gl}(t)\right)\right),$$

$$E\left(\sum_i Y_{igl}^2(t)\right)E\left(\sum_i Y_{i'k'l}^2(t)\right) = \sum_i E\left(Y_{igl}^2(t)\right)\sum_{i'} E\left(Y_{i'k'l}^2(t)\right).$$

Além disso,

$$\begin{aligned}
E(Y_{igl}(t)) &= \eta_{gl},\\
E(Y_{igl}^2(t)) &= \frac{\theta+\rho_{ggl}}{2} + \eta_{gl}^2,\\
E(Y_{igl}^3(t)) &= \frac{3\eta_{gl}(\theta+\rho_{ggl})}{2} + \eta_{gl}^3,\\
E\left(Y_{igl}^4(t)\right) &= \frac{3\theta^2}{2} + \frac{3(\theta+\rho_{ggl})^2}{4} + \frac{(\theta+\rho_{ggl})}{2} + 3\eta_{gl}^2(\theta+\rho_{ggl}) + \eta_{gl}^4,\\
&= \frac{3\theta^2}{2} + \frac{(\theta+\rho_{ggl})}{2} + 3\left(\eta_{gl}^2 + \frac{(\theta+\rho_{ggl})}{2}\right)^2 - 2\eta_{gl}^4,
\end{aligned}$$

pois, sabemos que

$$\begin{aligned}
E\left(Y_{igl}(t) - \eta_{gl}\right)^3 &= 0\\
\Longrightarrow E\left(Y_{igl}^3(t)\right) &= 3\eta_{gl}E\left(Y_{igl}^2(t)\right) - 3\eta_{gl}^2E\left(Y_{igl}(t)\right) + \eta_{gl}^3
\end{aligned}$$

$$
\begin{aligned}
&= \frac{3\eta_{gl}(\theta+\rho_{ggl})}{2}+\eta_{gl}^3 \quad \text{e}\\
E(Y_{igl}(t)-\eta_{gl})^4 &= \frac{3\theta^2}{2}+\frac{3(\theta+\rho_{ggl})^2}{4}+\frac{(\theta+\rho_{ggl})}{2}\\
\Longrightarrow E\left(Y_{igl}^4(t)\right) &= \frac{3\theta^2}{2}+\frac{3(\theta+\rho_{ggl})^2}{4}+\frac{(\theta+\rho_{ggl})}{2}+3\eta_{gl}^2(\theta+\rho_{ggl})+\eta_{gl}^4.
\end{aligned}
$$

Logo,

$$
E\left[\left(\sum_i Y_{igl}^2(t)\right)^2\right]=2n\left(\frac{3\theta^2}{2}+\frac{\theta+\rho_{ggl}}{2}-2\eta_{gl}^4\right)+4n(n+1)\left(\eta_{gl}^2+\frac{\theta+\rho_{ggl}}{2}\right)^2,
$$

$$
E\left(\sum_i Y_{igl}^2(t)\sum_i Y_{igl}\sum_{i'} Y_{i'k'l}(t)\right)=2n\eta_{k'l}\left[6n\eta_{gl}\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)-4n\eta_{gl}^3+\right.
$$

$$
\left.+2n(2n-1)\eta_{gl}\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)\right]=8n^2\eta_{k'l}\eta_{gl}\left[(n+1)\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)-\eta_{gl}^2\right],
$$

$$
E\left(\sum_i\sum_{i'} Y_{igl}(t)Y_{i'g'l}(t)\sum_i\sum_{i'} Y_{igl}(t)Y_{i'k'l}(t)\right)=4n^2\eta_{g'l}\eta_{k'l}\left[2n\frac{\theta+\rho_{ggl}}{2}+4n^2\eta_{gl}^2\right]
$$

$$
=8n^3\eta_{g'l}\eta_{k'l}\left[\frac{\theta+\rho_{ggl}}{2}+2n\eta_{gl}^2\right],
$$

$$
E\left(\sum_i Y_{igl}^2(t)\right)E\left(\sum_i Y_{i'k'l}^2(t)\right)=4n^2\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta_{k'l}^2\right).
$$

Assim,

$$
16n^4E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right)=
$$

$$
8n^3\left(\frac{3\theta^2}{2}+\frac{\theta+\rho_{ggl}}{2}-2\eta_{gl}^4\right)+16n^3(n+1)\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)^2+32n^3\eta_{k'l}\eta_{gl}^3+
$$

$$
-32n^3\eta_{k'l}\eta_{gl}(n+1)\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)+16n^4\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)\left(\frac{\theta+\rho_{k'k'}}{2}+\eta_{k'l}^2\right)+
$$

$$
-32n^3\eta_{g'l}\eta_{gl}(n+1)\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)+32n^3\eta_{g'l}\eta_{gl}^3+32n^3\eta_{g'l}\eta_{k'l}\left(\frac{\theta+\rho_{ggl}}{2}+2n\eta_{gl}^2\right)+
$$

$$
-32n^4\eta_{gl}\eta_{g'l}\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta_{k'l}^2\right)+16n^4\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta_{g'l}^2\right)\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)+
$$

$$
-32n^4\eta_{gl}\eta_{k'l}\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta_{g'l}^2\right)+16n^4\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta_{k'l}^2\right)\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta_{g'l}^2\right).
$$

Para o caso 4, da mesma forma que para o caso 3, temos que $E[\bar{D}_{gg'l}(t)] = \theta + \rho_{gg'l}$, $E[\bar{D}_{kg'l}(t)] = \theta + \rho_{kg'l}$ e

$$
\begin{aligned}
&16n^4 E[\bar{D}_{gg'l}(t)\bar{D}_{kg'l}(t)] = \\
&= 4n^2 \sum_i E\left[Y^2_{igl}(t)\right] \sum_i E\left[Y^2_{ikl}(t)\right] - 16n^3\eta_{kl}\eta_{g'l} \sum_i E\left[Y^2_{igl}(t)\right] + \\
&+4n^2 \sum_i E\left[Y^2_{igl}(t)\right] \sum_{i'} E\left[Y^2_{i'g'l}(t)\right] - 16n^3\eta_{gl}\eta_{g'l} \sum_i E\left[Y^2_{ikl}(t)\right] + \\
&+16n^2\eta_{gl}\eta_{kl} E\left[\left(\sum_{i'} Y_{i'g'l}(t)\right)^2\right] - 8n^2\eta_{gl} E\left[\sum_{i'} Y_{i'g'l}(t) \sum_{i'} Y^2_{i'g'l}(t)\right] + \\
&+4n^2 \sum_{i'} E\left[Y^2_{i'g'l}(t)\right] \sum_i E\left[Y^2_{ikl}(t)\right] - 8n^2\eta_{kl} E\left[\sum_{i'} Y_{i'g'l}(t) \sum_{i'} Y^2_{i'g'l}(t)\right] + \\
&+4n^2 E\left[\left(\sum_{i'} Y^2_{i'g'l}(t)\right)^2\right] \\
&= 16n^4 \left(\frac{\theta + \rho_{ggl}}{2} + \eta^2_{gl}\right)\left(\frac{\theta + \rho_{kkl}}{2} + \eta^2_{kl}\right) - 32n^4\eta_{g'l}\eta_{kl}\left(\frac{\theta + \rho_{ggl}}{2} + \eta^2_{gl}\right) + \\
&+16n^4 \left(\frac{\theta + \rho_{ggl}}{2} + \eta^2_{gl}\right)\left(\frac{\theta + \rho_{g'g'l}}{2} + \eta^2_{g'l}\right) - 32n^4\eta_{gl}\eta_{g'l}\left(\frac{\theta + \rho_{kkl}}{2} + \eta^2_{kl}\right) + \\
&+32n^3\eta_{gl}\eta_{kl}\left(\frac{\theta + \rho_{g'g'l}}{2} + 2n\eta^2_{g'l}\right) - 32n^3\eta_{gl}\eta_{g'l}(n+1)\left(\frac{\theta + \rho_{g'g'l}}{2} + \eta^2_{g'l}\right) + 32n^3\eta_{gl}\eta^3_{g'l} + \\
&+16n^4 \left(\frac{\theta + \rho_{g'g'l}}{2} + \eta^2_{g'l}\right)\left(\frac{\theta + \rho_{kkl}}{2} + \eta^2_{kl}\right) - 32n^3\eta_{kl}\eta_{g'l}(n+1)\left(\frac{\theta + \rho_{g'g'l}}{2} + \eta^2_{g'l}\right) + \\
&+32n^3\eta_{kl}\eta^3_{g'l} + 8n^3\left(\frac{3\theta^2}{2} + \frac{\theta + \rho_{g'g'l}}{2} - 2\eta^4_{g'l}\right) + 16n^3(n+1)\left(\frac{\theta + \rho_{g'g'l}}{2} + \eta^2_{g'l}\right)^2.
\end{aligned}
$$

E, para o caso 5, temos $E[\bar{D}_{gg'l}(t)] = \theta + \rho_{gg'l}$, $E[\bar{D}_{g'k'l}(t)] = \theta + \rho_{g'k'l}$ e

$$
\begin{aligned}
&16n^4 E[\bar{D}_{gg'l}(t)\bar{D}_{g'k'l}(t)] = \\
&= 4n^2 \sum_i E\left[Y^2_{igl}(t)\right] \sum_i E\left[Y^2_{ig'l}(t)\right] - 8n^2\eta_{gl} E\left[\sum_{i'} Y_{i'g'l}(t) \sum_i Y^2_{ig'l}(t)\right] + \\
&+4n^2 E\left[\sum_i Y^2_{ig'l}(t) \sum_{i'} Y^2_{i'g'l}(t)\right] - 16n^3\eta_{g'l}\eta_{k'l} \sum_i E\left[Y^2_{igl}(t)\right] +
\end{aligned}
$$

$$
\begin{aligned}
&+16n^2\eta_{gl}\eta_{k'l}E\left[\sum_i Y_{ig'l}(t)\sum_{i'} Y_{i'g'l}(t)\right] - 8n^2\eta_{k'l}E\left[\sum_{i'} Y^2_{i'g'l}(t)\sum_i Y_{ig'l}(t)\right] + \\
&+4n^2\sum_{i'} E\left[Y^2_{i'k'l}(t)\right]\sum_i E\left[Y^2_{igl}(t)\right] - 16n^3\eta_{gl}\eta_{g'l}\sum_{i'} E\left[Y_{i'k'l}(t)^2\right] + \\
&+4n^2\sum_{i'} E\left[Y^2_{i'g'l}(t)\right]\sum_{i'} E\left[Y^2_{i'k'l}(t)\right] \\
&= 16n^4\left(\frac{\theta+\rho_{ggl}}{2}+\eta^2_{gl}\right)\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta^2_{g'l}\right) - 32n^3\eta_{gl}\eta_{g'l}(n+1)\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta^2_{g'l}\right) + \\
&+32n^3\eta_{gl}\eta^3_{g'l} + 8n^3\left(\frac{3\theta^2}{2}+\frac{\theta+\rho_{g'g'l}}{2}-2\eta^4_{g'l}\right) + 16n^3(n+1)\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta^2_{g'l}\right)^2 + \\
&-32n^4\eta_{g'l}\eta_{k'l}\left(\frac{\theta+\rho_{ggl}}{2}+\eta^2_{gl}\right) + 32n^3\eta_{gl}\eta_{k'l}\left(\frac{\theta+\rho_{g'g'l}}{2}+2n\eta^2_{g'l}\right) + \\
&-32n^3\eta_{k'l}\eta_{g'l}(n+1)\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta^2_{g'l}\right) + 32n^3\eta_{k'l}\eta^3_{g'l} + \\
&+16n^4\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta^2_{k'l}\right)\left(\frac{\theta+\rho_{ggl}}{2}+\eta^2_{gl}\right) - 32n^4\eta_{gl}\eta_{g'l}\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta^2_{k'l}\right) + \\
&+16n^4\left(\frac{\theta+\rho_{g'g'l}}{2}+\eta^2_{g'l}\right)\left(\frac{\theta+\rho_{k'k'l}}{2}+\eta^2_{k'l}\right).
\end{aligned}
$$

Finalmente, temos que

$$
(QM_{Bl}(t) - (\theta+\bar{\rho}_{Bl})) \xrightarrow{\mathcal{D}} N\left(0, \vartheta^2+\varrho\right),
$$

em que

$$
\begin{aligned}
\vartheta^2 &= \frac{4}{2nG^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}\left\{4\theta^2+\theta+(2\theta+1)\left[2\rho_{gg'l}-\left(\frac{\rho_{ggl}+\rho_{g'g'l}}{2}\right)\right]+\right. \\
&+ \left. 2\left(\rho_{gg'l}-\frac{\rho_{ggl}}{2}\right)^2+2\left(\rho_{gg'l}-\frac{\rho_{g'g'l}}{2}\right)^2\right\}. \qquad (3.5.7)
\end{aligned}
$$

e para $G \geq 3$,

$$
\begin{aligned}
&\varrho = \frac{8}{G^2(G-1)^2}\left\{\sum_{g=k=1}^{G}\sum_{g'>g}\sum_{k'>g'} E\left[\bar{D}_{gg'l}(t)\bar{D}_{kk'l}(t)\right] - (\theta+\rho_{gg'l})(\theta+\rho_{kk'l}) + \right. \\
&+\sum_{g=1}^{G}\sum_{k>g}\sum_{g'=k'>k} E\left[\bar{D}_{gg'l}(t)\bar{D}_{kk'l}(t)\right] - (\theta+\rho_{gg'l})(\theta+\rho_{kk'l}) +
\end{aligned}
$$

$$+\sum_{g=1}^{G}\sum_{g'=k>g}\sum_{k'>k} E\left[\bar{D}_{gg'l}(t)\bar{D}_{kk'l}(t)\right] - (\theta+\rho_{gg'l})(\theta+\rho_{kk'l})\Bigg\}.$$

Note que, em ambos os casos a variância é assintótica e não são necessariamente iguais às obtidas na Seção 3.3. Pois, nesta Seção, supomos que a seqüência de variáveis formada por $\Delta_{ii'gg'l}(t)$ para $g' > g$ com $g = 1, \ldots, G$, eram não correlacionadas.

3.5.2 Distribuição de $QM_{Wl}(t)$ e $QM_{Bl}(t)$ para amostras não balanceadas

Para tamanhos amostrais diferentes, da mesma forma que na Seção 3.5.1 vamos encontrar a distribuição de $QM_{Wl}(t)$,

$$QM_{Wl}(t) = \frac{1}{G}\sum_{g=1}^{G}\bar{D}_{ggl}(t),$$

em que

$$\bar{D}_{ggl}(t) = \frac{1}{\begin{pmatrix} 2n_g \\ 2 \end{pmatrix}}\sum_{i=1}^{2n_g}\sum_{i'>i}\Delta^2_{ii'ggl}(t).$$

Da mesma forma que no caso anterior, $\Delta^2_{ii'ggl}(t)$ é o kernel para $\theta(F)$. Assim, a estatística U é dada por

$$\bar{D}_{ggl}(t) = \frac{1}{\begin{pmatrix} 2n_g \\ 2 \end{pmatrix}}\sum_{i=1}^{2n}\sum_{i'>i}\Delta^2_{ii'ggl}(t).$$

Da mesma forma, $E\left[\Delta^4_{ii'ggl}(t)\right] = 3\theta^2+3(\theta+\rho_{ggl})^2+\theta+\rho_{ggl} < \infty$ e ζ_1^g é dado pela equação (3.5.2). Com isso,

$$\sqrt{2n_g}\left(\bar{D}_{ggl}(t) - \theta - \rho_{ggl}\right) \xrightarrow{\mathcal{D}} N\left(0, 4\zeta_1^g\right).$$

E,

$$\left(QM_{Wl}(t) - \theta - \bar{\rho}_{Wl}\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{4}{G^2}\sum_{g=1}^{G}\frac{\zeta_1^g}{2n_g}\right).$$

Assim,

$$(QMP_{Wl}(t) - \theta - \bar{\rho}_{PWl}) \xrightarrow{\mathcal{D}} N\left(0, \frac{4\sum_{g=1}^{G} 2n_g\zeta_1^g}{\left(\sum_{g=1}^{G} 2n_g\right)^2}\right).$$

Podemos utilizar a teoria de estatística U generalizada para encontrar a distribuição de $QM_{Bl}(t)$,

$$QM_{Bl}(t) = \frac{2}{G(G-1)} \sum_{g=1}^{G} \sum_{g'>g} \bar{D}_{gg'l}(t),$$

em que

$$\bar{D}_{gg'l}(t) = \frac{1}{2n_g 2n_{g'}} \sum_{i=1}^{2n_g} \sum_{i'=1}^{2n_{g'}} \Delta^2_{ii'gg'l}(t).$$

Da mesma maneira que no caso de amostras balanceadas, $\Delta^2_{ii'gg'l}(t)$ é o kernel para θ^*. Assim, $\bar{D}_{gg'l}(t)$ é uma estatística U generalizada de grau $\mathbf{m} = (1,1)$. Com isso, utilizando os resultados de Hoeffding (1948),

$$\sqrt{2n_g 2n_{g'}}\left(\bar{D}_{gg'l}(t) - \theta - \rho_{gg'l}\right) \xrightarrow{\mathcal{D}} N\left(0, 2n_{g'}\zeta_{10} + 2n_g\zeta_{01}\right),$$

em que ζ_{01} e ζ_{10} são dados conforme as equações (3.5.5) e (3.5.6). Esse resultado é válido com a mesma restrição feita para amostras balanceadas.

Desta forma,

$$(QM_{Bl}(t) - \theta - \bar{\rho}_{Bl}) \xrightarrow{\mathcal{D}} N\left(0, \lambda^2 + \varsigma\right),$$

em que

$$\lambda^2 = \frac{4}{G^2(G-1)^2} \sum_{g=1}^{G} \sum_{g'>g} \left(\frac{\zeta_{01}}{2n_{g'}} + \frac{\zeta_{10}}{2n_g}\right) \tag{3.5.8}$$

e ς é a covariância entre as estatísticas $\bar{D}_{gg'l}(t)$ $g' > g$ para $g = 1, \ldots, G$, que pode ser obtida da mesma forma que a covariância de amostras balanceadas.

3.5.3 Distribuição da estatística do teste

Considerando a estatística do teste $QM_{Bl}(t) - QM_{Wl}(t)$, devemos encontrar a distribuição dessa estatística sob H_0 para podermos executar o teste de homogeneidade proposto. A hipótese nula é definida por $\rho_{gg'l} = 0$, para todo $g,\ g' = 1, \ldots, G$, com $\rho_{gg'l} \geq 0$. Temos que $E[QM_{Bl}(t) - QM_{Wl}(t)] = \bar{\rho}_{Bl} - \bar{\rho}_{Wl}$. As distribuições de $QM_{Bl}(t)$ e $QM_{Wl}(t)$ foram encontradas para amostras balanceadas e não balanceadas. Primeiramente, considere a estatística do teste para amostras balanceadas. Sabemos que

1. A distribuição de $QM_{Wl}(t)$ é assintoticamente $N\left(\theta + \bar{\rho}_{Wl}, \frac{4\zeta_1^*}{2nG}\right)$, em que ζ_1^* é dado pela relação (3.5.3),

2. A distribuição de $QM_{Bl}(t)$ é assintoticamente $N\left(\theta + \bar{\rho}_{Bl}, \vartheta^2 + \varrho\right)$, em que ϑ^2 é dado pela relação (3.5.7), sob a restrição

$$2\theta^2 + \frac{\theta}{2} + 2\left(\rho_{gg'l} - \frac{\rho_{ggl}}{2}\right)^2 > (2\theta + 1)\left(\frac{\rho_{ggl}}{2} - \rho_{gg'l}\right),$$

   para $\rho_{gg'l} < \rho_{gg},\ g = 1, 2, \ldots, G$ com $g \neq g'$.

A variância da estatística do teste é dada por

$$\begin{aligned} Var[QM_{Bl}(t) - QM_{Wl}(t)] &= Var[QM_{Bl}(t)] + Var[QM_{Wl}(t)] - 2\text{cov}(QM_{Bl}(t), QM_{Wl}(t)) \\ &= \vartheta^2 + \varrho + \frac{4\zeta_1^*}{2nG} - 2\text{cov}(QM_{Bl}(t), QM_{Wl}(t)), \end{aligned}$$

em que

$$\begin{aligned} \text{cov}(QM_{Bl}(t), QM_{Wl}(t)) &= \text{cov}\left(\frac{1}{\binom{G}{2}} \sum_{g=1}^{G} \sum_{g'>g} \bar{D}_{gg'l}(t), \frac{1}{G} \sum_{k=1}^{G} \bar{D}_{kkl}(t)\right) \\ &= \frac{1}{G\binom{G}{2}} \sum_{g=1}^{G} \sum_{g'>g} \sum_{k=1}^{G} \text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kkl}(t)\right) \end{aligned}$$

Para encontrar a covariância entre $QM_{Bl}(t)$ e $QM_{Wl}(t)$, vamos separar nos seguintes casos

1. Se $k = g$, então

$$\begin{aligned}
&\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{ggl}(t)\right) = \\
&= \frac{1}{(2n)^2 \begin{pmatrix} 2n \\ 2 \end{pmatrix}} \text{cov}\left(\sum_{i=1}^{2n}\sum_{i'=1}^{2n}(Y_{igl}(t) - Y_{i'g'l}(t))^2, \sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t) - Y_{i'gl}(t))^2\right) \\
&= \frac{1}{(2n)^2 \begin{pmatrix} 2n \\ 2 \end{pmatrix}} \left[\text{cov}\left(2n\sum_{i=1}^{2n} Y_{igl}^2(t), \sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t) - Y_{i'gl}(t))^2\right) + \right. \\
&\left. - 4\text{cov}\left(\sum_{i=1}^{2n}\sum_{i'=1}^{2n} Y_{igl}(t)Y_{i'g'l}(t), \sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t) - Y_{i'gl}(t))^2\right)\right],
\end{aligned}$$

$$\begin{aligned}
&\text{cov}\left(2n\sum_{i=1}^{2n} Y_{igl}^2(t), \sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t) - Y_{i'gl}(t))^2\right) = \\
&\qquad = 2n\sum_{j=1}^{2n}\sum_{i=1}^{2n}\sum_{i'>i}\text{cov}\left(Y_{jgl}^2(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right) \\
&\qquad = 2n\sum_{i=1}^{2n}\sum_{i'>i}\text{cov}\left(Y_{igl}^2(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right) + \\
&\qquad +2n\sum_{i=1}^{2n}\sum_{i'>i}\text{cov}\left(Y_{i'gl}^2(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right),
\end{aligned}$$

$$\begin{aligned}
\text{cov}\left(Y_{igl}^2(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right) &= Var\left[Y_{igl}^2(t)\right] - 2\left[E[Y_{igl}^3(t)]E[Y_{i'gl}(t)] + \right. \\
&- \left. E[Y_{igl}^2(t)]E[Y_{igl}(t)]E[Y_{i'gl}(t)]\right],
\end{aligned}$$

em que

$$\begin{aligned}
Var\left[Y_{igl}^2(t)\right] &= E[Y_{igl}^4(t)] - \left(E[Y_{igl}^2(t)]\right)^2 \\
&= \frac{3\theta^2}{2} + \frac{\theta + \rho_{ggl}}{2} + 3\left(\eta_{gl}^2 + \frac{\theta + \rho_{ggl}}{2}\right)^2 - 2\eta_{gl}^4 - \left(\frac{\theta + \rho_{ggl}}{2} + \eta_{gl}^2\right)^2 \\
&= \frac{3\theta^2}{2} + \frac{\theta + \rho_{ggl}}{2} + 2\left(\eta_{gl}^2 + \frac{\theta + \rho_{ggl}}{2}\right)^2 - 2\eta_{gl}^4.
\end{aligned}$$

Logo,

$$\begin{aligned}\operatorname{cov}\left(Y_{igl}^2(t),(Y_{igl}(t)-Y_{i'gl}(t))^2\right) &= Var\left[Y_{igl}^2(t)\right]-2\left[\left(\frac{3\eta_{gl}(\theta+\rho_{ggl})}{2}+\eta_{gl}^3\right)\eta_{gl}+\right.\\ &- \left.\left(\frac{\theta+\rho_{ggl}}{2}+\eta_{gl}^2\right)\eta_{gl}^2\right]\\ &= \frac{3\theta^2+\theta+\rho_{ggl}+(\theta+\rho_{ggl})^2}{2}.\end{aligned}$$

Da mesma forma, temos

$$\operatorname{cov}\left(Y_{i'gl}^2(t),(Y_{igl}(t)-Y_{i'gl}(t))^2\right) = \frac{3\theta^2+\theta+\rho_{ggl}+(\theta+\rho_{ggl})^2}{2}.$$

Temos também,

$$\begin{aligned}&\operatorname{cov}\left(\sum_{i=1}^{2n}\sum_{i'=1}^{2n}Y_{igl}(t)Y_{i'g'l}(t),\sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t)-Y_{i'gl}(t))^2\right)=\\ &\quad=\sum_{i=1}^{2n}\sum_{i'=1}^{2n}\sum_{j=1}^{2n}\sum_{j'>j}\operatorname{cov}\left(Y_{igl}(t)Y_{i'g'l}(t),(Y_{jgl}(t)-Y_{j'gl}(t))^2\right)\\ &\quad=\sum_{j'=1}^{2n}\sum_{i=1}^{2n}\sum_{i'>i}\operatorname{cov}\left(Y_{igl}(t)Y_{j'g'l}(t),(Y_{igl}(t)-Y_{i'gl}(t))^2\right)+\\ &\quad+\sum_{j'=1}^{2n}\sum_{i=1}^{2n}\sum_{i'>i}\operatorname{cov}\left(Y_{i'gl}(t)Y_{j'g'l}(t),(Y_{igl}(t)-Y_{i'gl}(t))^2\right).\end{aligned}$$

Temos, por definição

$$\begin{aligned}&\operatorname{cov}\left(Y_{igl}(t)Y_{j'g'l}(t),(Y_{igl}(t)-Y_{i'gl}(t))^2\right)=E\left[Y_{igl}(t)Y_{j'g'l}(t)(Y_{igl}(t)-Y_{i'gl}(t))^2\right]+\\ &\quad-E[Y_{igl}(t)Y_{j'g'l}(t)]\,E\left[(Y_{igl}(t)-Y_{i'gl}(t))^2\right]\\ &\quad=\eta_{g'l}\left\{E\left[Y_{igl}(t)(Y_{igl}(t)-Y_{i'gl}(t))^2\right]-\eta_{gl}(\theta+\rho_{ggl})\right\}\\ &\quad=\eta_{g'l}\left\{E\left[Y_{igl}^3(t)-2Y_{igl}^2(t)Y_{i'gl}(t)+Y_{igl}(t)Y_{i'gl}^2(t)\right]-\eta_{gl}(\theta+\rho_{ggl})\right\}\\ &\quad=\eta_{g'l}\left(\frac{3\eta_{gl}(\theta+\rho_{ggl})}{2}+\eta_{gl}^3-2\eta_{gl}\left(\frac{\theta+\rho_{ggl}}{2}\right)-2\eta_{gl}^3+\right.\\ &\quad\left.+\eta_{gl}\left(\frac{\theta+\rho_{ggl}}{2}\right)+\eta_{gl}^3-\eta_{gl}(\theta+\rho_{ggl})\right)\end{aligned}$$

$$=\eta_{g'l}\left(\eta_{gl}\left(\frac{\theta+\rho_{ggl}}{2}\right)-\eta_{gl}^3+\eta_{gl}\left(\frac{\theta+\rho_{ggl}}{2}\right)+\eta_{gl}^3-\eta_{gl}(\theta+\rho_{ggl})\right)=0.$$

Da mesma forma,

$$\text{cov}\left(Y_{i'gl}(t)Y_{j'g'l}(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right) = 0.$$

Com isso,

$$\begin{aligned}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{ggl}(t)\right) &= \frac{1}{2n\begin{pmatrix} 2n \\ 2 \end{pmatrix}}\left[\sum_{i=1}^{2n}\sum_{i'>i}\text{cov}\left(Y_{igl}^2(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right) + \right. \\ &+ \left. \sum_{i=1}^{2n}\sum_{i'>i}\text{cov}\left(Y_{i'gl}^2(t), (Y_{igl}(t) - Y_{i'gl}(t))^2\right)\right] \\ &= \frac{3\theta^2 + \theta + \rho_{ggl} + (\theta + \rho_{ggl})^2}{2n}.\end{aligned}$$

2. Se $k = g'$, da mesma maneira que o caso anterior, temos

$$\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{g'g'l}(t)\right) = \frac{3\theta^2 + \theta + \rho_{g'g'l} + (\theta + \rho_{g'g'l})^2}{2n}.$$

3. Se $k \neq g'$ e $k \neq g$, temos $\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kkl}(t)\right) = 0$, pois assume-se que as populações são independentes.

Assim,

$$\begin{aligned}&\text{cov}(QM_{Bl}(t), QM_{Wl}(t)) = \\ &= \frac{1}{G\begin{pmatrix} G \\ 2 \end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g}\sum_{k=1}^{G}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kkl}(t)\right) \\ &= \frac{1}{G\begin{pmatrix} G \\ 2 \end{pmatrix}}\left(\sum_{g=1}^{G}\sum_{g'>g}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{ggl}(t)\right) + \sum_{g=1}^{G}\sum_{g'>g}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{g'g'l}(t)\right)\right) \\ &= \frac{1}{G\begin{pmatrix} G \\ 2 \end{pmatrix}}\left[\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{3\theta^2 + \theta + \rho_{ggl} + (\theta + \rho_{ggl})^2}{2n}\right) + \right.\end{aligned}$$

$$+\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{3\theta^2+\theta+\rho_{g'g'l}+(\theta+\rho_{g'g'l})^2}{2n}\right)\Bigg]$$

$$=\frac{6\theta^2+2\theta}{2nG}+\frac{1}{G\begin{pmatrix}G\\2\end{pmatrix}}\left[\sum_{g=1}^{G}\sum_{g'>g}\frac{\rho_{ggl}+(\theta+\rho_{ggl})^2+\rho_{g'g'l}+(\theta+\rho_{g'g'l})^2}{2n}\right].$$

Seja $\Upsilon = Var[QM_{Bl}(t) - QM_{Wl}(t)]$, desta forma

$$\Upsilon = \vartheta^2+\varrho+\frac{4\zeta_1^*}{2nG}-\frac{12\theta^2+4\theta}{2nG}-\frac{2}{G\begin{pmatrix}G\\2\end{pmatrix}}\left[\sum_{g=1}^{G}\sum_{g'>g}\frac{\rho_{ggl}+(\theta+\rho_{ggl})^2+}{2n}\right.$$

$$+\left.\frac{\rho_{g'g'l}+(\theta+\rho_{g'g'l})^2}{2n}\right].$$

Proposição 3.2. *Sejam X_1, $X_2, \ldots,$ e X variáveis aleatórias e $g : \mathbb{R} \to \mathbb{R}$ uma função contínua. Então $X_n \xrightarrow{\mathcal{D}} X \Rightarrow g(X_n) \xrightarrow{\mathcal{D}} g(X)$.*

A estatística do teste dada por (3.5.1) é combinação linear de variáveis aleatórias assintoticamente Normais, então, conforme a Proposição 3.2

$$\frac{\left(QM_{(B-W)l}(t)-(\bar{\rho}_{Bl}-\bar{\rho}_{Wl})\right)}{\sqrt{\Upsilon}} \xrightarrow{\mathcal{D}} N(0,1).$$

Sob H_0, temos que $\rho_{gg'l} = 0$, para todo g, $g' = 1, 2, \ldots, G$. Se $\rho_{gg'l} = 0$, para $g \neq g'$ implica que $\tau_{gg'l} = 0$, indicando que não houve divisão populacional, ou seja, homogeneidade entre populações. Com isso e supondo o processo de coalescência proposto neste Capítulo, as populações descendem de um mesmo *locus* ancestral. Com isso, sob H_0, podemos intuitivamente considerar $\eta_{gl} = \eta_l$. Neste Capítulo, na Seção 3.5.1, já mostramos que

$$E\left[(\eta_{gl} - Y_{i'g'l}(t))^2\right] = \frac{\theta-\rho_{ggl}}{2} + \rho_{gg'l}.$$

Desta forma,

$$E\left[(\eta_{gl} - Y_{i'g'l}(t))^2\right] = \eta_{gl}^2 - 2\eta_{gl}E[Y_{i'g'l}(t)] + E[Y_{i'g'l}(t)^2] = \frac{\theta-\rho_{ggl}}{2} + \rho_{gg'l}$$

$$= \eta_{gl}^2 - 2\eta_{gl}\eta_{g'l} + \frac{\theta + \rho_{g'g'l}}{2} + \eta_{g'l}^2 = \frac{\theta - \rho_{ggl}}{2} + \rho_{gg'l}$$

$$\Longrightarrow \rho_{gg'l} - \frac{\rho_{ggl} + \rho_{g'g'l}}{2} = (\eta_{gl} - \eta_{g'l})^2$$

e se $\rho_{gg'l} = \rho_{g'g'l} = \rho_{ggl} = 0$ para $g \neq g'$, então $\eta_{gl} = \eta_{g'l}$.

Então,

$$\begin{aligned} Var_0[QM_{Bl}(t) - QM_{Wl}(t)] &= \frac{8\theta^2}{2nG(G-1)} + \frac{2\theta}{2nG(G-1)} - \frac{8\theta^2}{2nG} - \frac{2\theta}{2nG} + \\ &+ \frac{4(G-2)}{G(G-1)}\left(\frac{\theta^2}{n} + \frac{\theta}{4n}\right) = \Upsilon_0, \end{aligned}$$

para $G \geq 3$. A demonstração deste resultado pode ser visto no Apêndice A número 10.

Note que essa variância depende do parâmetro, então podemos estimá-la através da estimação do parâmetro θ. O estimador para θ pode ser encontrado sob H_0 por $\hat{\theta} = QM_{Totl(obs)}(t)$, em que

$$QM_{Totl(obs)}(t) = \frac{1}{\begin{pmatrix} 2nG \\ 2 \end{pmatrix}} S_{Totl(obs)}(t) = \frac{1}{\begin{pmatrix} 2nG \\ 2 \end{pmatrix}} \sum_{i=1}^{2nG} \sum_{i'>i} (y_{il}(t) - y_{i'l}(t))^2,$$

em que $y_{il}(t)$ e $y_{i'l}(t)$ são os valores observados do tamanho alélico da i e i'-ésima cópia, desconsiderando a população.

Desta forma, sob H_0,

$$\frac{QM_{Bl}(t) - QM_{Wl}(t)}{\sqrt{\Upsilon_0}} \xrightarrow{\mathcal{D}} N(0,1).$$

Agora, considere a estatística para amostras não balanceadas. Definimos a estatística $QM_{(B-W)l}(t) = QM_{Bl}(t) - QM_{Wl}(t)$. Sabemos que

1. A distribuição de $QM_{Wl}(t)$ é assintoticamente $N\left(\theta + \bar{\rho}_{Wl}, \frac{4}{G^2}\sum_{g=1}^{G}\frac{\zeta_1^g}{2n_g}\right)$, em que ζ_1^g é dado pela relação (3.5.2),

2. A distribuição de $QM_{Bl}(t)$ é assintoticamente $N\left(\theta + \bar{\rho}_{Bl}, \lambda^2 + \varsigma\right)$, em que λ^2 é dado pela relação (3.5.8) e ς é a covariância.

A variância da estatística do teste é dada por

$$Var[QM_{Bl}(t) - QM_{Wl}(t)] = \lambda^2 + \varsigma + \frac{4}{G^2}\sum_{g=1}^{G}\frac{\zeta_1^g}{2n_g} - 2\text{cov}\left(QM_{Bl}(t), QM_{Wl}(t)\right),$$

em que

$$\text{cov}(QM_{Bl}(t), QM_{Wl}(t)) = \frac{1}{G\begin{pmatrix} G \\ 2 \end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g}\sum_{k=1}^{G}\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kkl}(t)\right).$$

Para esse caso, da mesma forma para amostras balanceadas, para encontrar a covariância entre $QM_{Bl}(t)$ e $QM_{Wl}(t)$, vamos separar nos seguintes casos

1. Se $k = g$, então

$$\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{ggl}(t)\right) = \frac{3\theta^2 + \theta + \rho_{ggl} + (\theta + \rho_{ggl})^2}{2n_g},$$

cujo o cálculo segue o mesmo raciocínio para amostras balanceadas.

2. Se $k = g'$ temos

$$\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{g'g'l}(t)\right) = \frac{3\theta^2 + \theta + \rho_{g'g'l} + (\theta + \rho_{g'g'l})^2}{2n_{g'}}.$$

3. Se $k \neq g'$ e $k \neq g$, temos

$$\text{cov}\left(\bar{D}_{gg'l}(t), \bar{D}_{kkl}(t)\right) = 0,$$

pois assume-se que as populações são independentes.

Assim,

$$\begin{aligned}\text{cov}(QM_{Bl}(t), QM_{Wl}(t)) &= \frac{1}{G\begin{pmatrix} G \\ 2 \end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{3\theta^2 + \theta + \rho_{ggl} + (\theta + \rho_{ggl})^2}{2n_g} + \right. \\ &+ \left. \frac{3\theta^2 + \theta + \rho_{g'g'l} + (\theta + \rho_{g'g'l})^2}{2n_{g'}}\right).\end{aligned}$$

Chamaremos $Var[QM_{Bl}(t) - QM_{Wl}(t)]$ de Υ^*. Desta forma,

$$\Upsilon^* = \lambda^2 + \varsigma + \frac{4}{G^2}\sum_{g=1}^{G}\frac{\zeta_1^g}{2n_g} - 2\left\{\frac{1}{G\binom{G}{2}}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{3\theta^2+\theta+\rho_{ggl}+(\theta+\rho_{ggl})^2}{2n_g} + \right.\right.$$
$$\left.\left. + \frac{3\theta^2+\theta+\rho_{g'g'l}+(\theta+\rho_{g'g'l})^2}{2n_{g'}}\right)\right\}.$$

Temos que a estatística do teste é combinação linear de variáveis aleatórias Normais, com isso,

$$\frac{\left(QM_{(B-W)l}(t) - (\bar{\rho}_{Bl} - \bar{\rho}_{Wl})\right)}{\sqrt{\Upsilon^*}} \xrightarrow{\mathcal{D}} N(0,1). \tag{3.5.9}$$

Sob H_0, temos que

$$\lambda_0^2 = \frac{(8\theta^2+2\theta)}{G^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{1}{2n_g}+\frac{1}{2n_{g'}}\right), \tag{3.5.10}$$

$$\varsigma_0 = \frac{8}{G^2(G-1)^2}\left\{\sum_{g=1}^{G}\sum_{g'>g}\sum_{k'>g'}\left(\frac{\theta^2}{n_g}+\frac{\theta}{4n_g}\right) + \sum_{g=1}^{G}\sum_{k>g}\sum_{g'>k}\left(\frac{\theta^2}{n_{g'}}+\frac{\theta}{4n_{g'}}\right) + \right.$$
$$\left. + \sum_{g=1}^{G}\sum_{g'>g}\sum_{k'>g'}\left(\frac{\theta^2}{n_{g'}}+\frac{\theta}{4n_{g'}}\right)\right\}, \tag{3.5.11}$$

$$\zeta_1^g = 2\theta^2 + \frac{\theta}{2},$$

cuja demonstração se encontra no Apêndice A número 11.

Agora, considere a estatística do teste $QM_{(B-PW)l}(t) = QM_{Bl}(t) - QMP_{Wl}(t)$. Sabemos que

$$(QMP_{Wl}(t) - \theta - \bar{\rho}_{Wl}) \xrightarrow{\mathcal{D}} N\left(0, 4\sum_{g=1}^{G} w_g^2\zeta_1^g\right),$$

em que ζ_1^g é dado pela relação (3.5.2). Temos que, sob H_0

$$\mathrm{cov}_0(QM_{Bl}(t), QMP_{Wl}(t)) = \frac{4\theta^2+\theta}{\binom{G}{2}}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{w_g}{2n_g}+\frac{w_g'}{2n_{g'}}\right)$$

$$
\begin{aligned}
&= \frac{4\theta^2+\theta}{\binom{G}{2}} \sum_{g=1}^{G} \sum_{g'>g} \left(\frac{2}{\sum_{g=1}^{G} 2n_g} \right) \\
&= \frac{8\theta^2+2\theta}{\sum_{g=1}^{G} 2n_g}.
\end{aligned}
$$

Além disso, temos que sob H_0,

$$
(QMP_{Wl}(t) - \theta) \xrightarrow{\mathcal{D}} N\left(0, \frac{4\zeta_1}{\sum_{g=1}^{G} 2n_g}\right),
$$

em que $\zeta_1 = 2\theta^2 + \dfrac{\theta}{2}$.

Chamaremos $Var[QM_{Bl}(t) - QMP_{Wl}(t)]$ de ψ. Com isso

$$
\psi_0 = \lambda_0^2 + \varsigma_0 - \frac{8\theta^2+2\theta}{\sum_{g=1}^{G} 2n_g}, \tag{3.5.12}
$$

em que λ_0^2 e ς_0 são dados pelas relações (3.5.10) e (3.5.11), respectivamente.

Da mesma forma que (3.5.9), temos sob H_0

$$
\frac{QM_{(B-PW)l}(t)}{\sqrt{\psi_0}} \xrightarrow{\mathcal{D}} N(0,1).
$$

O p-valor para o teste de homogeneidade é obtido da seguinte maneira

$$
\text{p-valor} = P\left(\frac{QM_{(B-PW)l}(t)}{\sqrt{\psi_0}} > \frac{\bar{\rho}_{Bl(obs)} - \bar{\rho}_{Wl(obs)}}{\sqrt{\psi_{(obs)l}}} \right) = 1 - \Phi\left(\frac{\bar{\rho}_{Bl(obs)} - \bar{\rho}_{Wl(obs)}}{\psi_{(obs)l}} \right),
$$

em que, sob H_0

$$
\frac{QM_{(B-PW)l}(t)}{\sqrt{\psi_0}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{e} \quad \Phi\left(\frac{\bar{\rho}_{(obs)l}}{\psi_{(obs)l}} \right) = P\left(\frac{QM_{(B-PW)l}(t)}{\sqrt{\psi_0}} \leq \frac{\bar{\rho}_{(obs)l}}{\sqrt{\psi_{(obs)l}}} \right),
$$

com $\psi_{(obs)l}$ dado pela relação (3.5.12), ou seja, considerando H_0 verdadeira.

Com isso, definimos a distribuição da estatística do teste para amostras balanceadas e não balanceadas.

Capítulo 4

Distância genética para *loci* de microsatélites baseada nos desvios absolutos

4.1 Introdução

Shriver et al. (1995) consideraram o modelo de mutação "*stepwise*" para desenvolver uma medida de distância genética que fosse linear com respeito ao tempo.

Duas medidas, que são muito utilizadas em estatística genética, serão introduzidas neste Capítulo, a distância genética mínima (D_M) e a distância genética padrão (D_S), introduzidas por Nei (1972). Segundo Li (1976), essas duas medidas não são lineares com relação ao tempo de divergência entre duas populações, quando a taxa de mutação é alta. Sendo assim, Shriver et al. (1995) propuzeram uma nova medida de distância genética, D_{SM}, que é mais apropriada para *loci* com altas taxas de mutação, envolvido com o processo de mutação "*stepwise*".

Na Seção 4.2, introduziremos as medidas de distância de Shriver et al. (1995) e de Nei (1972). A medida de distância desenvolvida por Shriver et al. (1995) será chamada de medida de distância de Shriver. Adaptamos essa medida de distância para construir uma

medida à qual podemos utilizar, de forma análoga, a teoria desenvolvida no Capítulo 3 (na Seção 3.2).

Na Seção 4.3.1, vamos encontrar a distribuição das medidas estudadas em 4.2, utilizando a teoria de Estatística U.

4.2 Medidas de heterozigosidade e de distância

Shriver et al. (1993) consideram a medida de heterozigosidade numa única população por $1 - \sum_i p_i^2$, que é a medida de diversidade gênica, definida no Capítulo 1 (na Seção 1.2).

Seja p_{gi} e $p_{g'i'}$ as proporções dos alelos i e i' num determinado *locus* nas populações g e g', respectivamente. Seja,

$$\begin{aligned}
\omega_g &= \sum_{i=1}^{2n} \sum_{i' \neq i} p_{gi} p_{gi'} \\
\omega_{g'} &= \sum_{i=1}^{2n} \sum_{i' \neq i} p_{g'i} p_{g'i'} \\
\omega_{gg'} &= \sum_{i=1}^{2n} \sum_{i' \neq i} p_{gi} p_{g'i'},
\end{aligned}$$

em que a soma é sobre todas as combinações de diferentes alelos $i \neq i'$ num mesmo *locus*. Note que as quantidades ω_g, $\omega_{g'}$ e $\omega_{gg'}$ podem ser interpretadas como as probabilidades de que dois alelos sejam diferentes quando amostrados aleatoriamente de duas populações diferentes, $\omega_{gg'}$, ou de uma mesma população, ω_g e $\omega_{g'}$.

A distância genética mínima de Nei (1972), D_M, e a distância padrão de Nei (1972), D_S, são definidas por,

$$\begin{aligned}
D_M &= \omega_{gg'} - \frac{\omega_g + \omega_{g'}}{2} \\
D_S &= -\log(1 - \omega_{gg'}) + \frac{\log(1 - \omega_g) + \log(1 - \omega_{g'})}{2}.
\end{aligned}$$

A medida de distância utilizada por Shriver, D_{Sw}, é uma extensão de D_M, que foi primeiramente sugerida por Rao (1982c). Considerando um *locus* envolvido com o pro-

cesso de mutação "*stepwise*", descrito no Capítulo 2, ω_g, $\omega_{g'}$ e $\omega_{gg'}$ podem ser ponderados pelo valor absoluto da diferença do número de repetições entre dois alelos. Ou seja,

$$\begin{aligned}
\omega_{gw} &= \sum_{i=1}^{2n}\sum_{i'\neq i}\delta^{g}_{ii'}p_{gi}p_{gi'} \quad \text{medida dentro da população } g \\
\omega_{g'w} &= \sum_{i=1}^{2n}\sum_{i'\neq i}\delta^{g'}_{ii'}p_{g'i}p_{g'i'} \quad \text{medida dentro da população } g' \\
\omega_{gg'w} &= \sum_{i=1}^{2n}\sum_{i'\neq i}\delta^{gg'}_{ii'}p_{gi}p_{g'i'} \quad \text{medida entre as populações } g \text{ e } g',
\end{aligned}$$

em que $\delta^{g}_{ii'} = |Y_{ig} - Y_{i'g}|$, $\delta^{g'}_{ii'} = |Y_{ig'} - Y_{i'g'}|$ e $\delta^{gg'}_{ii'} = |Y_{ig} - Y_{i'g'}|$ com Y_{ig} representado o número de repetições do alelo i na população g.

Assim, Shriver et al. (1995) sugerem D_{Sw} como uma medida de distância geral entre populações

$$D_{Sw} = \omega_{gg'w} - \frac{\omega_{gw} + \omega_{g'w}}{2}.$$

Note que se $D_{Sw} > 0$, a variabilidade entre as populações g e g' é maior do que a média entre as variabilidades dentro de g e g'.

Vamos utilizar a distância genética entre o número de repetições entre dois alelos, e aplicar a teoria introduzida na Seção 3.2 do Capítulo 3. Para isso, considere a mesma medida $\Delta_{ii'ggl}(t) = Y_{igl}(t) - Y_{i'gl}(t)$, em que $Y_{igl}(t)$, o número de repetições da i-ésima cópia $(i = 1, 2, \ldots, 2n)$ na g-ésima população $(g = 1, 2, \ldots, G)$ no l-ésimo *locus* $(l = 1, 2, \ldots, L)$ no tempo t. Podemos definir a soma dos desvios absolutos na g-ésima população para o l-ésimo *locus* por

$$\mathrm{D}_{ggl}(t) = \sum_{i=1}^{2n}\sum_{i'>i}|\Delta_{ii'ggl}(t)|$$

e a distância média absoluta na população g é dada por

$$\bar{\mathrm{D}}_{ggl}(t) = \frac{\mathrm{D}_{ggl}(t)}{\begin{pmatrix} 2n \\ 2 \end{pmatrix}}.$$

A distância média absoluta dentre populações é dada por

$$AM_{Wl}(t) = \frac{1}{G}\sum_{g=1}^{G} \bar{\mathrm{D}}_{ggl}(t), \tag{4.2.1}$$

que é equivalente a $QM_{Wl}(t)$, descrita no Capítulo 3.

Podemos também definir uma medida de distância entre populações. Para isso, considere a soma dos desvios absolutos entre o número de repetições da população g e g',

$$\mathrm{D}_{gg'l}(t) = \sum_{i=1}^{2n}\sum_{i'=1}^{2n} |\Delta_{ii'gg'l}(t)|,$$

em que $\Delta_{ii'gg'l}(t) = Y_{igl}(t) - Y_{i'g'l}(t)$. A distância média absoluta entre a população g e g' no l-ésimo *locus* no tempo t é dada por

$$\bar{\mathrm{D}}_{gg'l}(t) = \frac{\mathrm{D}_{gg'l}(t)}{(2n)^2}.$$

A distância média absoluta entre populações é dada por

$$AM_{Bl}(t) = \frac{1}{\begin{pmatrix} G \\ 2 \end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g} \bar{\mathrm{D}}_{gg'l}(t),$$

que é equivalente a $QM_{Bl}(t)$, descrita no Capítulo 3.

Com isso a soma absoluta total é dada por

$$AS_{Totl}(t) = \sum_{g=1}^{G}\sum_{g'>g}(2n)^2\ \bar{\mathrm{D}}_{gg'l}(t) + \sum_{g=1}^{G}\begin{pmatrix} 2n \\ 2 \end{pmatrix}\bar{\mathrm{D}}_{ggl}(t)$$

e a sua média é

$$\begin{aligned} AM_{Totl}(t) &= \frac{AS_{Totl}(t)}{\begin{pmatrix} 2nG \\ 2 \end{pmatrix}} \\ &= \frac{2n(G-1)}{2nG-1}AM_{Bl}(t) + \frac{2n-1}{2nG-1}AM_{Wl}(t). \end{aligned}$$

Para encontrarmos o primeiro momento de $|\Delta_{ii'ggl}(t)|$ ou de $|\Delta_{ii'gg'l}(t)|$ devemos utilizar a teoria da Seção 3.2. Primeiramente, vamos encontrar o primeiro momento de $|\Delta_{ii'l}(t)| = |Y_{il}(t) - Y_{i'l}(t)|$. Considerando o Teorema de Révész (Teorema 3.1), sabemos que a distribuição de $\Delta_{ii'l}(t) \mid [T = t, N(t) = \alpha]$, é um passeio aleatório simétrico em 0 com α passos. Assim, considere a variável aleatória X_α representando o passeio aleatório simétrico em zero com α passos. Podemos encontrar a esperança de $|X_\alpha|$, ou seja,

$$E|X_\alpha| = \frac{1}{2^\alpha} \sum_{k=-\alpha}^{\alpha} |k| \begin{pmatrix} \alpha \\ \frac{\alpha-k}{2} \end{pmatrix},$$

lembrando que se α é par, k assume somente valores pares e, se α é ímpar, k assume somente valores ímpares. Assim, para α qualquer $k \in \{-n, -n+2, -n+4, \ldots, n-4, n-2, n\}$. Essa soma pode ser reescrita da seguinte maneira,

$$\frac{1}{2^\alpha} \sum_{k=-\alpha}^{\alpha} |k| \begin{pmatrix} \alpha \\ \frac{\alpha-k}{2} \end{pmatrix} = \frac{1}{2^\alpha} \sum_{u=0}^{\alpha} |\alpha - 2u| \begin{pmatrix} \alpha \\ u \end{pmatrix} \quad \text{com} \quad u = \frac{\alpha - k}{2}.$$

Por exemplo, consideremos $\alpha = 3$, então

$$\begin{aligned} E|X_3| &= \frac{1}{2^3} \sum_{k=-3}^{3} |k| \begin{pmatrix} 3 \\ \frac{3-k}{2} \end{pmatrix} \\ &= \frac{1}{8} \left[3 \begin{pmatrix} 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} + 3 \begin{pmatrix} 3 \\ 0 \end{pmatrix} \right] = \frac{3}{2} \end{aligned}$$

e para $\alpha = 4$

$$\begin{aligned} E|X_4| &= \frac{1}{2^4} \sum_{k=-4}^{4} |k| \begin{pmatrix} 4 \\ \frac{4-k}{2} \end{pmatrix} \\ &= \frac{1}{16} \left[4 \begin{pmatrix} 4 \\ 4 \end{pmatrix} + 2 \begin{pmatrix} 4 \\ 3 \end{pmatrix} + 2 \begin{pmatrix} 4 \\ 1 \end{pmatrix} + 4 \begin{pmatrix} 4 \\ 0 \end{pmatrix} \right] = \frac{3}{2}. \end{aligned}$$

Para $\alpha = m$ qualquer, temos,

$$E|X_m| = 2^{-m} \sum_{k=0}^{m} |m - 2k| \begin{pmatrix} m \\ k \end{pmatrix}$$

$$= m \underbrace{2^{-m} \sum_{k=0}^{m} \frac{|m-2k|}{m} \begin{pmatrix} m \\ k \end{pmatrix}}_{c_m}.$$

Mostraremos que $c_m < 1$, para $m > 1$. Para isso, vamos considerar um resultado de relação de ordem dos números reais. Esse resultado consiste no seguinte fato:

- Seja a seqüência de números reais, tal que, $a_1 \leq b_1, a_2 \leq b_2, \ldots, a_m \leq b_m$.
- Se $\exists\ i$ tal que $a_i < b_i$ para $i = 1, 2, \ldots, m$, então $a_1 + a_2 + \ldots + a_m < b_1 + b_2 + \ldots + b_m$.

Considere

$$a_m = \frac{|m-2k|}{m} \quad \text{e} \quad b_m = 1 \quad \text{para todo } m = 1, 2, \ldots,.$$

Para $m = 1$, temos que

$$\begin{aligned} |1-2k| &= 1, \quad \text{se} \quad k = 0 \\ |1-2k| &= 1, \quad \text{se} \quad k = 1, \end{aligned}$$

não valendo a condição do resultado.

Para $m > 1$, temos,

$$\begin{aligned} \frac{|m-2k|}{m} &= 1, \quad \text{se} \quad k = 0 \\ \frac{|m-2k|}{m} &< 1, \quad \text{para} \quad 0 < k < m \\ \frac{|m-2k|}{m} &= 1, \quad \text{se} \quad k = m, \end{aligned}$$

satifazendo a condição do resultado.

Então, para $m > 1$

$$c_m = 2^{-m} \sum_{k=0}^{m} \frac{|m-2k|}{m} \begin{pmatrix} m \\ k \end{pmatrix} < 2^{-m} \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} = 1,$$

em que a última igualdade se deve ao binômio de Newton. Assim, $c_m < 1$ para todo $m > 1$. Para encontrar c_m, devemos encontrar o resultado de

$$2^{-m} \sum_{k=0}^{m} \frac{|m-2k|}{m} \begin{pmatrix} m \\ k \end{pmatrix}.$$

Considere a seguinte identidade

$$\begin{pmatrix} m \\ k \end{pmatrix} = \frac{m!}{(m-k)!k!} = \begin{pmatrix} m \\ m-k \end{pmatrix}. \tag{4.2.2}$$

Separaremos os casos de m par e m ímpar.

Para m par temos,

$$\begin{aligned}
c_m &= 2^{-m} \left[\sum_{k=0}^{\frac{m}{2}} \frac{m-2k}{m} \begin{pmatrix} m \\ k \end{pmatrix} - \sum_{k=\frac{m}{2}+1}^{m} \frac{m-2k}{m} \begin{pmatrix} m \\ k \end{pmatrix} \right] \\
&= 2^{-m} \left[\sum_{k=0}^{\frac{m}{2}} \begin{pmatrix} m \\ k \end{pmatrix} - \frac{2}{m} \sum_{k=1}^{\frac{m}{2}} k \begin{pmatrix} m \\ k \end{pmatrix} - \sum_{k=\frac{m}{2}+1}^{m} \begin{pmatrix} m \\ k \end{pmatrix} + \frac{2}{m} \sum_{k=\frac{m}{2}+1}^{m} k \begin{pmatrix} m \\ k \end{pmatrix} \right] \\
&= 2^{-m} \left[\sum_{k=0}^{\frac{m}{2}-1} \begin{pmatrix} m \\ k \end{pmatrix} + \begin{pmatrix} m \\ m/2 \end{pmatrix} - 2 \sum_{k=1}^{\frac{m}{2}} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix} - \sum_{k=\frac{m}{2}+1}^{m} \begin{pmatrix} m \\ k \end{pmatrix} + \right. \\
&+ \left. 2 \sum_{k=\frac{m}{2}+1}^{m} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix} \right] \\
&= 2^{-m} \left[\sum_{k=0}^{\frac{m}{2}-1} \begin{pmatrix} m \\ k \end{pmatrix} + \begin{pmatrix} m \\ m/2 \end{pmatrix} - 2 \sum_{l=0}^{\frac{m}{2}-1} \begin{pmatrix} m-1 \\ l \end{pmatrix} - \sum_{k=\frac{m}{2}+1}^{m} \begin{pmatrix} m \\ k \end{pmatrix} + \right. \\
&+ \left. 2 \sum_{l=\frac{m}{2}}^{m-1} \begin{pmatrix} m-1 \\ l \end{pmatrix} \right].
\end{aligned}$$

Utilizando o resultado (4.2.2), temos que

$$\begin{aligned}
\sum_{k=0}^{\frac{m}{2}-1} \begin{pmatrix} m \\ k \end{pmatrix} &= \sum_{k=\frac{m}{2}+1}^{m} \begin{pmatrix} m \\ k \end{pmatrix} \\
2 \sum_{l=0}^{\frac{m}{2}-1} \begin{pmatrix} m-1 \\ l \end{pmatrix} &= 2 \sum_{k=\frac{m}{2}}^{m-1} \begin{pmatrix} m-1 \\ l \end{pmatrix}.
\end{aligned}$$

Logo, para m par $c_m = 2^{-m}\begin{pmatrix} m \\ \frac{m}{2} \end{pmatrix}$. Para m ímpar, temos

$$\begin{aligned}
c_m &= 2^{-m}\left[\sum_{k=0}^{\frac{m-1}{2}} \frac{m-2k}{m}\begin{pmatrix} m \\ k \end{pmatrix} - \sum_{k=\frac{m+1}{2}}^{m} \frac{m-2k}{m}\begin{pmatrix} m \\ k \end{pmatrix}\right] \\
&= 2^{-m}\left[\sum_{k=0}^{\frac{m-1}{2}} \begin{pmatrix} m \\ k \end{pmatrix} - \frac{2}{m}\sum_{k=1}^{\frac{m-1}{2}} k\begin{pmatrix} m \\ k \end{pmatrix} - \sum_{k=\frac{m+1}{2}}^{m} \begin{pmatrix} m \\ k \end{pmatrix} + \frac{2}{m}\sum_{k=\frac{m+1}{2}}^{m} k\begin{pmatrix} m \\ k \end{pmatrix}\right] \\
&= 2^{-m}\left[\sum_{k=0}^{\frac{m-1}{2}} \begin{pmatrix} m \\ k \end{pmatrix} - 2\sum_{k=1}^{\frac{m-1}{2}} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix} - \sum_{k=\frac{m+1}{2}}^{m} \begin{pmatrix} m \\ k \end{pmatrix} + 2\sum_{k=\frac{m+1}{2}}^{m} \begin{pmatrix} m-1 \\ k-1 \end{pmatrix}\right] \\
&= 2^{-m}\left[\sum_{k=0}^{\frac{m-1}{2}} \begin{pmatrix} m \\ k \end{pmatrix} - 2\sum_{l=0}^{\frac{m-1}{2}-1} \begin{pmatrix} m-1 \\ l \end{pmatrix} - \sum_{k=\frac{m+1}{2}}^{m} \begin{pmatrix} m \\ k \end{pmatrix} + 2\sum_{l=\frac{m+1}{2}}^{m-1} \begin{pmatrix} m-1 \\ l \end{pmatrix} + \right. \\
&+ \left. 2\begin{pmatrix} m-1 \\ \frac{m+1}{2}-1 \end{pmatrix}\right].
\end{aligned}$$

Utilizando o resultado (4.2.2) temos que para m ímpar $c_m = 2^{-m-1}\begin{pmatrix} m-1 \\ \frac{m+1}{2}-1 \end{pmatrix}$. Assim,

$$E|X_m| = \begin{cases} \frac{m}{2^m}\begin{pmatrix} m \\ \frac{m}{2} \end{pmatrix}, & \text{se } m \text{ é par;} \\ \\ \frac{m}{2^{m-1}}\begin{pmatrix} m-1 \\ \frac{m+1}{2}-1 \end{pmatrix}, & \text{se } m \text{ é ímpar.} \end{cases}$$

Com isso,

$$E[|\Delta_{ii'l}(t)| \mid N(t) = \alpha, T = t] = \begin{cases} \frac{\alpha}{2^\alpha}\begin{pmatrix} \alpha \\ \frac{\alpha}{2} \end{pmatrix}, & \text{se } \alpha \text{ é par;} \\ \\ \frac{\alpha}{2^{\alpha-1}}\begin{pmatrix} \alpha-1 \\ \frac{\alpha+1}{2}-1 \end{pmatrix}, & \text{se } \alpha \text{ é ímpar.} \end{cases}$$

Como $N(t) \mid T = t \sim Poisson(2\mu t)$. Então,

$$E[E[|\Delta_{ii'l}(t)| \mid N(t), T = t] \mid T = t] =$$

$$= \sum_{k=0}^{\infty} E\left[|\Delta_{ii'l}(t)| \mid N(t) = k, T = t\right] P[N(t) = k \mid T = t]$$

$$= \sum_{k \in \{0,2,\ldots\}} \frac{k}{2^k} \begin{pmatrix} k \\ \frac{k}{2} \end{pmatrix} P[N(t) = k \mid T = t] +$$

$$+ \sum_{k \in \{1,3,\ldots\}} \frac{k}{2^{k-1}} \begin{pmatrix} k-1 \\ \frac{k+1}{2} - 1 \end{pmatrix} P[N(t) = k \mid T = t].$$

Assim,

$$E[E[|\Delta_{ii'l}(t)| \mid N(t), T = t] \mid T = t] = E\left[\frac{N(t)}{2^{N(t)}} \begin{pmatrix} N(t) \\ \frac{N(t)}{2} \end{pmatrix},\ N(t) \text{ par} \mid T = t\right] +$$

$$+ E\left[\frac{N(t)}{2^{N(t)-1}} \begin{pmatrix} N(t)-1 \\ \frac{N(t)+1}{2} - 1 \end{pmatrix},\ N(t) \text{ ímpar} \mid T = t\right].$$

Calculando as esperanças, temos

$$E\left[\frac{N(t)}{2^{N(t)}} \begin{pmatrix} N(t) \\ \frac{N(t)}{2} \end{pmatrix},\ N(t) \text{ par} \mid T = t\right] = \sum_{k \in \{0,2,4,\ldots\}} \frac{k}{2^k} \begin{pmatrix} k \\ \frac{k}{2} \end{pmatrix} \frac{(2\mu t)^k \exp\{-2\mu t\}}{k!}$$

$$= \sum_{k \in \{2,4,\ldots\}} \frac{k}{2^k} \frac{(2\mu t)^k \exp\{-2\mu t\}}{\frac{k}{2}!\frac{k}{2}!}$$

$$= 2 \sum_{k \in \{2,4,\ldots\}} \left(\frac{2\mu t}{2}\right)^k \frac{\exp\{-2\mu t\}}{\left(\frac{k}{2}-1\right)!\frac{k}{2}!}$$

$$= 2\mu t \sum_{k \in \{2,4,\ldots\}} \left(\frac{2\mu t}{2}\right)^{\frac{k}{2}} \left(\frac{2\mu t}{2}\right)^{\frac{k}{2}-1} \frac{\exp\{-2\mu t\}}{\left(\frac{k}{2}-1\right)!\frac{k}{2}!}$$

$$\left(\text{fazendo } l = \frac{k}{2} - 1\right) = 2\mu t \sum_{l=0}^{\infty} \left(\frac{2\mu t}{2}\right)^l \left(\frac{2\mu t}{2}\right)^{l+1} \frac{\exp\{-2\mu t\}}{l!(l+1)!}$$

$$= \frac{(2\mu t)^2}{2} \exp\{-2\mu t\} \sum_{l=0}^{\infty} \frac{\left(\frac{2\mu t}{2}\right)^{2l}}{l!(l+1)!}$$

$$E\left[\frac{N(t)}{2^{N(t)-1}} \begin{pmatrix} N(t)-1 \\ \frac{N(t)+1}{2} - 1 \end{pmatrix},\ N(t) \text{ ímpar} \mid T = t\right] =$$

$$
\begin{aligned}
&= \sum_{k\in\{1,3,\ldots\}} \frac{k}{2^{k-1}} \left(\begin{array}{c} k-1 \\ \frac{k+1}{2}-1 \end{array} \right) \frac{(2\mu t)^k \exp\{-2\mu t\}}{k!} \\
&= 2\mu t \sum_{k\in\{1,3,\ldots\}} \left(\frac{2\mu t}{2}\right)^{k-1} \frac{\exp\{-2\mu t\}}{\left(\frac{k-1}{2}\right)!\left(\frac{k-1}{2}\right)!} \\
\left(\text{fazendo } l = \frac{k-1}{2}\right) &= 2\mu t \sum_{l=0}^{\infty} \left(\frac{2\mu t}{2}\right)^{2l} \frac{\exp\{-2\mu t\}}{l!l!} \\
&= (2\mu t)\exp\{-2\mu t\} \sum_{l=0}^{\infty} \frac{\left(\frac{2\mu t}{2}\right)^{2l}}{l!l!}.
\end{aligned}
$$

A função generalizada de Bessel modificada (veja Abramowitz & Stegun (1972)) é dada por

$$
I_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^k}{k!\Gamma(\nu+k+1)}.
$$

No Capítulo 3, introduzimos a definição dessa função e utilizamos esta quando $\nu = 0$. Para $\nu = 1$, temos

$$
I_1(z) = \frac{z}{2} \sum_{k=0}^{\infty} \frac{\left(\frac{z^2}{4}\right)^k}{k!(k+1)!}.
$$

Neste caso,

$$
E\left[\frac{N(t)}{2^{N(t)}} \left(\begin{array}{c} N(t) \\ \frac{N(t)}{2} \end{array} \right),\ N(t) \text{ par} \mid T = t\right] \quad = \quad (2\mu t)\exp\{-2\mu t\} I_1(2\mu t).
$$

Quando $\nu = 0$, a função é dada por (3.2.5). Com isso, temos

$$
E\left[\frac{N(t)}{2^{N(t)-1}} \left(\begin{array}{c} N(t)-1 \\ \frac{N(t)+1}{2}-1 \end{array} \right),\ N(t) \text{ ímpar} \mid T = t\right] = (2\mu t)\exp\{-2\mu t\} I_0(2\mu t).
$$

Logo,

$$
\begin{aligned}
E[E[|\Delta_{ii'l}(t)| \mid N(t), T=t] \mid T=t] &= (2\mu t)\exp\{-2\mu t\} I_1(2\mu t) + \\
+(2\mu t)\exp\{-2\mu t\} I_0(2\mu t) &= (2\mu t)\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t)).
\end{aligned}
$$

Com isso,

$$\begin{aligned} E|\Delta_{ii'l}(t)| &= E[E[E[|\Delta_{ii'l}(t)| \mid N(t), T] \mid T]] \\ &= E[(2\mu T)\exp\{-2\mu T\}(I_0(2\mu T) + I_1(2\mu T))] \\ &= \int_0^\infty (2\mu t)\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\frac{1}{2N_e}\exp\left\{-\frac{t}{2N_e}\right\}dt \\ &= \frac{2\mu}{2N_e}\int_0^\infty t\exp\left\{-\left(2\mu + \frac{1}{2N_e}\right)t\right\}(I_0(2\mu t) + I_1(2\mu t))dt, \end{aligned}$$

lembrando que $T \sim exp(1/2N_e)$.

Assim, temos que calcular

$$\int_0^\infty t\exp\left\{-\left(2\mu + \frac{1}{2N_e}\right)t\right\}(I_0(2\mu t) + I_1(2\mu t))dt.$$

Segundo Abramowitz & Stegun (1972), a tranformada de Laplace, definida por

$$M(s) = \mathfrak{L}\{f(t)\} = \int_0^\infty e^{-st}f(t)dt,$$

para função $f(t)$ dada por

$$\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a-b}{2}t\right) + I_1\left(\frac{a-b}{2}t\right)\right]$$

é $M(s) = \dfrac{1}{(s+a)^{\frac{1}{2}}(s+b)^{\frac{3}{2}}}$. Em particular, se $s = 1/2N_e$, $a = 4\mu$ e $b = 0$, então,

$$\begin{aligned} M(1/2N_e) &= \int_0^\infty \exp\left\{-\left(\frac{1}{2N_e}\right)t\right\}t\exp(-2\mu t)(I_0(2\mu t) + I_1(2\mu t))dt \\ &= \frac{1}{\left(\frac{1}{2N_e} + 4\mu\right)^{\frac{1}{2}}\left(\frac{1}{2N_e}\right)^{\frac{3}{2}}} = \frac{(2N_e)^2}{(1 + 8N_e\mu)^{\frac{1}{2}}} = \frac{(2N_e)^2}{(1+2\theta)^{\frac{1}{2}}}, \end{aligned}$$

para $\theta = 4N_e\mu$. Temos,

$$E|\Delta_{ii'l}(t)| = \frac{2\mu}{2N_e}\frac{(2N_e)^2}{(1+2\theta)^{\frac{1}{2}}} = \frac{\theta}{(1+2\theta)^{\frac{1}{2}}} = f^*(\theta).$$

De acordo com $E|\Delta_{ii'l}(t)|$, obtemos $E|\Delta_{ii'ggl}(t)|$, no entanto temos que $T_{\tau_{ggl}}$ é uma variável aleatória com distribuição exponencial, $T_{\tau_{ggl}} = T + \tau_{ggl}$, em que T tem distribuição

exponencial com média $2N_e$ e $\rho_{ggl} = 2\mu\tau_{ggl}$, definidos no Capítulo 3. Com isso,

$$\begin{aligned}
E|\Delta_{ii'ggl}(t)| &= E[E[E[|\Delta_{ii'ggl}(t)| \mid N(t), T] \mid T]] \\
&= E[(2\mu T)\exp\{-2\mu T\}(I_0(2\mu T) + I_1(2\mu T))] \\
&= \int_0^\infty (2\mu t)\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\frac{1}{2N_e}\exp\left\{-\frac{t-\tau_{ggl}}{2N_e}\right\}\mathbb{I}(t)_{(\tau_{ggl},\infty)}dt \\
&= \int_{\tau_{ggl}}^\infty (2\mu t)\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\frac{1}{2N_e}\exp\left\{-\frac{t-\tau_{ggl}}{2N_e}\right\}dt \\
&= \frac{2\mu\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}}{2N_e}\int_{\tau_{ggl}}^\infty t\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\exp\left\{-\frac{t}{2N_e}\right\}dt.
\end{aligned}$$

Em particular, se $a = 4\mu$, $b = 0$, $s = 1/2N_e$ e $k = \tau_{ggl}$, temos

$$\begin{aligned}
E|\Delta_{ii'ggl}(t)| &= \frac{2\mu}{2N_e}\int_k^\infty \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right) + I_1\left(\frac{a}{2}t\right)\right]dt \\
&= \frac{2\mu}{2N_e}\int_0^\infty \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right) + I_1\left(\frac{a}{2}t\right)\right]u(t-k)dt,
\end{aligned}$$

em que

$$u(t-k) = \begin{cases} 0, & \text{se } t < k; \\ \frac{1}{2}, & \text{se } t = k; \\ 1, & \text{se } t > k. \end{cases}$$

Assim,

$$\begin{aligned}
E|\Delta_{ii'ggl}(t)| &= \underbrace{\frac{2\mu}{2N_e}\int_0^k \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right) + I_1\left(\frac{a}{2}t\right)\right]dt}_{\text{Zero}} + \\
&+ \frac{2\mu}{2N_e}\int_k^\infty \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right) + I_1\left(\frac{a}{2}t\right)\right]dt \\
&= \frac{2\mu\exp\{sk\}}{2N_e}\left[\int_0^\infty \exp\{-st\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right) + I_1\left(\frac{a}{2}t\right)\right]dt\right] \\
&= \frac{2\mu\exp\{sk\}}{2N_e}\left[\frac{1}{(s+a)^{\frac{1}{2}}s^{\frac{3}{2}}}\right] = \frac{\theta\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}}{(1+2\theta)^{\frac{1}{2}}} \\
&= \exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}f^*(\theta),
\end{aligned}$$

em que $f^*(\theta) = \dfrac{\theta}{(1+2\theta)^{\frac{1}{2}}}$.

Isso implica que

$$E[AM_{Wl}(t)] = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{\begin{pmatrix} 2n \\ 2 \end{pmatrix}}\sum_{i=1}^{2n}\sum_{i'>i}E|\Delta_{ii'ggl}(t)| = f^*(\theta)\frac{1}{G}\sum_{g=1}^{G}\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}.$$

A variância de $AM_{Wl}(t)$ é dada por

$$Var[AM_{Wl}(t)] = \frac{1}{G^2}\sum_{g=1}^{G}\frac{1}{\begin{pmatrix} 2n \\ 2 \end{pmatrix}^2}\sum_{i=1}^{2n}\sum_{i'>i}Var|\Delta_{ii'ggl}(t)|,$$

assumindo que no tempo t as populações sejam independentes e as cópias também o são. E, $Var|\Delta_{ii'ggl}(t)|$ é dada por

$$\begin{aligned} Var|\Delta_{ii'ggl}(t)| &= E|\Delta_{ii'ggl}(t)|^2 - (E|\Delta_{ii'ggl}(t)|)^2 \\ &= E[\Delta^2_{ii'ggl}(t)] - (E|\Delta_{ii'ggl}(t)|)^2 \\ &= \theta + \rho_{ggl} - \frac{\theta^2}{(1+2\theta)}\left(\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right)^2. \end{aligned}$$

Da mesma maneira que $E|\Delta_{ii'ggl}(t)|$, podemos calcular $E|\Delta_{ii'gg'l}(t)|$. Lembre-se que o tempo de coalescência para comparação entre duas populações (g e g', $g \neq g'$), $T_{\tau_{gg'l}}$, é uma variável aleatória com distribuição exponencial, $T_{\tau_{gg'l}} = T + \tau_{gg'l}$, em que T tem distribuição exponencial com média $2N_e$ e $\rho_{gg'l} = 2\mu\tau_{gg'l}$, definidos no Capítulo 3. Com isso,

$$\begin{aligned} E|\Delta_{ii'gg'l}(t)| &= E[E[E[|\Delta_{ii'gg'l}(t)| \mid N(t), T] \mid T]] \\ &= E[(2\mu T)\exp\{-2\mu T\}(I_0(2\mu T) + I_1(2\mu T))] \\ &= \int_0^\infty (2\mu t)\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\frac{1}{2N_e}\exp\left\{-\frac{t-\tau_{gg'l}}{2N_e}\right\}\mathbb{I}(t)_{(\tau_{gg'l},\infty)}dt \\ &= \int_{\tau_{gg'l}}^\infty (2\mu t)\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\frac{1}{2N_e}\exp\left\{-\frac{t-\tau_{gg'l}}{2N_e}\right\}dt \\ &= \frac{2\mu\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}}{2N_e}\int_{\tau_{gg'l}}^\infty t\exp\{-2\mu t\}(I_0(2\mu t) + I_1(2\mu t))\exp\left\{-\frac{t}{2N_e}\right\}dt. \end{aligned}$$

Em particular, se $a = 4\mu$, $b = 0$, $s = 1/2N_e$ e $k = \tau_{gg'l}$, temos

$$\begin{aligned} E|\Delta_{ii'gg'l}(t)| &= \frac{2\mu}{2N_e}\int_k^\infty \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right)+I_1\left(\frac{a}{2}t\right)\right]dt \\ &= \frac{2\mu}{2N_e}\int_0^\infty \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right)+I_1\left(\frac{a}{2}t\right)\right]u(t-k)dt, \end{aligned}$$

em que

$$u(t-k) = \begin{cases} 0, & \text{se } t < k; \\ \frac{1}{2}, & \text{se } t = k; \\ 1, & \text{se } t > k. \end{cases}$$

Assim,

$$\begin{aligned} E|\Delta_{ii'gg'l}(t)| &= \frac{2\mu}{2N_e}\underbrace{\int_0^k \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right)+I_1\left(\frac{a}{2}t\right)\right]dt}_{\text{Zero}} + \\ &+ \frac{2\mu}{2N_e}\int_k^\infty \exp\{-s(t-k)\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right)+I_1\left(\frac{a}{2}t\right)\right]dt \\ &= \frac{2\mu\exp\{sk\}}{2N_e}\left[\int_0^\infty \exp\{-st\}t\exp\left\{-\frac{a}{2}t\right\}\left[I_0\left(\frac{a}{2}t\right)+I_1\left(\frac{a}{2}t\right)\right]dt\right] \\ &= \frac{2\mu\exp\{sk\}}{2N_e}\left[\frac{1}{(s+a)^{\frac{1}{2}}s^{\frac{3}{2}}}\right] = \frac{\theta\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}}{(1+2\theta)^{\frac{1}{2}}} \\ &= \exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta), \end{aligned}$$

em que $f^*(\theta) = \dfrac{\theta}{(1+2\theta)^{\frac{1}{2}}}$.

Isso implica que

$$\begin{aligned} E[AM_{Bl}(t)] &= \frac{1}{\binom{G}{2}}\sum_{g=1}^{G}\sum_{g'>g}\frac{1}{(2n)^2}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}E|\Delta_{ii'gg'l}(t)| \\ &= \frac{f^*(\theta)}{\binom{G}{2}}\sum_{g=1}^{G}\sum_{g'>g}\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}. \end{aligned}$$

A variância de $AM_{Bl}(t)$ é dada por

$$Var[AM_{Bl}(t)] = \frac{1}{\begin{pmatrix} G \\ 2 \end{pmatrix}^2} \sum_{g=1}^{G} \sum_{g'>g} \frac{1}{(2n)^4} \sum_{i=1}^{2n} \sum_{i'=1}^{2n} Var|\Delta_{ii'gg'l}(t)|,$$

assumindo que no tempo t as populações sejam independentes e as cópias também o são, com $\Delta_{ii'gg'l}(t)$ não correlacionados. A $Var|\Delta_{ii'gg'l}(t)|$ é dada por

$$\begin{aligned} Var|\Delta_{ii'gg'l}(t)| &= E|\Delta_{ii'gg'l}(t)|^2 - (E|\Delta_{ii'gg'l}(t)|)^2 \\ &= E[\Delta^2_{ii'ggl}(t)] - (E|\Delta_{ii'gg'l}(t)|)^2 \\ &= \theta + \rho_{gg'l} - \frac{\theta^2}{(1+2\theta)} \left(\exp\left\{ \frac{\tau_{gg'l}}{2N_e} \right\} \right)^2. \end{aligned}$$

Podemos definir $AM_{Wl}(t)$ e $AM_{Bl}(t)$ para diferentes tamanhos amostrais entre as populações, ou seja, $2n_g$ para $g = 1, \ldots, G$. Assim,

$$\mathrm{D}_{ggl}(t) = \sum_{i=1}^{2n_g} \sum_{i'>i} |\Delta_{ii'ggl}(t)|.$$

Então a distância média absoluta para a população g no l-ésimo *locus* no tempo t é

$$\bar{\mathrm{D}}_{ggl}(t) = \frac{\mathrm{D}_{ggl}(t)}{\begin{pmatrix} 2n_g \\ 2 \end{pmatrix}}.$$

A distância média absoluta dentre populações é dada por

$$AM_{Wl}(t) = \frac{1}{G} \sum_{g=1}^{G} \bar{\mathrm{D}}_{ggl}(t).$$

Da mesma forma podemos definir $AM_{Bl}(t)$. Seja

$$\mathrm{D}_{gg'l}(t) = \sum_{i=1}^{2n_g} \sum_{i'=1}^{2n_{g'}} |\Delta_{ii'gg'l}(t)|.$$

Então, a distância média absoluta entre as populações g e g' no l-ésimo *locus* no tempo t é

$$\bar{\mathrm{D}}_{gg'l}(t) = \frac{\mathrm{D}_{gg'l}(t)}{2n_g 2n_{g'}}.$$

A distância média absoluta entre populações é dada por

$$AM_{Bl}(t) = \frac{1}{\binom{G}{2}} \sum_{g'>g} \bar{\mathrm{D}}_{gg'l}(t).$$

Define-se a estatística $AMP_{Wl}(t)$ como a distância média absoluta ponderada pelo tamanho amostral das populações. Isto é,

$$AMP_{Wl}(t) = \sum_{g=1}^{G} w_g \, \bar{\mathrm{D}}_{ggl}(t), \tag{4.2.3}$$

em que $w_g = \dfrac{2n_g}{\sum_{g=1}^{G} 2n_g} \quad g = 1, \ldots, G$. Note que $\sum_g w_g = 1$.

4.3 Teste de homogeneidade

Como já foi discutido, o teste de homogeneidade se traduziria em testar se existem diferenças entre as variações entre e dentre populações. Da mesma forma, o teste de homogeneidade é $H_0 : \rho_{gg'l} = 0$ para todo $g,\ g' = 1, \ldots, G$. Isso implica que sob H_0,

$$\frac{1}{\binom{G}{2}} \sum_{g=1}^{G} \sum_{g'>g} \exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\} = 1 \quad \text{e} \quad \frac{1}{G} \sum_{g=1}^{G} \exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} = 1.$$

A hipótese alternativa é $H_1 : \rho_{gg'} > 0$, para algum $g' \neq g$, ou seja, existe alguma diferença em pelo menos duas populações.

Então, sob $H_0 : \rho_{gg'l} = \rho_{ggl} = \rho_{g'g'l} = \rho = 0$ para todo $g\ , g' = 1, \ldots, G$. Desta forma, sob $H_0 : E(AM_{Totl}(t)) = f^*(\theta)$, ou seja, tenho homogeneidade populacional.

Para esse propósito, define-se a estatística do teste

$$AM_{(B-W)l}(t) = AM_{Bl}(t) - AM_{Wl}(t), \tag{4.3.1}$$

que é a diferença dos desvios absolutos entre populações e dentre populações.

4.3.1 Distribuição de $AM_{Wl}(t)$ e $AM_{Bl}(t)$

Para encontrar a distribuição de $AM_{Wl}(t)$ e $AM_{Bl}(t)$, utilizaremos a teoria de estatística U, estudada na Seção 3.4. Primeiramente, estudaremos as distribuições destas estatísticas para amostras balanceadas. Considere a estatística, $AM_{Wl}(t)$, já definido na equação (4.2.1), temos que $|\Delta_{ii'ggl}(t)|$ é o kernel para o parâmetro $f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}$. A estatística U correspondente é

$$\bar{\mathrm{D}}_{ggl}(t) = \frac{1}{\begin{pmatrix} 2n \\ 2 \end{pmatrix}} \sum_p |\Delta_{ii'ggl}(t)|,$$

em que $\sum_p$ é a soma sobre todas combinações duas a duas das cópias $\{1, 2, \ldots, 2n\}$ e é igual a diferença média de Gini (ver Seção 3.4). Logo $\bar{\mathrm{D}}_{ggl}(t)$ é um estimador não viesado para $f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}$.

Primeiramente, temos

$$\begin{aligned} E|Y_{igl}(t) - Y_{i'gl}(t)| &= E|Y_{igl}(t) - \eta_{gl} + \eta_{gl} - Y_{i'gl}(t)| \\ &\leq E|Y_{igl}(t) - \eta_{gl}| + E|\eta_{gl} - Y_{i'gl}(t)| \quad \text{desigualdade triangular} \\ \Rightarrow E|Y_{igl}(t) - \eta_{gl}| &\geq \frac{f^*(\theta)}{2}\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}. \end{aligned} \tag{4.3.2}$$

Além disso, $Y_{igl}(t) - \eta_{gl}$ é integrável, se somente se, $E|Y_{igl}(t) - \eta_{gl}| < \infty$. Temos que $E[Y_{igl}(t) - \eta_{gl}] = 0$, então o primeiro momento de $Y_{igl}(t) - \eta_{gl}$ existe, logo

$$0 < \frac{f^*(\theta)}{2}\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} \leq E|Y_{igl}(t) - \eta_{gl}| < \infty.$$

Se $\theta < \infty$ e $\rho_{ggl} < \infty$, então $E[h^2_{ii'ggl}(t)] = E[\Delta^2_{ii'ggl}(t)] = \theta + \rho_{ggl} < \infty$. Aplicando a teoria para estatística U, temos que para $c = 1$

$$\begin{aligned} h_1(x) &= E[|Y_{igl}(t) - Y_{i'gl}(t)| \mid Y_{i'gl}(t) = x] \\ &= E|Y_{igl}(t) - x|, \end{aligned}$$

lembrando que $Y_{1gl}(t), \ldots, Y_{2ngl}(t)$, $(g = 1, 2, \ldots, G)$ são independentes com distribuição comum F, com média η_{gl} e variância $\sigma^2 = (\theta + \rho_{ggl})/2$. Então

$$h_1(x) = E|Y_{igl}(t) - x|$$

$$\begin{aligned} &= E|Y_{igl}(t) - \eta_{gl} + \eta_{gl} - x| \geq |E[(Y_{igl}(t) - \eta_{gl}) + (\eta_{gl} - x)]| \qquad (4.3.3) \\ &= |\eta_{gl} - x|, \end{aligned}$$

em que a relação (4.3.3) se deve a Desigualdade de Jensen, Definição A.1. Com isso,

$$\begin{aligned} \tilde{h}_1(x) &= h_1 - f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} \geq |\eta_{gl} - x| - f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} \\ \tilde{h}_1^2(x) &\geq (x - \eta_{gl})^2 - 2f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}|\eta_{gl} - x| + \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right)^2. \end{aligned}$$

Então

$$\begin{aligned} \zeta_1^g &= E[\tilde{h}_1^2(Y_{igl}(t))] = Var[h_1(Y_{igl}(t))] \geq E(Y_{igl}(t) - \eta_{gl})^2 + \\ &- 2f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}E|\eta_{gl} - Y_{igl}(t)| + \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right)^2 \\ &\geq \frac{\theta + \rho_{ggl}}{2} - 2f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}E|\eta_{gl} - Y_{igl}(t)| + \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right)^2. \end{aligned}$$

Logo,

$$\begin{aligned} \zeta_1^g &\geq \frac{\theta + \rho_{ggl}}{2} - 2f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}E|\eta_{gl} - Y_{igl}(t)| + \left(f^*(\theta)\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}\right)^2 \\ &= \frac{\theta + \rho_{ggl}}{2} + (f^*(\theta) - E|\eta_{gl} - Y_{igl}(t)|)^2 - (E|\eta_{gl} - Y_{igl}(t)|)^2. \end{aligned}$$

Se a função $\varphi(x) = |x|^p$, com $p \geq 1$, então $E|Y|^p \geq |EY|^p$ pois, fazendo $X = |Y|$ e aplicando a desigualdade de Jensen, obtemos $E[X]^p \geq (EX)^p$. Assim,

$$E|Y|^p \geq (E|Y|)^p \geq |E(Y)|^p, \qquad (4.3.4)$$

em que a última desigualdade se deve ao fato de que $E|Y| \geq |EY|$. Se $p = 2$ e aplicando esse fato, temos

$$\begin{aligned} \zeta_1^g &\geq \frac{\theta + \rho_{ggl}}{2} + \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} - E|\eta_{gl} - Y_{igl}(t)|\right)^2 - (E|\eta_{gl} - Y_{igl}(t)|)^2 \\ &\geq \frac{\theta + \rho_{ggl}}{2} + \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} - E|\eta_{gl} - Y_{igl}(t)|\right)^2 - E|\eta_{gl} - Y_{igl}(t)|^2 \\ &= \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} - E|\eta_{gl} - Y_{igl}(t)|\right)^2 \geq 0, \end{aligned}$$

$\zeta_1^g = 0$, se e somente se $f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} = E|\eta_{gl} - Y_{igl}(t)|$. Assim a distribuição da estatística U será degenerada. Se $\zeta_1^g > 0$ e $E[h^2_{ii'ggl}(t)] < \infty$, satisfazem as duas condições do Teorema 3.3 do Capítulo 3. Portanto, para $m = 2$

$$\sqrt{2n}\left(\bar{\mathrm{D}}_{ggl}(t) - f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right) \xrightarrow{\mathcal{D}} N\left(0, 4\zeta_1^g\right).$$

Não sabemos qual é o valor de ζ_1^g, somente que $\zeta_1^g \geq \left(f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\} - E|\eta_{gl} - Y_{igl}(t)|\right)^2$ e que depende da população. Assim, temos

$$\sqrt{2nG}\left(AM_{Wl}(t) - f^*(\theta)\frac{1}{G}\sum_{g=1}^{G}\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{4}{G}\sum_{g=1}^{G}\zeta_1^g\right),$$

pois considera-se que no tempo t as populações sejam independentes.

Agora, consideremos $|\Delta_{ii'gg'l}(t)|$ o kernel para o parâmetro $\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\} f^*(\theta)$. A estatística U correspondente é

$$\bar{\mathrm{D}}_{gg'l}(t) = \frac{1}{(2n)^2}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}|\Delta_{ii'gg'l}(t)|,$$

a qual é um estimador não viesado para $f^*(\theta)\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}$. Assim, $\bar{\mathrm{D}}_{gg'l}(t)$ é uma estatística U generalizada de grau $\mathbf{m} = (1, 1)$.

Aplicando a teoria de estatística U, temos que $E\left[\Delta^2_{ii'gg'l}(t)\right] = \theta + \rho_{gg'l} < \infty$, o que é razoável considerando que o número médio de mutações é na ordem de 10^{-2} a 10^{-5}. Assim,

$$\begin{aligned}
\Psi_{01}(x) &= E[|Y_{igl}(t) - Y_{i'g'l}(t)| \mid Y_{i'g'l}(t) = x] \\
&= E[|Y_{igl}(t) - x|] = E[|Y_{igl}(t) - \eta_{gl} + \eta_{gl} - x|] \\
&\geq |E[(Y_{igl}(t) - \eta_{gl}) + (\eta_{gl} - x)] = |\eta_{gl} - x| \quad \text{conforme a relação (4.3.3).}
\end{aligned}$$

Logo, $\Psi_{01}(Y_{i'g'l}(t)) \geq |\eta_{gl} - Y_{i'g'l}(t)|$. Da mesma forma, $\Psi_{10}(Y_{igl}(t)) \geq |Y_{igl}(t) - \eta_{g'l}|$. Com isso,

$$\zeta_{01} = E\left[\Psi_{01}(Y_{i'g'l}(t)) - \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\} f^*(\theta)\right)\right]^2$$

$$
\begin{aligned}
&= E\left[\Psi_{01}^2(Y_{i'g'l}(t)) - 2\Psi_{01}(Y_{i'g'l}(t))\left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right) + \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right)^2\right] \\
&\geq E(Y_{igl}(t) - \eta_{g'l})^2 - 2E|Y_{igl}(t) - \eta_{g'l}|\left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right) + \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right)^2.
\end{aligned}
$$

Sabemos que pela relação (3.5.4),

$$E(Y_{igl}(t) - \eta_{g'l})^2 = \frac{\theta - \rho_{g'g'l}}{2} + \rho_{gg'l}.$$

Assim,

$$
\begin{aligned}
\zeta_{01} &\geq \frac{\theta - \rho_{g'g'l}}{2} + \rho_{gg'l} - 2E|Y_{igl}(t) - \eta_{g'l}|\left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right) + \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right)^2 \\
&= \frac{\theta - \rho_{g'g'l}}{2} + \rho_{gg'l} + \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta) - E|Y_{igl}(t) - \eta_{g'l}|\right)^2 - (E|Y_{igl}(t) - \eta_{g'l}|)^2
\end{aligned}
$$

e utilizando a desigualdade (4.3.4), temos

$$
\begin{aligned}
\zeta_{01} &\geq \frac{\theta - \rho_{g'g'l}}{2} + \rho_{gg'l} + \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta) - E|Y_{igl}(t) - \eta_{g'l}|\right)^2 - (E|Y_{igl}(t) - \eta_{g'l}|)^2 \\
&\geq \frac{\theta - \rho_{g'g'l}}{2} + \rho_{gg'l} + \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta) - E|Y_{igl}(t) - \eta_{g'l}|\right)^2 - E|Y_{igl}(t) - \eta_{g'l}|^2 \\
&\geq \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta) - E|Y_{igl}(t) - \eta_{g'l}|\right)^2 \geq 0,
\end{aligned}
$$

com igualdade se somente se $\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta) = E|Y_{igl}(t) - \eta_{g'l}|$. Da mesma forma

$$\zeta_{10} \geq \left(\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta) - E|Y_{i'g'l}(t) - \eta_{gl}|\right)^2 \geq 0.$$

Usando os resultados de Hoeffding (1948), temos que para $\zeta_{01} > 0$, $\zeta_{10} > 0$ e $E\left[\Delta^2_{ii'gg'l}(t)\right] = \theta + \rho_{gg'l} < \infty$, vale a extensão do Teorema 3.3 para a estatística U generalizada, e

$$\sqrt{2n}\left(\bar{\mathrm{D}}_{gg'l}(t) - \exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right) \xrightarrow{\mathcal{D}} N\left(0, \zeta_{10} + \zeta_{01}\right).$$

Para encontrar a distribuição de $AM_{Bl}(t)$, temos que encontrar a covariância entre $\bar{\mathrm{D}}_{gg'l}(t)$ e $\bar{\mathrm{D}}_{kk'l}(t)$, que nã será calculada. Temos que

$$Var\left[\sum_{g=1}^{G}\sum_{g'>g}\bar{\mathrm{D}}_{gg'l}(t)\right] = \frac{4}{G^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}\sum_{k=1}^{G}\sum_{k'>k}\mathrm{cov}\left(\bar{\mathrm{D}}_{gg'l}(t), \bar{\mathrm{D}}_{kk'l}(t)\right).$$

Conforme visto na Seção 3.5.1, temos as seguintes situações:

1. quando $k = g$ e $k' = g'$, temos a variância, ou seja,

$$\sum_{g=1}^{G}\sum_{g'>g} \text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{gg'l}(t)\right) = \sum_{g=1}^{G}\sum_{g'>g} Var\left[\bar{\text{D}}_{gg'l}(t)\right];$$

que é dada em termos de ζ_{01} e ζ_{10};

2. $\text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{kk'l}(t)\right) = 0$ para $g \neq g' \neq k \neq k'$, pois assume-se que as populações são independentes;

3. quando $g = k$, temos

$$\sum_{g=1}^{G}\sum_{g'>g}\sum_{k'>g'} \text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{gk'l}(t)\right) \quad \text{ou} \quad \sum_{g=1}^{G}\sum_{k'>g}\sum_{g'>k'} \text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{gk'l}(t)\right);$$

4. quando $g' = k'$, temos

$$\sum_{g=1}^{G}\sum_{k>g}\sum_{g'>k} \text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{kg'l}(t)\right) \quad \text{ou} \quad \sum_{k=1}^{G}\sum_{g>k}\sum_{g'>k} \text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{kg'l}(t)\right);$$

5. quando $g' = k$, temos

$$\sum_{g=1}^{G}\sum_{g'>g}\sum_{k'>k} \text{cov}\left(\bar{\text{D}}_{gg'l}(t), \bar{\text{D}}_{g'k'l}(t)\right);$$

6. quando $g = k'$, corresponde ao caso em que $g' = k$.

Com isso,

$$\left(AM_{Bl}(t) - \frac{f^*(\theta)}{\begin{pmatrix} G \\ 2 \end{pmatrix}} \sum_{g=1}^{G}\sum_{g'>g} \exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\} \right) \xrightarrow{\mathcal{D}}$$
$$N\left(0, \frac{4}{G^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{\zeta_{10}+\zeta_{01}}{2n}\right) + \chi\right),$$

em que χ representa a covariância para amostras balanceadas.

Para amostras não balanceadas, os mesmos resultados valem, ou seja, se $\zeta_1^g > 0$ e $E[h^2_{ii'ggl}(t)] < \infty$,

$$\sqrt{2n_g}\left(\bar{\mathrm{D}}_{ggl}(t) - f^*(\theta)\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right) \xrightarrow{\mathcal{D}} N\left(0, 4\zeta_1^g\right),$$

caso contrário, se $\zeta_1^g = 0$, esta será degenerada. Com isso,

$$\left(AM_{Wl}(t) - \frac{f^*(\theta)}{G}\sum_{g=1}^{G}\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{4}{G^2}\sum_{g=1}^{G}\frac{\zeta_1^g}{2n_g}\right).$$

Como também, temos que se $\zeta_{01} > 0$, $\zeta_{10} > 0$ e $E\left[\Delta^2_{ii'gg'l}(t)\right] = \theta + \rho_{gg'l} < \infty$ e

$$\left(\bar{\mathrm{D}}_{gg'l}(t) - \exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}f^*(\theta)\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{\zeta_{10}}{2n_g} + \frac{\zeta_{01}}{2n_{g'}}\right).$$

Além disso,

$$\left(AM_{Bl}(t) - \frac{f^*(\theta)}{\begin{pmatrix} G \\ 2 \end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g}\exp\left\{\frac{\tau_{gg'l}}{2N_e}\right\}\right) \xrightarrow{\mathcal{D}}$$

$$N\left(0, \frac{4}{G^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{\zeta_{10}}{2n_g} + \frac{\zeta_{01}}{2n_{g'}}\right) + \chi^*\right),$$

em que χ^* representa a covariância para amostras não balanceadas.

Para a estatística $AMP_{Wl}(t)$, temos

$$\left(AMP_{Wl}(t) - \frac{f^*(\theta)}{G}\sum_{g=1}^{G}\exp\left\{\frac{\tau_{ggl}}{2N_e}\right\}\right) \xrightarrow{\mathcal{D}} N\left(0, 4\sum_{g=1}^{G}w_g^2\zeta_1^g\right).$$

Note que não encontramos os valores de ζ_1^g, ζ_{01} e ζ_{10}, somente mostramos que estes são maiores ou iguais a zero. Para estudos futuros é interessante encontrar formas de estimar esses valores. Sen (1960) e Arvesen (1969) utilizaram o estimador jackknife para encontrar a variância de estatísticas U baseado em uma amostra, ou no nosso caso, uma população. Considere a estatística U_n, a variância estimada é dada por

$$\widehat{Var(U_n)} = n^2(n-1)\begin{pmatrix} n-1 \\ m \end{pmatrix}^{-2}\sum_{c=0}^{m}(cn - m^2)S_c, \tag{4.3.5}$$

em que $Sc = \sum \phi(\mathbf{X}_{i1}, \mathbf{X}_{i2})\,\phi(\mathbf{X}_{i3}, \mathbf{X}_{i4})$ para alguma reamostra $\{i_1, i_2, i_3, i_4\}$ de $\{1, \ldots, n\}$ de forma que tenhamos c índices coincidentes; $c = 1, \ldots, m$, ou seja, por exemplo, se $c = 0$ então não temos nenhum índice coincidente. Pinheiro et al. (2006) aplicaram essa variância para uma determinada estatística U para um tamanho de seqüência $n = 48$. No caso de nossa aplicação, o tamanho da amostra é igual à 3228 (ver Capítulo 5), o que exigiria um esforço e um tempo computacional muito grande para fazer todas as combinações. No entanto, se o tamanho da amostra fosse menor, poderíamos utilizar esse estimador para encontrar a variância da estatística do teste sob H_0, pois sob essa hipótese não teríamos divisão populacional.

A estatística do teste, $AM_{(B-W)l}(t)$, dada por (4.3.1) é combinação linear de variáveis aleatórias assintoticamente Normais, com as restrição de que $\zeta_1^g > 0$, $\zeta_{10} > 0$ e $\zeta_{01} > 0$, então esta terá também distribuição Normal, conforme Proposição 3.2.

Conforme já foi discutido no Capítulo 3, sob H_0, temos que $\rho_{gg'} = 0$ para $g, g' = 1, 2, \ldots, G$ e $\eta_{gl} = \eta_l$. Com isso, sob H_0, $\zeta_1^g = \zeta_1$ e

$$\zeta_1 \;\geq\; \left(f^*(\theta) - E|\eta_l - Y_{il}(t)|\right)^2 \geq 0.$$

Se $\zeta_1 > 0$, então

$$\sqrt{2nG}\left(AM_{Wl}(t) - f^*(\theta)\right) \xrightarrow{\mathcal{D}} N\left(0, 4\zeta_1\right).$$

O mesmo para a estatística $AM_{Wl}(t)$ definida para amostras não balanceadas e para $AMP_{Wl}(t)$, ou seja,

$$\left(AMP_{Wl}(t) - f^*(\theta)\right) \xrightarrow{\mathcal{D}} N\left(0, 4\sum_{g=1}^{G} w_g^2 \zeta_1\right).$$

Temos, também,

$$\zeta_{01} = \zeta_{10} \;\geq\; \left(f^*(\theta) - E|Y_{i'l}(t) - \eta_l|\right)^2 \geq 0.$$

Se $\zeta_{01} > 0$, vale

$$\left(AM_{Bl}(t) - f^*(\theta)\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{4\zeta_{10}}{2nG(G-1)} + \chi_0\right),$$

em que χ_0 representa a covariância para amostras balanceadas sob H_0. Esse resultado vale para $AM_{Bl}(t)$ definido para amostras não balanceadas.

Como a distribuição da estatística do teste é combinação linear de estatísticas com distribuições assintóticas Normais, então, sob H_0, $AM_{(B-W)}(t)$ tem distribuição assintoticamente Normal conforme Proposição 3.2.

Capítulo 5

Simulação e Aplicação

5.1 Simulação

Através de simulação iremos mostrar o comportamento da distribuição da estatística do teste, verificando sua convergência para a distribuição Normal padrão (com média 0 e variância 1) de acordo com diferentes tamanhos de amostra. Esse resultado foi demonstrado analiticamente nas Seções 3.5.3 e 4.3.1. Utilizaremos a simulação descrita a seguir, cujos programas foram feitos no MATLAB versão 6.5.

Primeiramente, simularemos um processo de coalescência com mutação para microsatélites sob o modelo de mutação "*stepwise*" de um passo. Shriver et al. (1993) utilizaram o tamanho populacional efetivo (N_e) igual à 5000. Então a população tem 10000 ($2N_e$) cromossomos com um tamanho alélico fixado de maneira arbitrária, com isso nenhum limite foi imposto no número de repetições. Uma geração consiste de uma amostragem aleatória com reposição de $2N_e$ cromossomos da população da geração precedente. Para cada alelo, um número aleatório é gerado com distribuição Uniforme (0,1) para identificar se esse alelo sofrerá mutação ou não. Dados uma taxa de mutação μ e um número, x, gerado com distribuição Uniforme (0,1):

- se $x < \mu/2$, então o alelo torna-se uma unidade maior,

- se $x > 1 - \mu/2$, então o alelo torna-se uma unidade menor e
- caso contrário, o alelo continua do mesmo tamanho.

A reamostra da população continua para 20000 ($4N_e$) gerações, ou seja, repetimos esse processo 20000 vezes. Vamos considerar 19 taxas de mutação variando no intervalo 10^{-2} a 10^{-5}. Utilizando essa simulação vamos constituir a população sob H_0, ou seja, sob essa hipótese temos uma única população.

Na Figura 5.1 podemos ver a distribuição dos números distintos de tamanhos alélicos conforme a taxa de mutação. Vemos que, conforme aumenta a taxa, o número de tamanhos alélicos distintos aumenta, conforme podemos ver também na Figura 3.3. Assim, para continuar a simulação consideremos as seguintes taxas de mutação 10^{-2}, 0,0005 e 10^{-5}, em que temos 36, 16 e 3 tamanhos alélicos distintos, respectivamente. Para a taxa de mutação igual à $0,01$ temos $\theta = 200$, ou seja, o número esperado de mutações por *locus* por geração é 200. Da mesma forma, para as taxas de mutações $0,0005$ e 10^{-5}, temos $\theta = 10$ e $\theta = 0,2$, respectivamente.

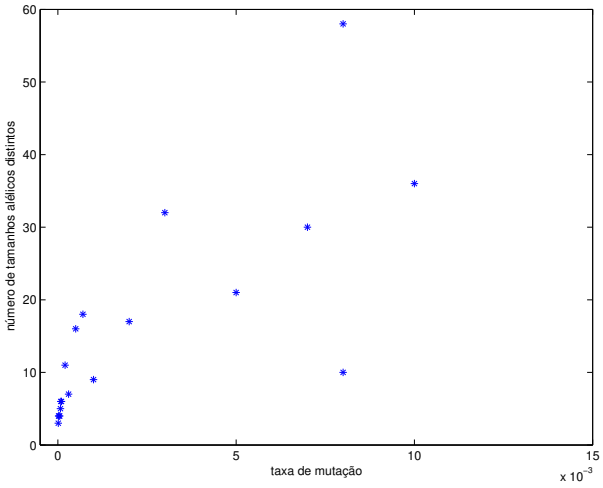

Figura 5.1: Número de tamanhos alélicos distintos para taxas de mutações variando de 10^{-2} a 10^{-5}.

Suponha que fixamos o tamanho amostral em n indivíduos com N_e igual à 5000. Simularemos 1000 amostras, cada uma de tamanho n e sem reposição das populações

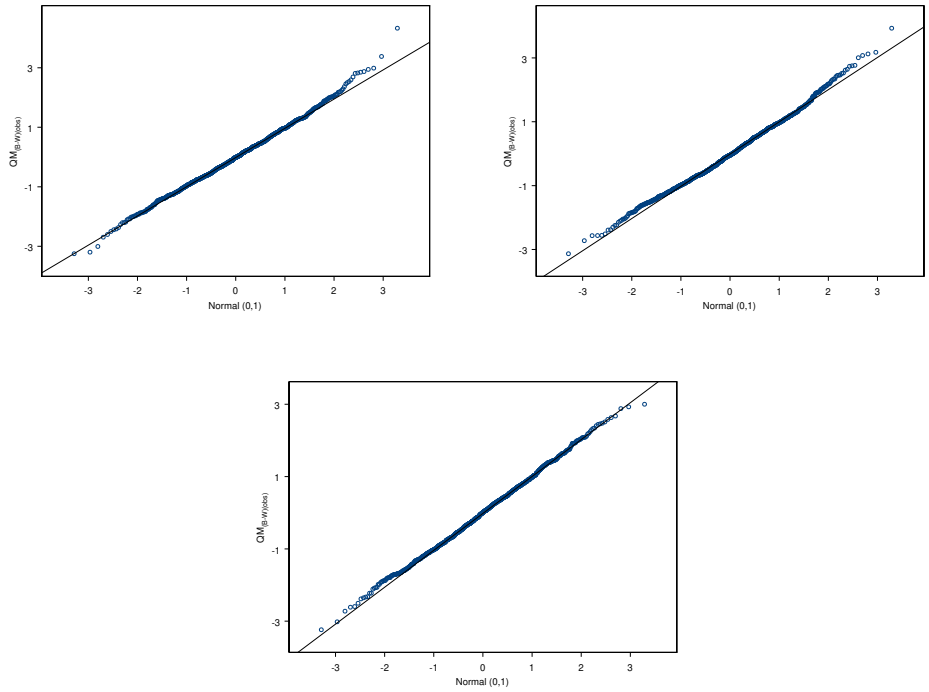

Figura 5.2: Q-Q Normal da estatística $QM_{(B-W)l}(t)$ para $\theta = 200, 10$ e 0,2 e $n = 300$, respectivamente.

constituídas pelas taxas de mutação 10^{-2}, 0,0005 e 10^{-5}. Consideramos que temos 3 populações distintas. Para cada uma das 1000 amostras definimos os tamanhos amostrais de cada população n_1, n_2 e n_3 e calculamos os valores das estatísticas $QM_{(B-W)l}(t) = QM_{Bl}(t) - QM_{Wl}(t)$ e $AM_{(B-W)l}(t) = AM_{Bl}(t) - AM_{Wl}(t)$, para amostras balanceadas e $QM_{(B-PW)l}(t) = QM_{Bl}(t) - QMP_{Wl}(t)$ e $AM_{(B-PW)l}(t) = AM_{Bl}(t) - AMP_{Wl}(t)$, para amostras não balanceadas. Ao final, para cada estatística subtraímos a sua média e dividimos pelo seu desvio padrão e comparamos com a distribuição Normal padrão através de um gráfico Q-Q Normal.

Primeiramente, considere amostras balanceadas e $n = 300$. Os gráficos de Q-Q Normal de $QM_{(B-W)}(t)$ podem ser vistos na Figura 5.2, para $\theta = 200, 10$ e 0,2. O comportamento com respeito a normalidade é razoável para essa estatística para todos θ's. Apesar disso, para $\theta = 200$ e 10, temos um leve desvio nas caudas, mas é pequeno.

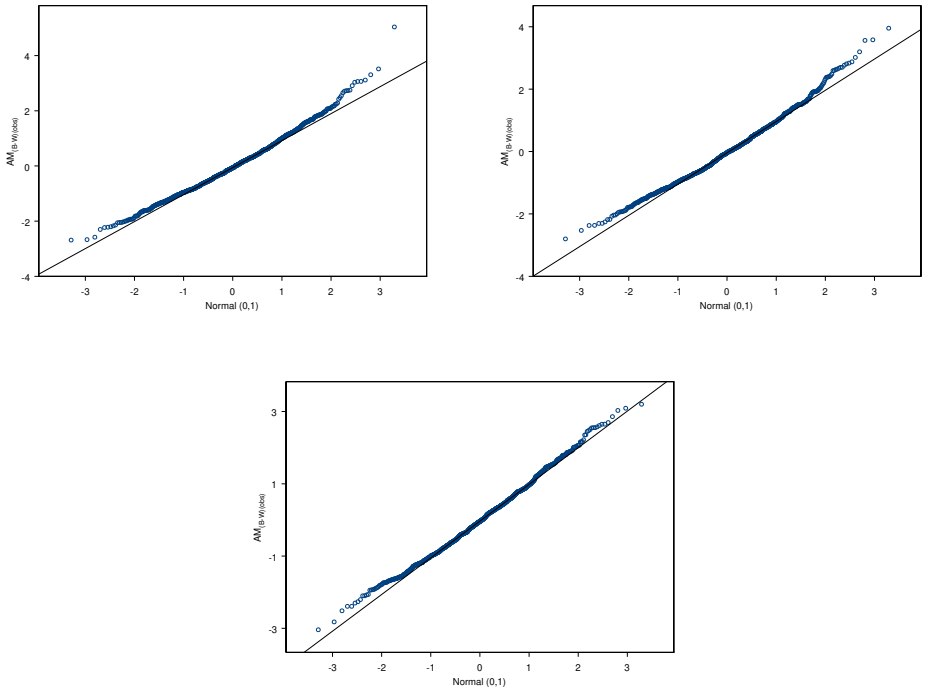

Figura 5.3: Q-Q Normal da estatística $AM_{(B-W)l}(t)$ para $\theta = 200, 10$ e 0,2 e $n = 300$, respectivamente.

Os gráficos Q-Q Normal para a estatística do teste baseada nos desvio absolutos, $AM_{(B-W)}(t)$, podem ser observados na Figura 5.3. Para $\theta = 200$ e 10, temos maiores desvios nas caudas. No entanto, temos um comportamento razoável com respeito a normalidade.

Comparando as duas Figuras 5.2 e 5.3, temos que os resultados para normalidade das estatística $AM_{(B-W)}(t)$ e $QM_{(B-W)}(t)$ são razoáveis para $n = 300$ e são melhores para $\theta = 0, 2$.

Para amostras não balanceadas, fixemos 3 tamanhos amostrais $n = 300$, $n = 400$ e $n = 500$. Para $n = 300$, temos no total 600 cromossomos, sendo que $n_1 = 100$, $n_2 = 125$ e $n_3 = 75$. Para $n = 400$, no total são 800 cromossomos, sendo que $n_1 = 150$, $n_2 = 175$ e $n_3 = 75$. Para $n = 500$, no total são 1000 cromossomos, sendo que $n_1 = 200$, $n_2 = 250$ e $n_3 = 50$.

Na Figura 5.4, vemos os gráficos Q-Q Normal da estatística do teste baseada nos desvios quadráticos para $\theta = 200$. No primeiro gráfico, para $n = 300$, vemos um desvio da reta de distribuição Normal padrão, o que indica caudas pesadas. Conforme aumentamos o tamanho amostral esse desvio desaparece.

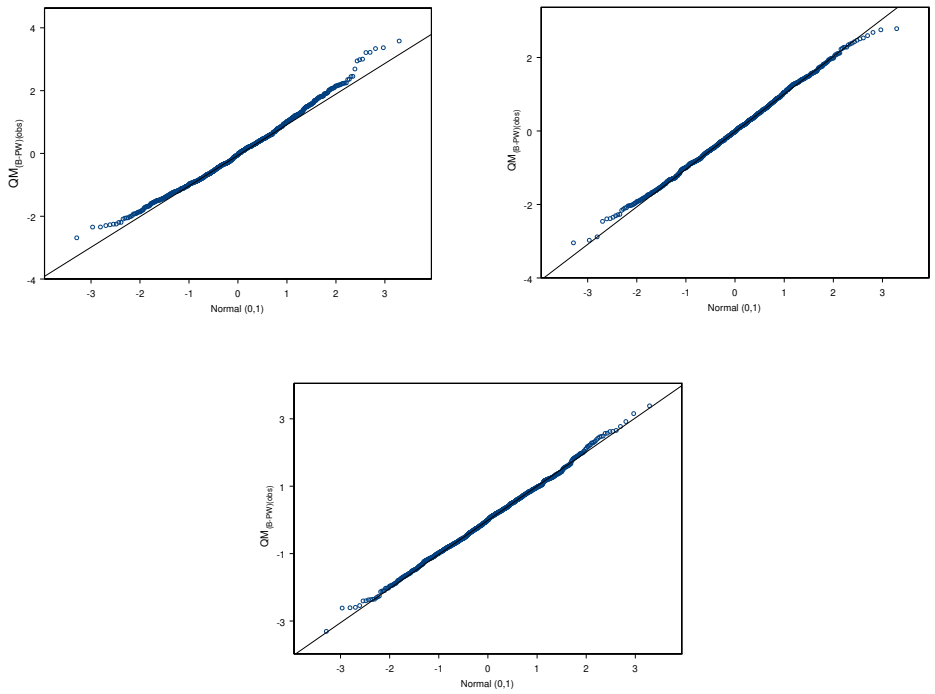

Figura 5.4: Q-Q Normal da estatística $QM_{(B-PW)l}(t)$ para $\theta = 200$ e tamanhos amostrais 300, 400 e 500, respectivamente.

Os gráficos Q-Q Normal da estatística do teste baseada nos desvios absolutos podem ser observados na Figura 5.5, para $\theta = 200$. Assim como a estatística baseada nos desvios quadráticos, para $n = 300$ temos um desvio maior da reta de distribuição Normal padrão, o que indica caudas pesadas.

Note que, para esse valor de θ a distribuição da estatística $QM_{(B-PW)l}(t)$ se aproxima melhor da distribuição Normal do que a distribuição de $AM_{(B-PW)l}(t)$, principalmente quando o tamanho amostral é igual a 300. Para ambas estatísticas temos desvios nas caudas, que são mais acentuados quando o tamanho amostral é igual a 300. Em ambos

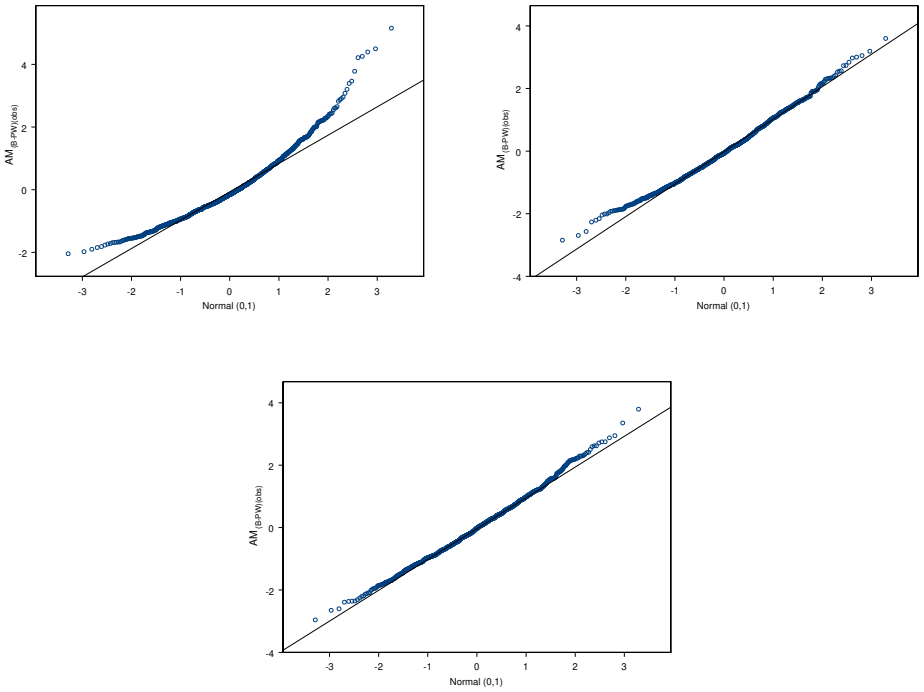

Figura 5.5: Q-Q Normal da estatística $AM_{(B-PW)l}(t)$ para $\theta = 200$ e tamanhos amostrais 300, 400 e 500, respectivamente.

os casos, para o tamanho amostral de 300, as caudas das distribuições das estatísticas do teste são mais pesadas do que as caudas da distribuição Normal padrão. Isso diminui quando o tamanho amostral aumenta.

Na Figura 5.6, observamos os gráficos Q-Q Normal da estatística do teste baseada nos desvios quadráticos para o valor de $\theta = 10$. No primeiro gráfico, para $n = 300$, vemos um desvio da reta de distribuição Normal padrão, indicando caudas pesadas. Para $n = 400$ esse desvio diminui, mas continua. Para $n = 500$, podemos ver um leve desvio na cauda à esquerda.

Os gráficos Q-Q Normal da estatística do teste baseada nos desvios absolutos para $\theta = 10$ encontram-se na Figura 5.7. Semelhante à estatística baseada nos desvios quadráticos, para $n = 300$ e $n = 400$ temos um desvio da reta de distribuição Normal padrão, indicando caudas pesadas. Para $n = 500$ esse desvio diminui.

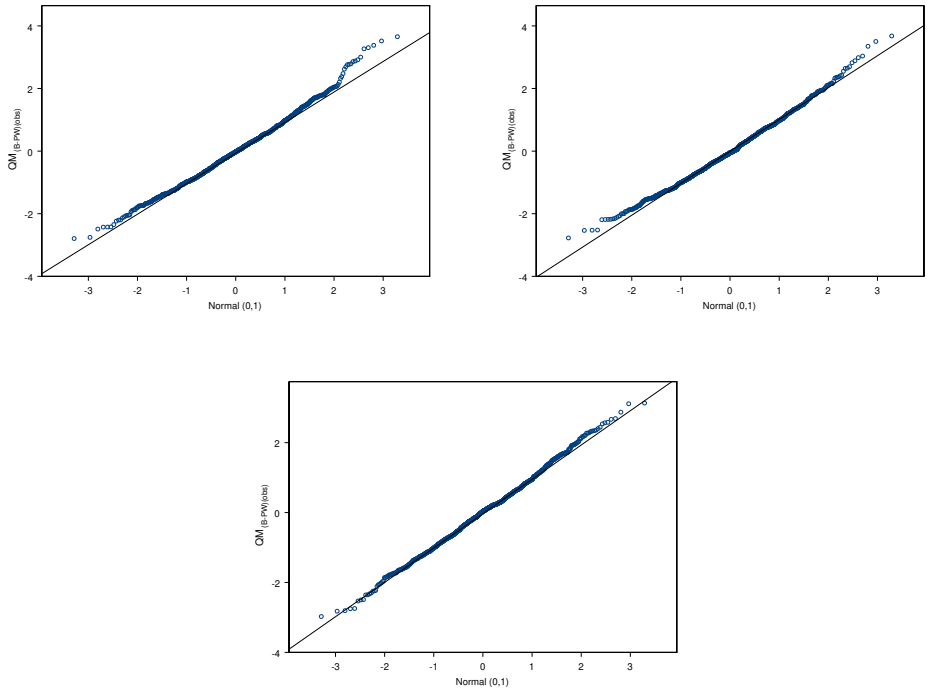

Figura 5.6: Q-Q Normal da estatística $QM_{(B-PW)l}(t)$ para $\theta = 10$ e tamanhos amostrais 300, 400 e 500, respectivamente.

Para $\theta = 10$, similarmente a $\theta = 200$, a distribuição da estatística $QM_{(B-PW)l}(t)$ se aproxima melhor da distribuição Normal do que a distribuição de $AM_{(B-PW)l}(t)$. Os desvios são maiores para $n = 300$ e $n = 400$, em que ambos os casos, as caudas são mais pesadas do que as caudas da distribuição Normal padrão. Para $n = 500$, o desvio diminui, mas ainda ocorre.

Na Figura 5.8, encontram-se os gráficos Q-Q Normal da estatística do teste baseada nos desvios quadráticos para o valor de $\theta = 0,2$. Nos dois primeiros gráficos, para $n = 300$ e $n = 400$, respectivamente, vemos um leve desvio da reta de distribuição Normal padrão, o que indica caudas levemente pesadas. Para $n = 500$, esse desvio é muito pequeno.

Os gráficos Q-Q Normal da estatística do teste baseada nos desvios absolutos podem ser vistos na Figura 5.9, para $\theta = 0,2$. Nos dois primeiros gráficos, para $n = 300$ e $n = 400$, respectivamente, há um desvio da reta de distribuição Normal padrão, indicando caudas

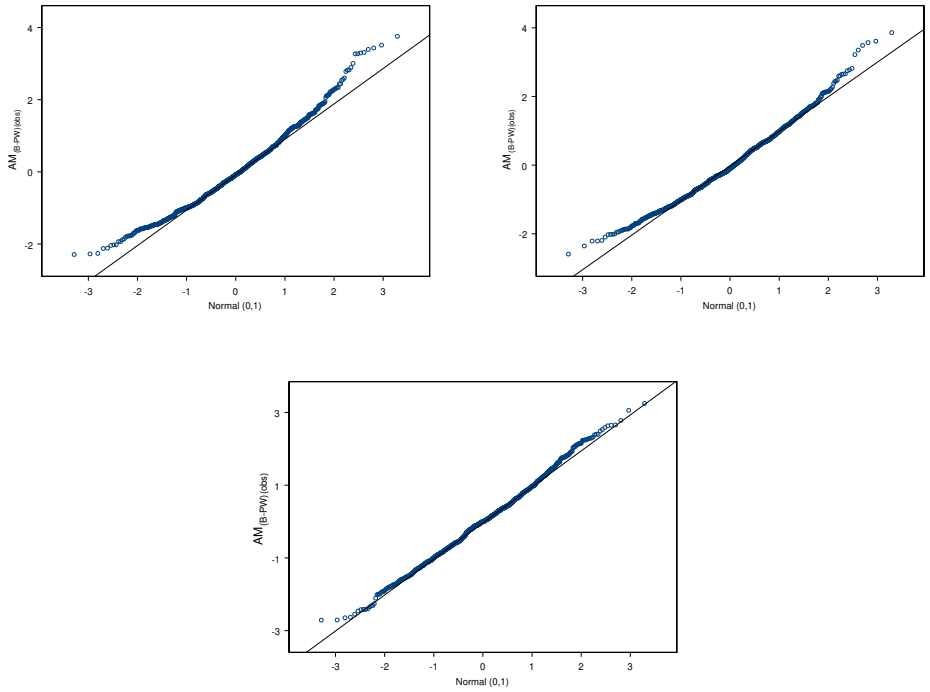

Figura 5.7: Q-Q Normal da estatística $AM_{(B-PW)l}(t)$ para $\theta = 10$ e tamanhos amostrais 300, 400 e 500, respectivamente.

pesadas. Esse desvio é um pouco maior comparado com a estatística baseada nos desvios quadráticos. Para $n = 500$, esse desvio diminui.

Similarmente a $\theta = 200$ e a $\theta = 10$, a distribuição da estatística $QM_{(B-PW)l}(t)$ se aproxima melhor da distribuição Normal do que a distribuição de $AM_{(B-PW)l}(t)$. Para a estatística $QM_{(B-PW)l}(t)$ temos um leve desvio na cauda à direita, quando o tamanho amostral é 300 e 400. Quando consideramos a estatística $AM_{(B-PW)l}(t)$ as caudas são mais pesadas do que as caudas da distribuição Normal padrão.

No geral, os resultados com relação à normalidade para $\theta = 0,2$ foram melhores do que os resultados obtidos para $\theta = 200$ e $\theta = 10$, tanto para a amostra balanceadas com para amostras não balanceadas.

Os resultados com respeito à normalidade para amostras balanceadas são razoáveis para $n = 300$, tanto para $QM_{(B-W)}(t)$ como $AM_{(B-W)l}(t)$.

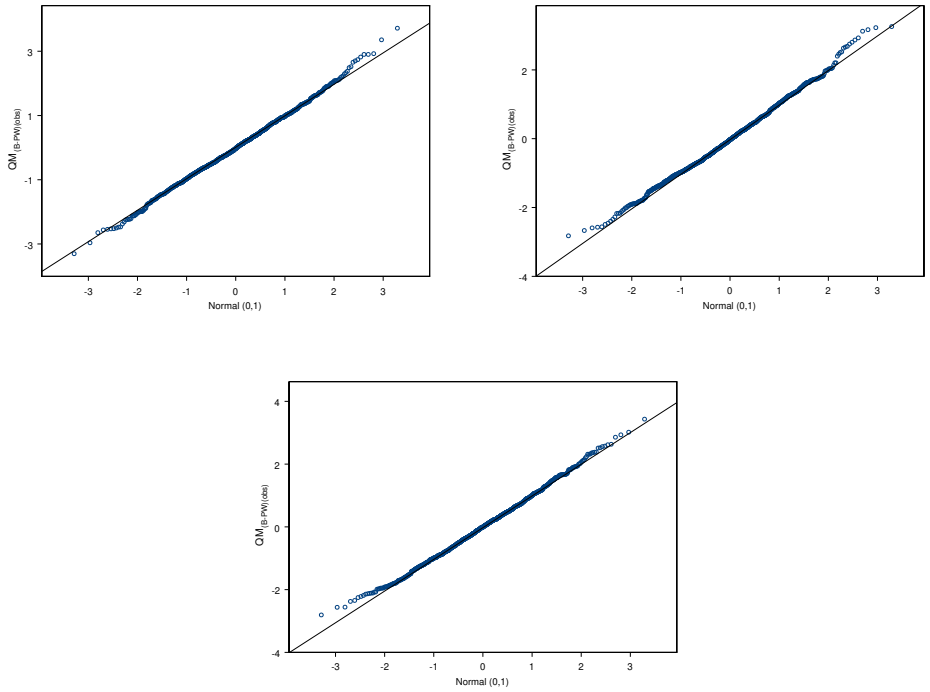

Figura 5.8: Q-Q Normal da estatística $QM_{(B-PW)l}(t)$ para $\theta = 0,2$ e tamanhos amostrais 300, 400 e 500, respectivamente.

Considerando amostras não balanceadas, a estatística $QM_{(B-PW)l}(t)$ se aproxima melhor de uma Normal padrão do que $AM_{(B-PW)l}(t)$, sendo que o melhor resultado foi obtido para $\theta = 0,2$ com respeito à normalidade, para o tamanho amostral 300. Quando considero a estatística $QM_{(B-PW)l}(t)$ e o tamanho amostral igual a 500, os resultados obtidos são equivalentes para os três θ's.

Como a estatística $AM_{(B-PW)l}(t)$ teve piores resultados com respeito à normalidade, repetimos o processo para $n = 600$ e para $\theta = 200,\ 10$ e 0,2. Desta forma, temos 1200 cromossomos com $n_1 = 300$, $n_2 = 100$ e $n_3 = 200$. Na Figura 5.10, vemos os gráficos Q-Q Normal da estatística do teste baseada nos desvios absolutos.

Para todos os valores de θ, a situação parece a mesma daquela obtida quando $n = 500$. Nas caudas temos um leve desvio da reta de distribuição Normal padrão.

Podemos aplicar o teste de Kolmogorov-Smirnov (KS) para verificar se as estatísticas

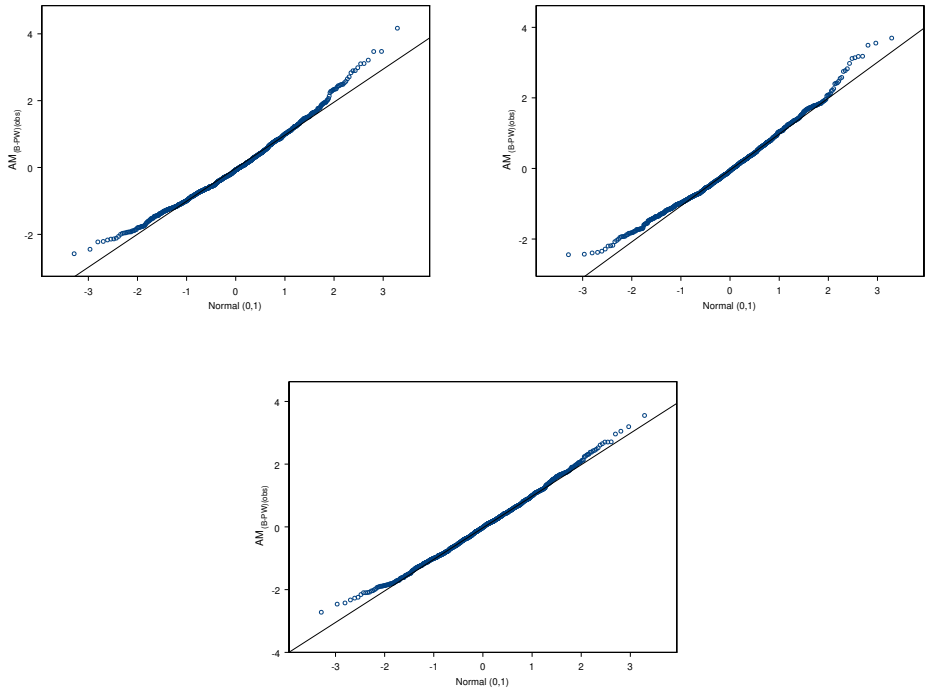

Figura 5.9: Q-Q Normal da estatística $AM_{(B-PW)l}(t)$ para $\theta = 0,2$ e tamanhos amostrais 300, 400 e 500, respectivamente.

do teste para amostras não balanceadas têm distribuição Normal padrão. A hipótese nula do teste de Kolmogorov-Smirnov é que as estatísticas têm distribuição Normal padrão e a alternativa é que não tem.

Seja x_i, $i = 1, \ldots,$ 1000 valores da estatística contida nos dados simulados. Para cada valor x_i, o teste de Kolmogorov-Smirnov compara a proporção de valores menores ou iguais a x_i com a proporção esperada de uma distribuição Normal padrão. A estatística do teste é o máximo da diferença sobre todos $\mathbf{x}$. Matematicamente, temos

$$\max_i(|F(x_i) - G(x_i)|),$$

em que $F(x_i)$ é a proporção de valores de X menores ou iguais a x_i e $G(x_i)$ é a função distribuição acumulada de uma Normal padrão avaliada em x_i.

A Tabela 5.1 mostra os resultados dos testes de Kolmogorov-Smirnov e os p-valores.

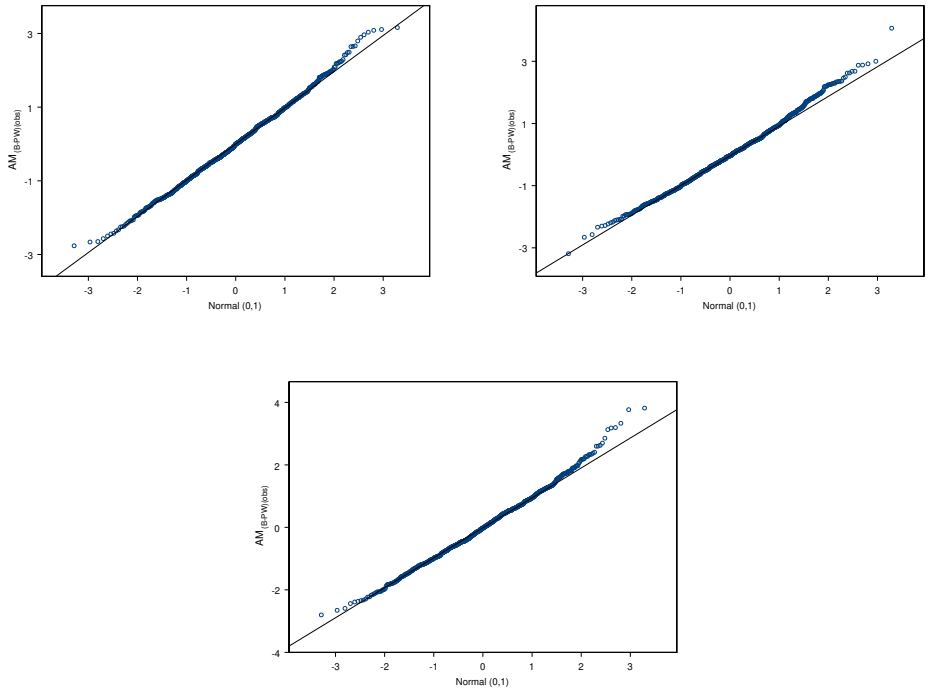

Figura 5.10: Q-Q Normal da estatística $AM_{(B-PW)l}(t)$ para $\theta = 0, 2, 10$ e 200 e tamanho amostral 600, respectivamente.

Para a estatística $QM_{(B-PW)l}(t)$ e para todos valores de θ, não rejeitamos H_0, ou seja, temos evidência de que os dados têm distribuição Normal padrão, a um nível de significância de 5%. No entanto, pelos gráficos Q-Q Normal, para $n = 300$ e $\theta = 10$ e 200 temos caudas levemente pesadas. Sendo assim, para $n > 300$ temos melhores resultados com respeito à normalidade.

Para a estatística $AM_{(B-PW)l}(t)$ temos uma situação diferente. Para $\theta = 0, 2$ não rejeitamos H_0, ou seja, temos evidência que os dados têm distribuição Normal padrão para todos os tamanhos amostrais, considerando um nível de significância em 5%. Para $\theta = 200$, rejeitamos H_0 para $n = 300$ e para $\theta = 10$, rejeitamos H_0 para $n = 400$. Avaliando os gráficos, vimos que para $n > 400$ temos melhores resultados quanto à normalidade.

Considere um tamanho efetivo populacional menor, ou seja, $N_e = 1000$. Construiremos populações com esse tamanho efetivo e simularemos 1000 amostras dessa população,

Tabela 5.1: Teste Kolmogorov-Smirnov para os dados simulados

Estatística	θ	n	Estatística KS	p-valor
$QM_{(B-PW)l}(t)$	200	300	0,0334	0,215
		400	0,0167	0,9443
		500	0,0181	0,899
	10	300	0,0274	0,439
		400	0,0342	0,193
		500	0,0252	0,548
	0,2	300	0.0172	0,928
		400	0,0224	0,698
		500	0,016	0,959
$AM_{(B-PW)l}(t)$	200	300	0,0543	0,006
		400	0,0293	0,356
		500	0,024	0,613
		600	0,0271	0,455
	10	300	0,0407	0,0728
		400	0,0451	0,0343
		500	0,0293	0,389
		600	0,0333	0,217
	0,2	300	0,0403	0,0782
		400	0,0357	0,155
		500	0,0191	0,859
		600	0,0181	0,899

para as taxas de mutação 0,01, 0,0005 e 10^{-5}, considerando amostras não balanceadas.

Para a taxa de mutação 10^{-5} não houve variação nos tamanhos alélicos e desta forma não foi possível calcular o valor da estatística do teste.

Para a taxa de mutação 0,01, temos $\theta = 40$. Fixemos o tamanho amostral em 600

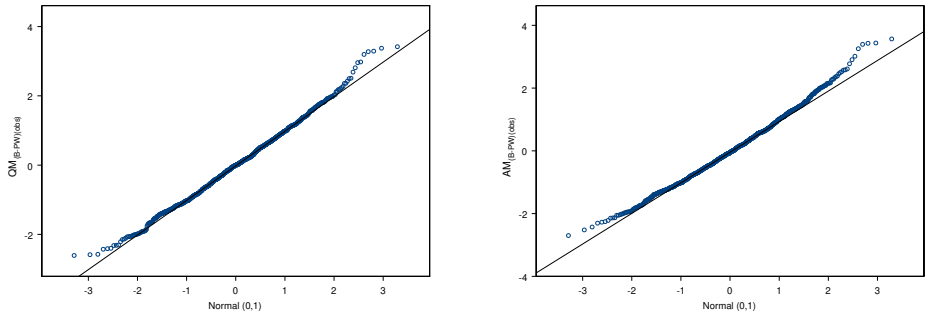

Figura 5.11: Q-Q Normal das estatísticas $QM_{(B-PW)l}(t)$ e $AM_{(B-PW)l}(t)$ para $\theta = 40$ e tamanho amostral 600, respectivamente.

com $n_1 = 300$, $n_2 = 100$ e $n_3 = 200$. A Figura 5.11 mostra os gráficos Q-Q Normal das estatísticas do teste baseadas nos desvios quadráticos e absolutos, respectivamente.

A Figura 5.12 mostra os gráficos Q-Q Normal das estatísticas baseadas nos desvios quadráticos e absolutos, respectivamente, para $\theta = 2$ (taxa de mutação $0,0005$).

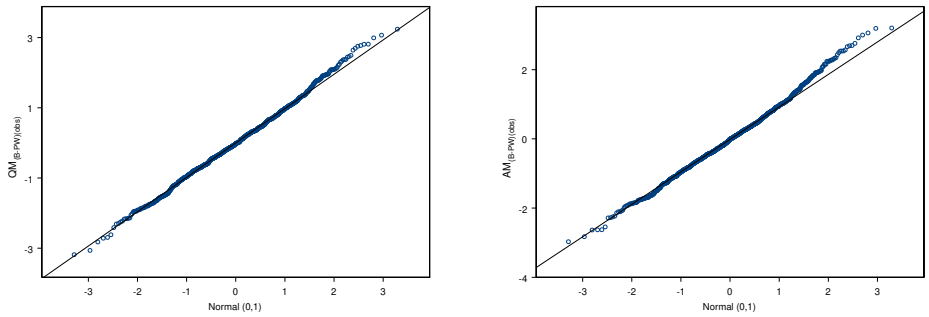

Figura 5.12: Q-Q Normal das estatísticas $QM_{(B-PW)l}(t)$ e $AM_{(B-PW)l}(t)$ para $\theta = 2$ e tamanho amostral 600, respectivamente.

Apesar de leves desvios nas caudas, não há porque não acreditar que as estatísticas $QM_{(B-PW)l}(t)$ e $AM_{(B-PW)l}(t)$ tenham distribuição Normal para $\theta = 2$ e 40.

Um fato interessante da simulação é que para $\theta = 0,2$ o resultado para normalidade é melhor do que para valores de θ maiores. Para esse valor de θ temos pouca variação

do tamanho alélico na população, pois só temos 3 tamanhos alélicos distintos. Como resultado, os desvios nas caudas da distribuição das estatísticas do teste (baseadas nos desvios absolutos e quadráticos) com relação a distribuição Normal padrão são menores quando comparados com os desvios obtidos para os outros θ's.

Outro resultado é o fato da estatística do teste baseada nos desvios quadráticos, $QM_{(B-PW)l}(t) = QM_{Bl}(t) - QMP_{Wl}(t)$, ter um comportamento melhor com relação à normalidade que a estatística baseada em desvios absolutos, $AM_{(B-PW)l}(t) = AM_{Bl}(t) - AMP_{Wl}(t)$. No entanto, esse resultado não é tão evidente quando considero amostras balanceadas, pois os comportamentos são semelhantes.

5.2 Aplicação a dados reais

Os dados, que foram analisados nesta dissertação de mestrado, tratam-se de um estudo feito pelo COGA (Collaborative Study on the Genetics of Alcoholism). Estes dados estão disponíveis através do GAW (Genetic Analysis Workshop) número 14 e podem ser obtidos pelo site www.niaaa.nih.gov/ResearchInformation/ExtramuralResearch/SharedResources /projcoga.htm (Mais informações sobre esses dados ver Edenberg et al. (2005)). O interesse deste estudo é aprender mais sobre como o alcoolismo é transmitido de geração para geração. Desta forma, a proposta é verificar características hereditárias que poderiam aumentar o risco de desenvolver o alcoolismo.

Mais de 2000 indivíduos participaram desse estudo, a partir de 366 famílias afetadas ou não pelo alcoolismo. A inclusão do indivíduo no estudo seguia certas regras como:

- não estar tomando nenhum medicamento que afete o funcionamento mental ou que seja prejudicial à saúde;
- não estar tomando álcool pelo menos 5 dias antes dos testes;
- não estar fazendo uso de drogas ilícitas pelo menos 5 dias antes dos testes.

Foram feitos vários testes com os pacientes selecionados, como testes cognitivos, teste de urina, teste de onda cerebral e teste do movimento dos olhos. Uma amostra de sangue foi colhida para a análise genética. O conteúdo genético foi retirado das células brancas do sangue.

Tabela 5.2: Freqüências étnicas dos indivíduos em estudo

etnia	freqüência	%
Branca não hispânica	1074	66,5
Negra não hispanica	191	11,9
Branca hispânica	78	4,9
Negra hispânica	14	0,9
Índio americano	12	0,7
Polinésios	4	0,2
Outro	12	0,7
Sem informação	229	14,2
Total	2000	

Para contextualização, destes 2000 indivíduos estudados, só há informação de 1614. Destes 1614, 788 eram do sexo feminino e 826 do sexo masculino. A média de idade dos indivíduos no estudo é aproximadamente 34 anos e um desvio padrão de 20.

Nosso principal interesse são os dados genéticos, que se apresentam com 23 pares de cromossomos homólogos, dos quais em cada cromossomo há uma análise de diversos marcadores para 1614 indivíduos. Desta forma, cada cromossomo apresenta um certo número de marcadores de microsatélites e em cada marcador estima-se a freqüência do tamanho alélico. No total são 388 *loci* de microsatélites que se tem informação neste estudo. Os marcadores baseados em microsatélites são altamente polimórficos e, segundo Valdes et al. (1993), têm se tornado o estudo principal para o desenvolvimento de mapas genéticos. Esses marcadores "marcam" a região com seqüência repetida (microsatélites). Biologicamente, os marcadores baseados em microsatélites identificam um único *locus* de

maneira aleatória (Griffiths et al. (2004)).

A informação sobre a etnia foi coletada e pode-se notar pela Tabela 5.2, que a maioria (66,5%) dos indivíduos são brancos não hispânicos.

Outra informação importante deste estudo é o número de indivíduos afetados pelo alcoolismo. Neste estudo foi utilizado um índice para classificar os indivíduos em quatro categorias, chamado de ALDX1. As categorias são: Afetado, não afetado (com alguns sintomas de alcoolismo), puramente não afetado e nunca bebeu. Na Tabela 5.3, podemos ver a distribuição de freqüência para este índice.

Tabela 5.3: Freqüências de ALDX1 dos indivíduos em estudo

ALDX1	freqüência	%
Afetado	643	39,8
Não afetado (com alguns sintomas)	431	26,7
Puramente não afetado	285	17,7
Nunca bebeu	29	1,8
Sem informação	226	14

No Capítulo 3 e 4 definimos duas estatísticas para aplicar o teste de homogeneidade, uma baseada nos desvios quadráticos e outra baseada nos desvios absolutos. Assim, podemos verificar se os grupos definidos pela etnia e pelo índice ALDX1 são ou não homogêneos em um determinado *locus*. Para isso, na etnia desconsideramos os grupos sem informação, polinésios e outros, por haver pouca informação sobre eles. Assim, têm-se 5 grupos: Branco não-hispânico, Negro não hispanico, Branco hispânico, Negro hispânico e Índio americano.

No caso do índice ALDX1, desconsideramos os grupos sem informação e juntaremos os grupos "Puramente não afetado" e "Nunca bebeu", pois entendemos que esses dois grupos não apresentam indivíduos afetados pelo alcoolismo. Desta forma, 3 grupos serão avaliados: Puramente não Afetado/Nunca bebeu, Não afetado (com alguns sintomas) e Afetado.

Para aplicar o teste de homogeneidade utilizando as estatísticas definidas nos Capítulos 3 e 4, estamos supondo que as população estão bem definidas e que não há migração entre elas. Apesar de se tratar de dados de família, vamos supor, também, que os indivíduos são independentes.

Podemos nos perguntar: Por que não utilizar a análise de variância para um critério de classificação (ANOVA)? O fato é que $Y_{igl}(t)$ ($i = 1, \ldots, 2n$, $g = 1, \ldots, G$ e $l = 1, \ldots, L$) é uma variável aleatória quantitativa discreta, então esse modelo não se aplica diretamente, pois a suposição de normalidade de $Y_{igl}(t)$ não é razoável. Segundo Valdes et al. (1993) a distribuição de $Y_{igl}(t)$ é irregular, podendo ser bimodal ou trimodal.

Outra consideração que deve ser feita previamente é que os *loci* considerados para o teste são aqueles que apresentam número de tamanhos alélicos distintos maiores que 8 e amplitude entre o valor mínimo e o máximo de tamanho alélico na população maior que 20 unidades de repetição. Isso reduz o número de *loci* que iremos analisar. Temos, então, 219 *loci* para aplicar o teste de homogeneidade.

Vamos utilizar o método Bootstrap para encontrar o nível de significância para cada *loci* dos testes de homogeneidade definidos para as estatísticas baseadas nos desvios quadráticos e absolutos. Sendo assim, na Seção 5.2.1 introduzimos esse método, enfatizando-o no propósito de encontrar o nível de significância.

Como temos 219 *loci*, temos 219 testes de hipóteses. Queremos obter uma visão global sobre o que acontece nos grupos levando em consideração esses 219 testes. Desta forma, os p-valores obtidos devem ser corrigidos para comparações múltiplas. Para isso, vamos assumir que os *loci* são independentes. Pode-se argumentar que para os *loci* de microsatélites nos mesmos cromossomos, a suposição de independência é muito forte. Essa suposição é considerada mais forte em regiões codantes cuja ordem de nucleotídeos é importante para a codificação de proteínas. Como 5% do genoma humano codificam proteínas, (Hartl 2000), e os *loci* de microsatélites estão dispersos aleatoriamente no genoma, poucos destes *loci* estarão em região codante. Além disso, a maneira de indentificar essas estruturas é aleatória. Com isso, a suposição de independência é razoável.

Na Seção 5.2.2, iremos introduzir um método chamado FDR ("*false discovery rate*") para correção dos p-valores quando temos testes múltiplos, Benjamini & Hochberg (1995). Em todas as análises rejeitamos H_0 a um nível de significância de 5%.

Na Seção 5.2.3, iremos aplicar a medida de distância baseada nos desvios quadráticos estudada no Capítulo 3 e na Seção 5.2.4, iremos aplicar a medida de distância baseada nos desvios absolutos, estudada no Capítulo 4.

5.2.1 Teste de hipótese com o método bootstrap

O Bootstrap foi introduzido em 1979 como um método computacional para estimar erro padrão de um estimador cuja a distribuição é desconhecida (Efron & Tibshirani (1993)). O método Bootstrap depende da noção de amostra Bootstrap. Considere $\hat{F}$ a distribuição empírica, seja $1/n$ a probabilidade de observar os valores x_i, $i = 1, 2, \ldots, n$, uma amostra aleatória cuja função de probabilidade é desconhecida (F) e cujo parâmetro θ desejamos estimar. A amostra Bootstrap é definida por uma amostra aleatória de tamanho n retirada de $\hat{F}$, ou seja

$$\hat{F} \longrightarrow (x_1^*, x_2^*, \ldots, x_n^*).$$

O vetor $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ representa a reamostra do vetor $\mathbf{x}$ observado na amostra. Ou seja, os dados $x_1^*, x_2^*, \ldots, x_n^*$ são amostras aleatórias de tamanho n retiradas com reposição da amostra $(x_1, x_2, \ldots, x_n)$ de uma população.

Suponha que temos duas populações e temos amostras $\mathbf{z}$ (de tamanho n) e $\mathbf{y}$ (de tamanho m) supostamente com distribuições diferentes F e F°. Suponha que o interesse é testar a seguinte hipótese nula $H_0 : F = F^\circ$. O teste de hipótese Bootstrap é baseado em uma estatística de teste, que denotamos por $t(\cdot)$. Procuramos encontrar o p-valor $= P_{H_0}[t(\mathbf{x}^*) \geq t(\mathbf{x})]$. A quantidade $t(\mathbf{x})$ é o valor observado da estatística do teste na amostra e $\mathbf{x}^*$ tem distribuição especificada sob H_0.

Considere o algoritmo para testar $F = F^\circ$.

1. Faça B amostras de tamanho $n + m$ com reposição de $\mathbf{x}$, em que $\mathbf{x}$ é o vetor com

os valores de $\mathbf{z}$ e $\mathbf{y}$ misturados. Chame as n primeiras observações de $\mathbf{z}^{*b}$ e os m restantes de $\mathbf{y}^{*b}$, para $b = 1, 2, \ldots, B$.

2. Calcule a estatística $t(\cdot)$ para cada amostra $b = 1, 2, \ldots, B$.

3. Calcule o p-valor$_{boot}$ da seguinte maneira

$$\text{p-valor}_{boot} = \frac{\sharp\{t(\mathbf{x}^{*b}) \geq t_{obs}\}}{B},$$

em que $t_{obs} = t(\mathbf{x})$ é o valor observado da estatística do teste.

5.2.2 Correção dos p-valores para comparações múltiplas

Suponha que temos o seguinte teste:

$$H_0 : \theta \in \Theta \quad \text{vs} \quad H_1 : \theta \notin \Theta,$$

baseado na estatística $T(\mathbf{X})$. Para uma dada região de rejeição Γ, rejeitamos H_0 quando $T(\mathbf{X}) \in \Gamma$. O erro tipo I ocorre quando $T(\mathbf{X}) \in \Gamma$, mas H_0 é verdadeira. O erro tipo II ocorre quando $T(\mathbf{X}) \notin \Gamma$, mas H_1 é verdadeira. Para escolha de Γ, fixamos o erro tipo I em α, ou seja, controlamos a probabilidade do erro tipo I para encontrar a região de rejeição.

Nos testes de hipótese múltiplos temos uma situação muito mais complicada que em testes simples, pois em cada teste temos erros tipo I e tipo II e desta forma, não fica claro como devemos medir a taxa de erro total. Suponha que, por exemplo, estejamos fazendo 10 testes de hipóteses com nível de significância 5% e assumimos que a distribuição dos p-valores para esses testes seja Uniforme (0,1) e que os testes sejam independentes. Assim, seja R a variável aleatória representando o número de testes significantes dentre 10 e seja $X_1, X_2, \ldots, X_{10}$ variáveis aleatórias independendentes e identicamente distribuídas Uniforme (0,1). A probabilidade de declarar um teste significante sob a hipótese nula é 0,05, mas a probabilidade de declarar pelo menos um teste significante dentre os 10 é

$$P(R \geq 1) \;=\; 1 - P(R = 0) = 1 - P\left(\bigcap_{i=1}^{10} [X_i > 0,05]\right)$$

$$= \quad 1 - \prod_{i=1}^{10} P(X_i > 0,05) = 1 - (0,95)^{10} = 0,401.$$

A primeira medida para a taxa de erro foi FWER ("*familywise error rate*"), que é a probabilidade de cometer um ou mais erros do tipo I dentre todas as hipóteses. Ao invés de fixar a probabilidade do erro tipo I para cada teste de hipótese no nível α, a medida FWER é controlada no nível α. Assim, α é escolhido de maneira que FWER$\leq \alpha$ e a região de rejeição Γ é obtida de maneira que o nível α é mantido.

Em um trabalho pioneiro, Benjamini & Hochberg (1995) introduziram uma medida de erro para testes de hipóteses múltiplos chamada "*false discovery rate*" (FDR). Essa quandidade é a proporção esperada de falsos positivos encontrada em todas as hipóteses rejeitadas (ou seja, a proporção esperada de rejeitar H_0 quando esta é verdadeira). Em várias situações o FWER é mais restrito, especialmente quando o número de testes é grande, o que não acontece com o FDR.

A Tabela 5.4 representa situações que podem ocorrer quando testamos m hipóteses. Seja R_0 a variável aleatória representando o número de resultados falsos positivos ou de erros tipo I. Assim, FWER é definido por $P(R_0 \geq 1)$. Em geral, quando o número de testes aumenta, o poder decresce quando controlamos o FWER. O FDR é definido por

$$\text{FDR} = E\left(\frac{R_0}{R}\middle| R > 0\right) P(R > 0),$$

em que R é a variável aleatória representando o número de testes rejeitados dentre todos os testes. Isto é, FDR é a proporção esperada de falsos positivos encontrada sobre todas as hipóteses rejeitadas vezes a probabilidade de se fazer pelo menos uma rejeição. Na prática, os valores de r e b, representados na Tabela 5.4, são observados e por isso estão assim representados.

Benjamini & Hochberg (1995) e Benjamini & Liu (1999) introduziram um método que fornece uma seqüência de p-valores controlando o FDR. Este método utiliza os dados observados, estima a região de rejeição de forma que na média FDR$\leq \alpha$ para algum α escolhido. Com isso, o método sequencial de p-valores nos permite fixar uma taxa de erro de antemão e estimar a região de rejeição.

Tabela 5.4: Testes múltiplos

Hipótese	não rejeita	rejeita	Total
H_0 é verdadeira	B_0	R_0	M_0
H_1 é verdadeira	B_1	R_1	M_1
	b	r	m

Para fazer a correção dos p-valores, utilizaremos o PROC MULTTEST do SAS system versão 9.1 controlando FDR com o nível de significância definido em 0,05.

5.2.3 Aplicação da medida de distância baseada nos desvios quadráticos

Primeiramente considere a medida de distância ponderada da relação (3.3.3), $QMP_{Wl(obs)}(t)$. Utilizaremos esta, pois o tamanho amostral de cada grupo é muito diferente. Adaptamos esta medida, da seguinte maneira

$$QMP_{Wl(obs)}(t) = \sum_{g=1}^{G} w_g \frac{2}{2n_g(2n_g-1)} \sum_{i=1}^{2n_g} \sum_{i'>i} (y_{igl}(t) - y_{i'gl}(t))^2 \, n_{y_{igl}(t)} n_{y_{i'gl}(t)},$$

em que $n_{y_{i'gl}(t)}$ é a freqüência do tamanho alélico $y_{i'gl}(t)$ $(i' = 1, \ldots, 2n_g,\ g = 1, \ldots, G)$ e w_g é a ponderação, ou seja, $w_g = 2n_g / \sum_{g=1}^{G} 2n_g$. Da mesma forma,

$$QM_{Bl(obs)}(t) = \frac{1}{\begin{pmatrix} G \\ 2 \end{pmatrix}} \sum_{g=1}^{G} \sum_{g'>g} \frac{1}{2n_g 2n_{g'}} \sum_{i=1}^{2n_g} \sum_{i'=1}^{2n_{g'}} (y_{igl}(t) - y_{i'g'l}(t))^2 \, n_{y_{igl}(t)} n_{y_{i'g'l}(t)},$$

em que $n_{y_{igl}(t)}$ e $n_{y_{i'g'l}(t)}$ são as freqüências dos tamanhos alélicos $y_{igl}(t)$ e $y_{i'g'l}(t)$, nas populações g e g', respectivamente.

Utilizando $QMP_{Wl}(t)$ reduzimos a informação contida em amostras de tamanhos menores, por serem menos informativas e aumentamos a informação das populações de amostras maiores.

Sob H_0, podemos encontrar a estimativa de $\theta = 4N_2\mu$, o número esperado de mutações por *locus* por geração, para o modelo de mutação de um passo, encontrando $QM_{Totl(obs)}$, supondo que não haja divisão populacional. Assim, suponha que $\sum_{g=1}^{G} n_g = N$,

$$QM_{Totl(obs)}(t) = \frac{1}{\begin{pmatrix} 2N \\ 2 \end{pmatrix}} \sum_{i=1}^{2N} \sum_{i'>i} (y_{il}(t) - y_{i'l}(t))^2 \, n_{y_{il}(t)} n_{y_{i'l}(t)}, \tag{5.2.1}$$

em que $n_{y_{il}(t)}$ e $n_{y_{i'l}(t)}$ são as freqüências dos tamanhos alélicos $y_{il}(t)$ e $y_{i'l}(t)$, desconsiderando os grupos.

Aplicaremos o método Bootstrap definido na Seção 5.2.1 para encontrar os p-valores dos 219 testes. Para cada *locus* faremos 1000 e 10000 reamostras Bootstrap, ou seja, $B = 1000$ e $B = 10000$.

Começamos com a análise avaliando a etnia. Seja $d_{obs} = QM_{B(obs)l} - QMP_{W(obs)l}$. Para cada *locus*, o algoritmo consiste em

1. Fazer $B = 1000$ e $B = 10000$ amostras dos 1369 indivíduos com reposição. Desta forma, teremos $\sum_{g=1}^{5} 2n_g = 2738$ tamanhos alélicos. Para $b = 1, 2, \ldots, B$, defina:
   - $n_1 = 1074$ indivíduos como pertencentes ao grupo Branco não-hispânico,
   - $n_2 = 191$ indivíduos como pertencentes ao grupo Negro não-hispânico,
   - $n_3 = 78$ indivíduos como pertencentes ao grupo Branco hispânico,
   - $n_4 = 14$ indivíduos como pertencentes ao grupo Negro hispânico e
   - $n_5 = 12$ indivíduos como pertencentes ao grupo Índio americano.
2. Calcule a estatística do teste $d^{*b} = QM^{*b}_{B(obs)l} - QMP^{*b}_{W(obs)l}$ para cada amostra $(b = 1, 2, \ldots, B)$.
3. Calcule o p-valor$_{boot}$ da seguinte maneira

$$\text{p-valor}_{boot} = \frac{\sharp\{d^{*b}) \geq d_{obs}\}}{B}.$$

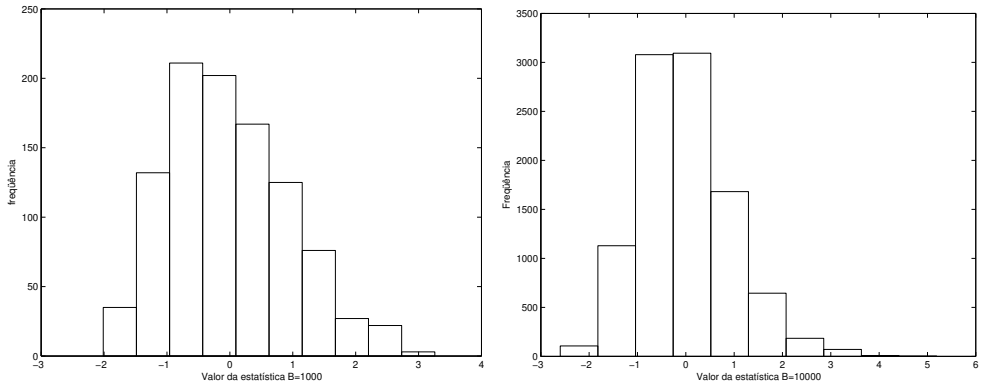

Figura 5.13: Distribuição da estatística do teste para o *locus* D4S1558 da análise de etnia (B=1000 e B=10000, respectivamente).

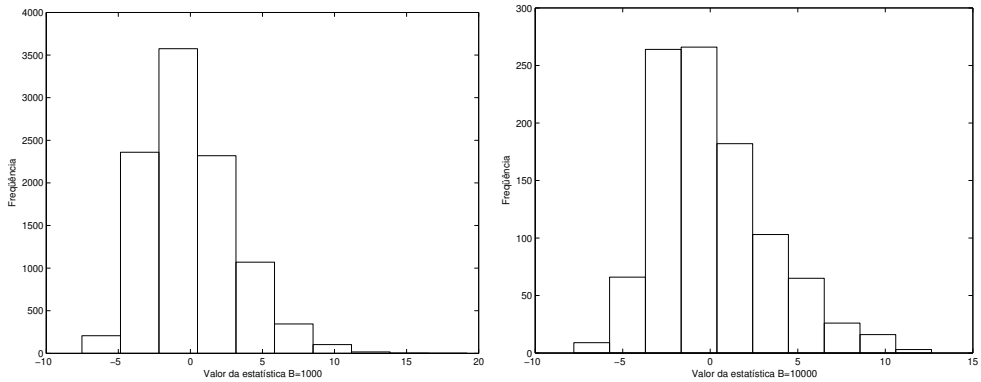

Figura 5.14: Distribuição da estatística do teste para o *locus* D2S2283 da análise de etnia (B=1000 e B=10000, respectivamente).

Pela Figura 5.13, vemos a distribuição da estatística do teste para amostras Bootstrap (B=1000 e B=10000, respectivamente) para o *locus* D4S1558, cujo p-valor, sem correção, deu significativo (p-valor<0,05). O valor da estatística observado na amostra é 3 e a estimativa de θ, sob H_0, dada pela relação (5.2.1), é 7,7. Para esse *locus* rejeitamos H_0, ou seja, existe evidência de diferenças significativas em pelo menos dois grupos de etnias (valor de p sem correção igual à 0,001 para B=10000). Analisando a distribuição dos tamanhos alélicos, Figura 5.15, e a matriz de distâncias, tem-se que as diferenças estão nas distribuições dos grupos negros não hispânicos e Negros hispânicos, comparadas com

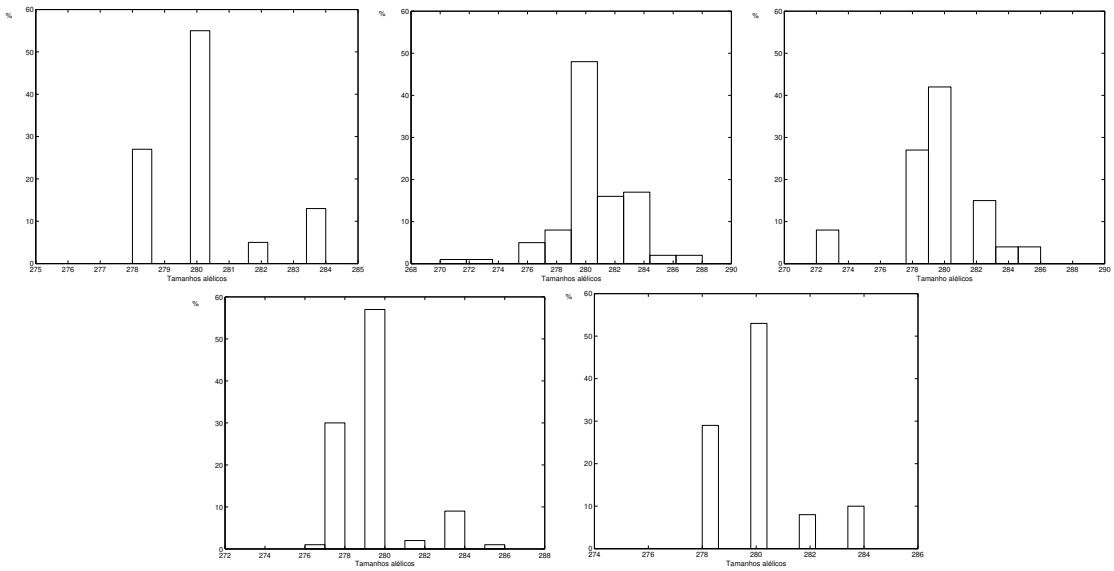

Figura 5.15: Distribuição dos tamanhos alélicos do *locus* D4S1558 para os grupos de etnia: índio americano, negros não hispânicos, negros hispânicos, brancos não hispânicos e brancos hispânicos, respectivamente.

a dos outros grupos. Apesar das distribuições por grupo terem a mesma moda, para esses dois grupos, elas se diferenciam principalmente pela dispersão.

Pela Figura 5.14, vemos a distribuição da estatística do teste para amostras Bootstrap (B=1000 e B=10000, respectivamente) para o *locus* D2S2283, cujo p-valor, sem correção, não deu significativo (p-valor>0,05). O valor da estatística observado na amostra é -5,43 e a estimativa de θ, sob H_0, é 27,98. Para esse *locus* não há evidência para rejeitar H_0, ou seja, não há evidência significativa de diferenças nas distribuições dos tamanhos alélicos nos 5 grupos de etnias (valor de p sem correção igual à 0,99 para B=10000). Assim, neste *locus* há evidência de homogeneidade populacional.

Vamos analisar somente os p-valores obtidos para $B = 10000$. Na Figura 5.16 podemos ver o histograma dos 219 p-valores e o gráfico Q-Q Uniforme comparando a distribuição dos 219 p-valores com a distribuição Uniforme em que $t = \dfrac{i}{219}, \quad i = 1, \ldots, 219$. Os p-valores estão mais concentrados no intervalo de 0 a 0,1. Pelo gráfico Q-Q Uniforme fica claro que a distribuição dos p-valores não se assemelha à distribuição Uniforme. Nos

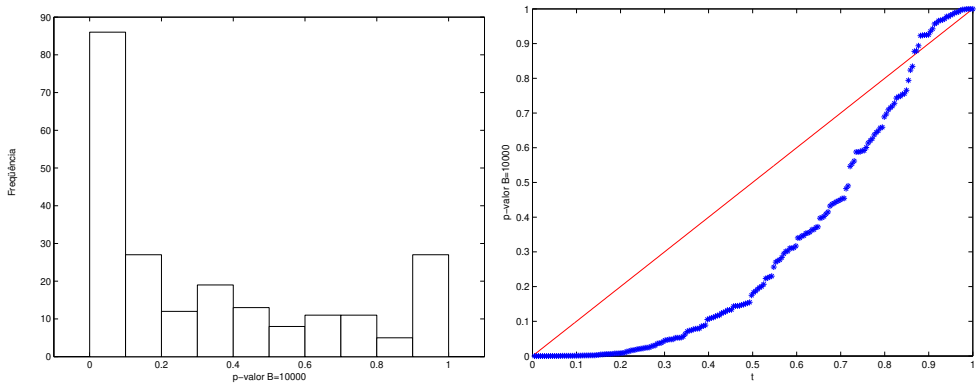

Figura 5.16: P-valores Bootstrap para 219 testes relacionados à etnia com $t = \frac{i}{219}$, $i = 1, \ldots, 219$.

valores maiores que 0,9 a distribuição está um pouco acima da distribuição Uniforme, ou seja, os p-valores obtidos são maiores que os esperados pela distribuição Uniforme. Nos valores menores que 0,9 a distribuição dos p-valores está bem abaixo da distribuição Uniforme, ou seja, os p-valores obtidos são menores que os esperados pela distribuição Uniforme.

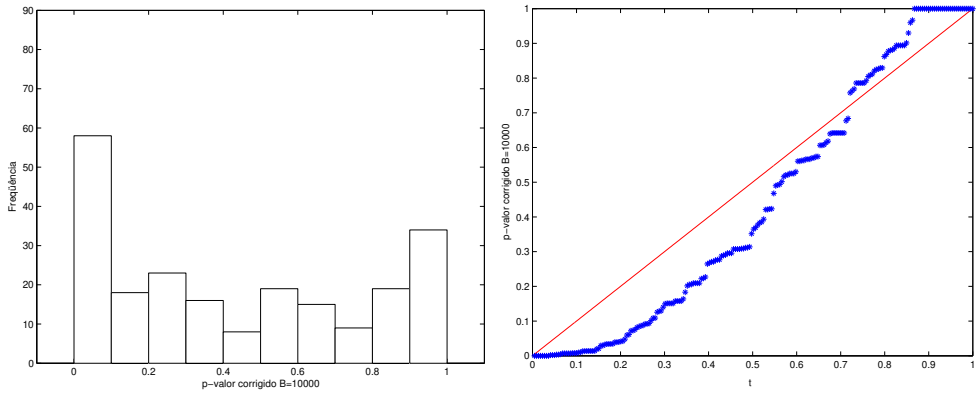

Figura 5.17: P-valores Bootstrap corrigidos para 219 testes relacionados à etnia com $t = \frac{i}{219}$, $i = 1, \ldots, 219$.

Como já foi discutido, devemos corrigir os p-valores para controlar o erro tipo I. A correção foi feita utilizando PROC MULTTEST do SAS system e controlando FDR.

A distribuição dos 219 p-valores corrigidos pode ser vista na Figura 5.17 e o gráfico Q-Q Uniforme comparando a distribuição dos 219 p-valores corrigidos com a distribuição Uniforme. O número de p-valores que estavam mais concentrados no intervalo de 0 a 0,1 diminuiu. A distribuição dos p-valores ficou mais próxima da distribuição Uniforme e além disso, no gráfico Q-Q Uniforme podemos ver que dimimuiram os valores que se encontravam abaixo da distribuição Uniforme. Nos valores maiores que 0,7 a distribuição está um pouco acima da distribuição Uniforme, ou seja, aumentou o número de p-valores maiores que os esperados pela distribuição Uniforme. E abaixo de 0,7, a distribuição dos p-valores está abaixo da distribuição Uniforme, mas não tão abaixo como antes da aplicação da correção.

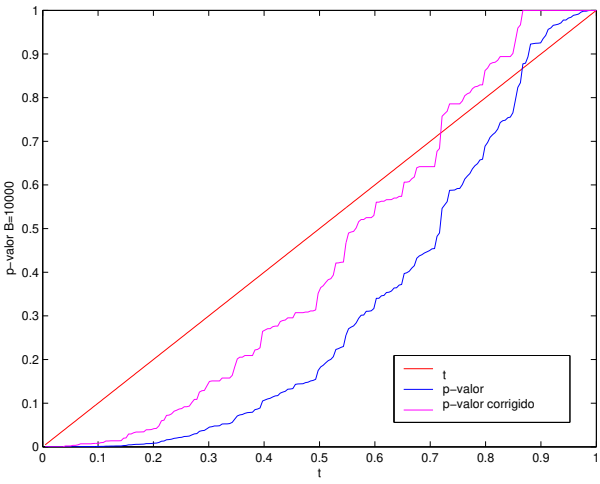

Figura 5.18: Comparação dos p-valores da estatística do teste baseada nos desvios quadráticos relacionados à etnia com $t = \dfrac{i}{219}, \quad i = 1, \ldots, 219$.

A comparação das distribuições dos p-valores (sem correção e corrigidos) versus a distribuição Uniforme pode ser vista na Figura 5.18. Realmente a distribuição dos p-valores corrigidos está acima da distribuição dos p-valores sem correção e a distribuição dos p-valores corrigidos se aproxima mais da distribuição Uniforme. Considerando um nível de significância de 5%, para $B = 1000$, temos 72 testes significativos dos 219, ou seja, 72 *loci* apresentaram evidência de diferenças significativas em pelo menos dois grupos

e para $B = 10000$, 70 testes significativos. Com a correção, para $B = 1000$, em 45 testes ou *loci* rejeitamos H_0, ou seja, encontramos evidência de diferenças significativas em pelo menos dois grupos de etnias. Para $B = 10000$, em 46 testes ou *loci* encontramos evidência para rejeitar H_0.

Para verificar em que grupos encontramos diferenças significativas, vamos analisar a forma da distribuição dos tamanhos alélicos entre grupos quanto às suas média, moda e dispersão e a matriz de distâncias. Considerando os p-valores corrigidos, pela Tabela 5.5, temos o número de *loci* e quais diferenças encontradas entre os grupos (NH=não hispânicos, H=hispânicos, IA=índio americano e $\sim$ significa que a distribuição de um grupo é semelhante à do outro grupo). Na maioria dos testes encontramos diferenças nas distribuições de todos os grupos. Em alguns casos encontramos semelhanças entre negros hispânicos e não hispânicos e como também entre brancos hispânicos e não hispânicos. Em dois *loci* encontramos semelhanças entre hispânicos brancos e negros e como também entre não hispânicos brancos e negros.

No Capítulo 3, encontramos a distribuição da estatística do teste e discutimos que podemos encontrar a variância sob H_0. Desta forma, podemos aplicá-la e encontrar os 219 p-valores e também corrigí-los para comparações múltiplas. A variância da estatística é encontrada sob H_0. Utilizaremos a estimativa de θ sob H_0 e a variância é dada pela relação (3.5.12). Para $G = 5$, temos

$$\lambda_0^2 = \frac{(4\theta^2+\theta)}{50}\sum_{g=1}^{5}\left(\frac{1}{2n_g}+\frac{1}{2n_{g'}}\right) \quad \text{e} \quad \varsigma_0 \;=\; \frac{3}{25}\left\{\sum_{g=1}^{5}\frac{\theta^2}{n_g}+\frac{\theta}{4n_g}\right\}.$$

Podemos ver, na Figura 5.19, o histograma dos 219 p-valores obtidos pela distribuição assintótica e o gráfico Q-Q Uniforme comparando a distribuição dos 219 p-valores com a distribuição Uniforme. A distribuição dos p-valores fica abaixo da distribuição Uniforme. Os p-valores estão mais concentrados no intervalo de 0 a 0,2 e de 0,4 a 0,5.

Na Figura 5.20 podemos ver a distribuição dos p-valores corrigidos e a sua comparação com a distribuição dos p-valores antes da correção. No intervalo de 0,15 a 0,7 os p-valores ficaram acima da distribuição Uniforme, indicando que aumentaram o valor dos p-valores

Tabela 5.5: Diferenças entre grupos de etnias para a estatística baseada nos desvios quadráticos

Grupos	n^o de *loci*
Negros NH $\neq$ Negros H $\neq$ Brancos NH $\neq$ Brancos H $\neq$ IA	11
Negros NH $\sim$ Negros H $\neq$ Brancos NH $\sim$ Brancos H $\sim$ IA	5
Negros NH $\sim$ Negros H $\neq$ Brancos NH $\neq$ Brancos H $\neq$ IA	4
Negros NH $\sim$ Negros H $\sim$ AI $\neq$ Brancos NH $\sim$ Brancos H	4
Negros NH $\neq$ Negros H $\sim$ Brancos NH $\sim$ Brancos H $\sim$ IA	4
Negros NH $\neq$ Negros H $\neq$ Brancos NH $\neq$ Brancos H $\sim$ IA	3
Negros H $\neq$ Negros NH $\sim$ Brancos NH $\sim$ Brancos H $\sim$ IA	3
Negros NH $\sim$ AI $\neq$ Negros H $\sim$ Brancos NH $\sim$ Brancos H	2
Negros H $\neq$ Negros NH $\neq$ Brancos NH $\sim$ Brancos H $\sim$ IA	2
Negros H $\sim$ Brancos H $\neq$ Negros NH $\sim$ Brancos NH $\neq$ IA	2
Negros NH $\neq$ Negros H $\sim$ Brancos H $\neq$ Brancos NH $\sim$ IA	1
Negros NH $\sim$ Brancos NH $\neq$ Negros H $\sim$ Brancos H $\sim$ IA	1
Negros NH $\sim$ AI $\neq$ Negros H $\neq$ Brancos NH $\sim$ Brancos H	1
Negros NH $\sim$ Negros H $\neq$ Brancos NH $\sim$ Brancos H $\neq$ IA	1
Negros NH $\neq$ Negros H $\neq$ Brancos NH $\sim$ Brancos H $\neq$ IA	1
Negros NH $\neq$ Negros H $\sim$ AI $\neq$ Brancos NH $\sim$ Brancos H	1

corrigidos, pois antes estes estavam todos abaixos da distribuição Uniforme.

Considerando um nível de significância de 5%, temos 38 testes significativos dos 219, ou seja, 38 *loci* apresentaram evidência de diferenças significativas em pelo menos dois grupos. Com a correção, em 18 testes ou *loci* rejeitamos H_0, ou seja, encontramos evidência de diferenças significativas em pelo menos dois grupos de etnias. Apesar de encontrar menos *loci* significativos comparado com o método Bootstrap, apenas 2 *loci* que encontramos evidência de diferenças significativas utilizando a distribuição assintótica não apresenta diferença significativa pelo método Bootstrap, os *loci* são D5S1473 e D6S1052.

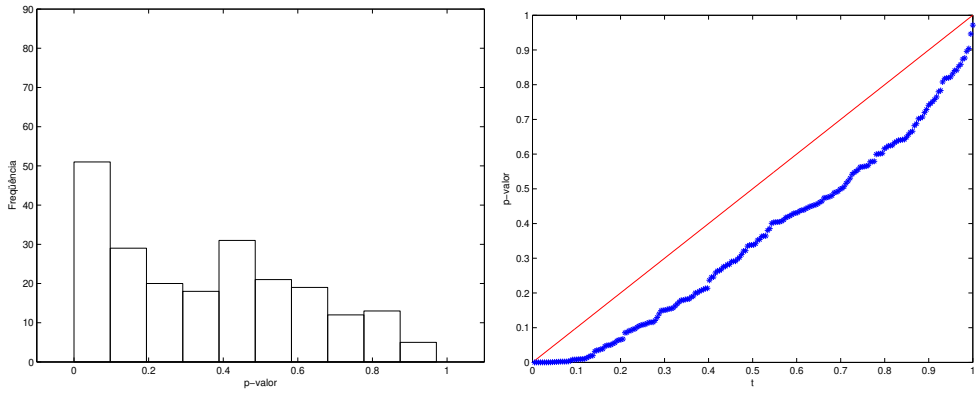

Figura 5.19: P-valores obtidos utilizando a distribuição assintótica relacionados à etnia com $t = \dfrac{i}{219}$, $i = 1, \ldots, 219$.

No *locus* D5S1473, as distribuições dos tamanhos alélicos apresentam diferenças entre todos os grupos. Na Seção 5.3 discutiremos mais sobre isso. No *locus* D6S1052, temos que os grupos Negros (hispânicos e não hispânicos) apresentam diferenças quando comparados com os outros grupos de etnia, no entanto eles são semelhantes entre si.

Considerando o índice ALDX1 e o mesmo algoritmo do método Bootstrap utilizado para a etnia, com as seguintes modificações:

1. são 1388 indivíduos, ou seja, $\sum_{g=1}^{5} 2n_g = 2776$ tamanhos alélicos e,

2. $n_1 = 643$ indivíduos como pertencentes ao grupo Afetado, $n_2 = 431$ indivíduos como pertencentes ao grupo Não afetado (com alguns sintomas) e $n_3 = 314$ indivíduos como pertencentes ao grupo Puramente não afetado e Nunca bebeu.

Pela Figura 5.21, vemos a distribuição da estatística do teste para amostras Bootstrap (B=1000 e B=10000, respectivamente) para o *locus* D5S1473, cujo p-valor, sem correção, deu significativo. O valor da estatística observado na amostra é 21,39 e a estimativa de θ sob H_0 é 92,03. Para esse *locus* rejeitamos H_0, ou seja, existe evidência de diferença significativa em pelo menos dois grupos (valor de p sem correção igual à 0,0002 para B=10000). Mais adiante, identificamos os grupos que apresentam essas diferenças através

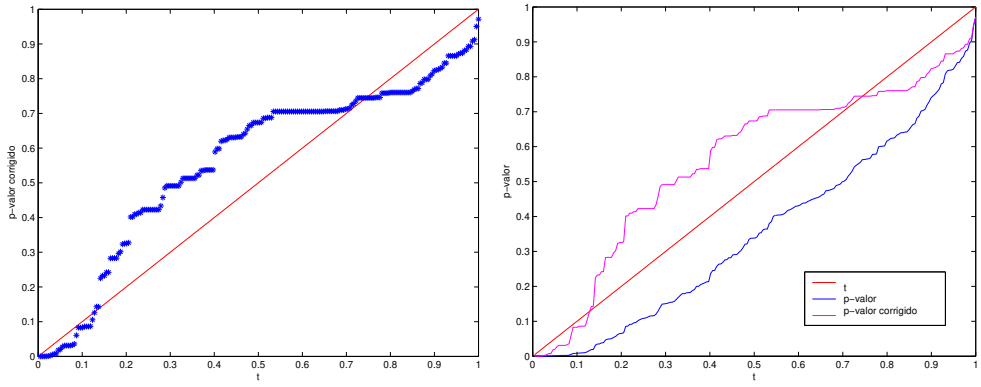

Figura 5.20: P-valores corrigidos e a comparação entre os p-valores relacionados à etnia com $t = \dfrac{i}{219}$, $i = 1, \ldots, 219$, respectivamente.

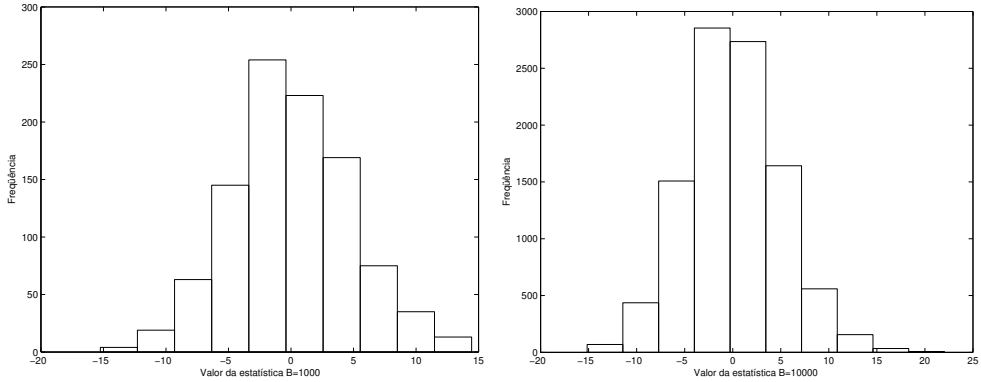

Figura 5.21: Distribuição da estatística do teste para o *locus* D5S1473 da análise de ALDX1 (B=1000 e B=10000, respectivamente).

da análise empírica da distribuição dos tamanhos alélicos.

Para o *locus* D20S448, cujo p-valor, sem correção, não deu significativo, temos a distribuição da estatística do teste para amostra Bootstrap (B=1000 e B10000, respectivamente) na Figura 5.22. O valor da estatística observado na amostra é -0,05 e a estimativa de θ sob H_0 é 76,82. Para esse *locus* não há evidência para rejeitar H_0, ou seja, não há evidência significativa de diferenças nas distribuições dos tamanhos alélicos nos 3 grupos. Assim, neste *locus* há evidência de homogeneidade populacional (valor de p sem correção igual à 0,5121 para B=10000).

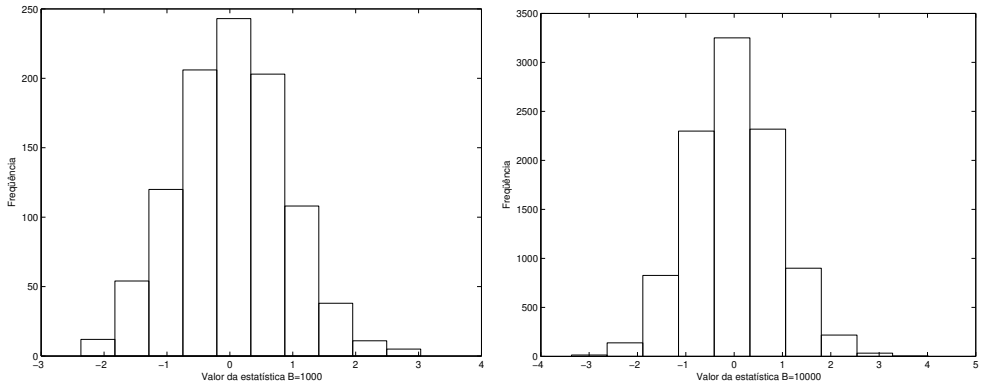

Figura 5.22: Distribuição da estatística do teste para o *locus* D20S448 da análise de ALDX1 (B=1000 e B=10000, respectivamente).

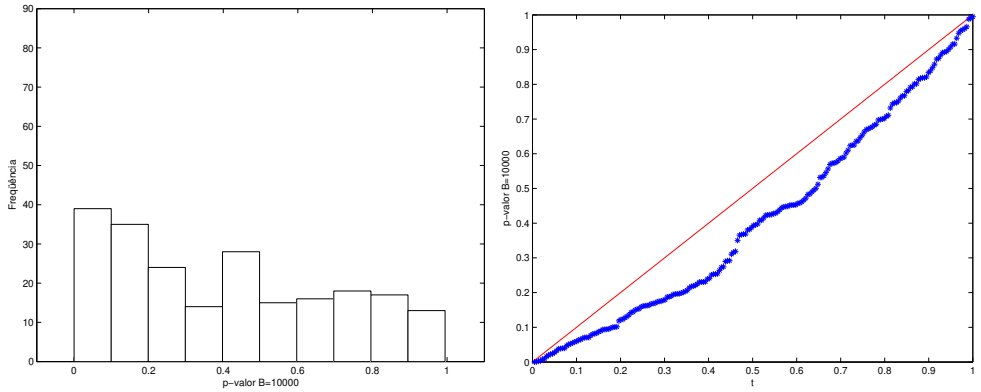

Figura 5.23: P-valores Bootstrap para 219 testes relacionados à ALDX1 com $t = \dfrac{i}{219}$, $i = 1, \ldots, 219$.

Conforme fizemos para analisar a etnia, vamos estudar somente os resultados obtidos para $B = 10000$. Na Figura 5.23 podemos ver os histogramas dos 219 p-valores e o gráfico Q-Q Uniforme comparando a distribuição dos 219 p-valores com a distribuição Uniforme. Os p-valores estão ligeiramente mais concentrados no intervalo de 0 a 0,3 e de 0,4 a 0,5. A distribuição dos p-valores está toda abaixo da distribuição Uniforme, ou seja, os p-valores obtidos são menores que os esperados pela distribuição Uniforme.

Após a correção, vemos pela Figura 5.24, que a distribuição dos p-valores ficou mais distante da distribuição Uniforme. Podemos ver, também, que a distribuição dos p-valores

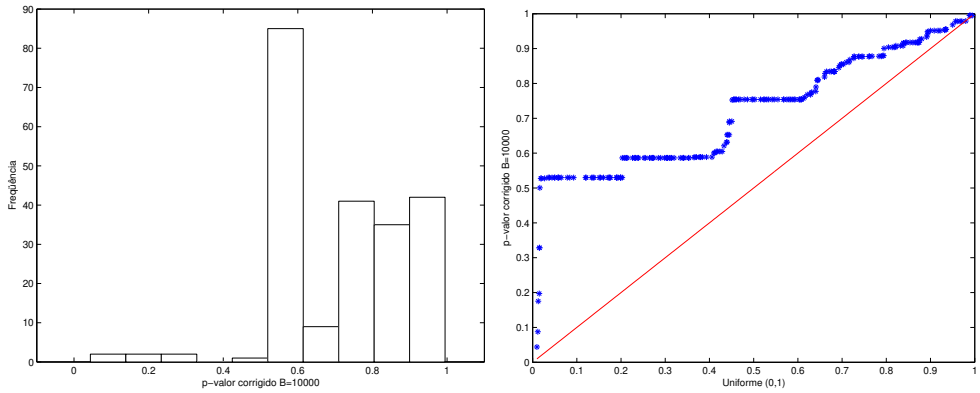

Figura 5.24: P-valores Bootstrap corrigidos para 219 testes relacionados à ALDX1 com $t = \dfrac{i}{219}, \quad i = 1, \ldots, 219.$

corrigidos está toda acima da distribuição Uniforme. Assim, os p-valores corrigidos são maiores que os esperados pela distribuição Uniforme. Pelo histograma, vemos que os p-valores corrigidos estão mais concentrados nos intervalos de 0,5 a 0,6 e acima dos valores 0,7.

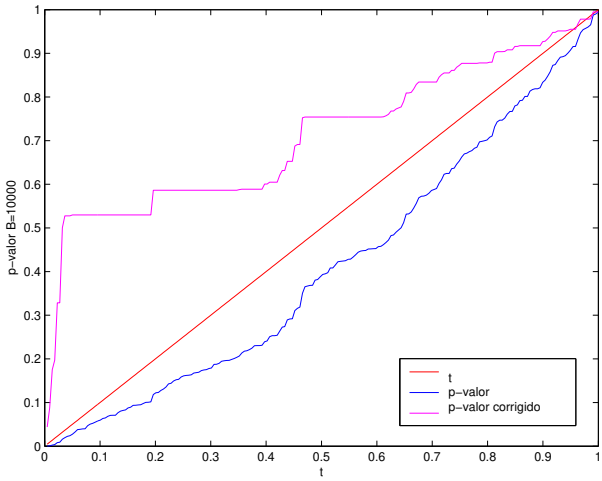

Figura 5.25: Comparação dos p-valores da estatística do teste baseada nos desvios quadráticos relacionados à ALDX1 com $t = \dfrac{i}{219}, \quad i = 1, \ldots, 219.$

A Figura 5.25 mostra a comparação das distribuições dos p-valores (sem correção e

corrigidos) versus a distribuição Uniforme. A distribuição dos p-valores corrigidos está acima da distribuição dos p-valores sem correção, que se aproxima mais da distribuição Uniforme, enquanto que com a correção a distribuição se distancia completamente da distribuição Uniforme. A distribuição mudou completamente, estava abaixo da linha da distribuição Uniforme e passou para cima.

Considerando um nível de significância de 5%, para $B = 1000$, temos 19 testes significativos dos 219 e para $B = 10000$, 17 testes significativos. Com a correção, para $B = 1000$, em 1 teste ou *locus* rejeitamos H_0, ou seja, encontramos evidência de diferenças significativas em pelo menos dois grupos. Da mesma forma, para $B = 10000$, em 1 teste ou *locus* rejeitamos H_0. O único *locus* que deu significativo foi D5S1473.

Analisando as distribuições dos tamanhos alélicos por grupos e suas respectivas matrizes de distância, temos diferenças em todos os grupos: Puramente não afetado/Nunca bebeu, afetado com alguns sintomas e afetado (mais detalhes ver Seção 5.3).

Da mesma forma que foi feito para etnia, podemos aplicar o teste utilizando a distribuição assintótica e encontrar os 219 p-valores e também corrigí-los para comparações múltiplas. Para $G = 3$, a estimativa da variância é dada pela relação (3.5.12), em que seus componentes podem ser encontrados por

$$\lambda_0^2 = \frac{(4\theta^2 + \theta)}{9} \sum_{g=1}^{3} \left(\frac{1}{2n_g} + \frac{1}{2n_{g'}} \right) \quad \text{e} \quad \varsigma_0 = \frac{2}{9} \left\{ \sum_{g=1}^{3} \frac{\theta^2}{n_g} + \frac{\theta}{4n_g} \right\}.$$

Na Figura 5.26 podemos ver o histograma dos 219 p-valores obtidos pela distribuição assintótica e o gráfico Q-Q Uniforme comparando a distribuição dos 219 p-valores com a distribuição Uniforme. Os valores maiores que 0,4 a distribuição está abaixo da distribuição Uniforme, ou seja, os p-valores obtidos são menores que os esperados pela distribuição Uniforme. Abaixo de 0,4, a distribuição dos p-valores está acima da distribuição Uniforme, ou seja, os p-valores obtidos são maiores que os esperados pela distribuição Uniforme.

A distribuição dos p-valores corrigidos pode ser vista na Figura 5.27, como também a sua comparação com a distribuição dos p-valores antes da correção. Após a correção,

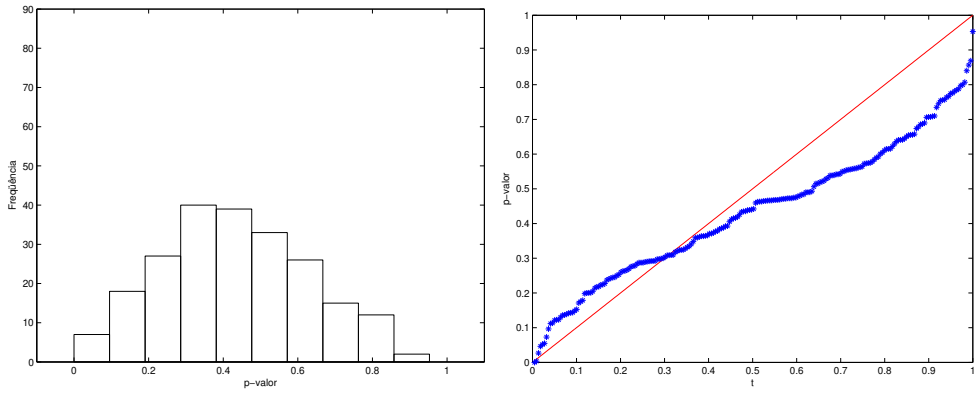

Figura 5.26: P-valores obtidos utilizando a distribuição assintótica relacionados à ALDX1 com $t = \dfrac{i}{219}, \quad i = 1, \ldots, 219$.

no intervalo de 0 a 0,7 os p-valores ficaram acima da distribuição Uniforme, indicando que aumentaram o valor dos p-valores corrigidos. Além disso, antes da correção, a distribuição dos p-valores cruzava a reta da distribuição Uniforme em 0,3. Com a correção, a distribuição cruza em 0,7. Excluindo valores perto do 1, os p-valores corrigidos são maiores que os p-valores sem correção.

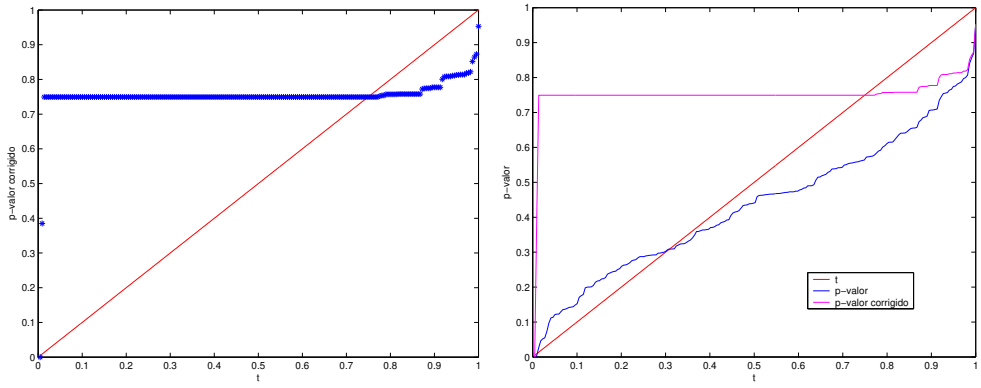

Figura 5.27: P-valores corrigidos e a comparação entre os p-valores relacionados à ALDX1 com $t = \dfrac{i}{219}, \quad i = 1, \ldots, 219$, respectivamente.

Considerando um nível de significância de 5%, antes da correção, temos 4 testes significativos dos 219. Com a correção, em apenas 1 teste ou *locus* rejeitamos H_0, ou seja,

encontramos evidência de diferenças significativas em pelo menos dois grupos. O único *locus* que deu significativo foi D5S1473, que é o mesmo obtido pelo método Bootstrap.

5.2.4 Aplicação da medida de distância baseada nos desvios absolutos

Da mesma maneira que a Seção anterior, considere a medida de distância ponderada definida pela relação (4.2.3), $AMP_{Wl(obs)}(t)$. Adaptamos esta medida, da seguinte maneira

$$AMP_{Wl(obs)}(t) = \sum_{g} w_g \frac{2}{2n_g(2n_g-1)} \sum_{i=1}^{2n_g} \sum_{i'>i} |y_{igl}(t) - y_{i'gl}(t)|\ n_{y_{igl}(t)} n_{y_{i'gl}(t)},$$

em que $n_{y_{i'gl}(t)}$ é a freqüência do tamanho alélico $y_{i'gl}(t)$ e w_g é a ponderação, ou seja, $w_g = 2n_g / \sum_{g=1}^{G} 2n_g$. Da mesma forma,

$$AM_{Bl(obs)}(t) = \frac{1}{\begin{pmatrix} G \\ 2 \end{pmatrix}} \sum_{g=1}^{G} \sum_{g'>g} \frac{1}{2n_g 2n_{g'}} \sum_{i=1}^{2n_g} \sum_{i'=1}^{2n_{g'}} |y_{igl}(t) - y_{i'g'l}(t)|\ n_{y_{igl}(t)} n_{y_{i'g'l}(t)},$$

em que $n_{y_{igl}(t)}$ e $n_{y_{i'g'l}(t)}$ são as freqüências dos tamanhos alélicos $y_{igl}(t)$ e $y_{i'g'l}(t)$, nas populações g e g' respectivamente.

Sob H_0, podemos encontrar a estimativa de $f^*(\theta) = \dfrac{\theta}{(1+2\theta)^{\frac{1}{2}}}$ encontrando $AM_{Totl(obs)}$ e supondo que não haja divisão populacional. Assim, suponha que $\sum_{g=1}^{G} n_g = N$,

$$AM_{Totl(obs)}(t) = \frac{1}{\begin{pmatrix} 2N \\ 2 \end{pmatrix}} \sum_{i=1}^{2N} \sum_{i'>i} |y_{il}(t) - y_{i'l}(t)|\ n_{y_{il}(t)} n_{y_{i'l}(t)}, \tag{5.2.2}$$

em que $n_{y_{il}(t)}$ e $n_{y_{i'l}(t)}$ são as freqüências dos tamanhos alélicos $y_{il}(t)$ e $y_{i'l}(t)$, desconsiderando os grupos.

Aplicamos também os algoritmos para etnia e ALDX1 definidos na Seção anterior para encontrar os p-valores dos 219 testes. Para cada *locus* faremos 10000 reamostras Bootstrap, ou seja, $B = 10000$. Além disso, temos que $d_{obs} = AM_{B(obs)l} - AMP_{W(obs)l}$.

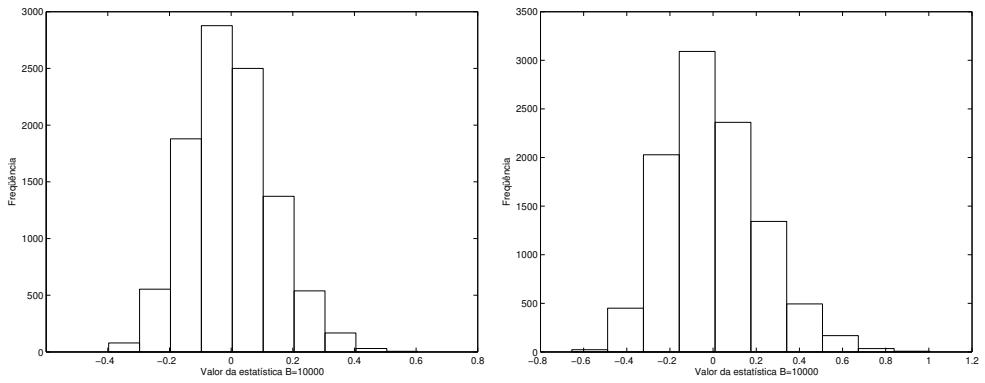

Figura 5.28: Distribuição da estatística do teste para o *locus* D4S1558 e D2S2283 da análise de etnia, respectivamente).

Primeiramente, consideremos a etnia. Pela Figura 5.28, vemos as distribuições da estatística do teste para amostra Bootstrap para o *locus* D4S1558, cujo p-valor, sem correção, deu significativo e para o *locus* D2S2283, cujo p-valor, sem correção, não deu significativo. Para o *locus* D4S1558, o valor da estatística observado na amostra é 0,42 e a estimativa de $f^*(\theta)$ sob H_0, dada pela relação (5.2.2), é 1,91. Para esse *locus* rejeitamos H_0, ou seja, existe evidência de diferenças significativas em pelo menos dois grupos de etnias (valor de p sem correção igual à 0,003). Neste mesmo *locus*, diferenças significativas foram encontradas no teste pela estatística baseada nos desvios quadráticos e na Figura 5.15 temos a distribuição dos tamanhos alélicos por grupos.

Para o *locus* D2S2283, o valor da estatística observado na amostra é -0,33 e a estimativa de $f^*(\theta)$ sob H_0 é 4,09. Para esse *locus* não há evidência para rejeitar H_0, ou seja, não há evidência significativa de diferenças nas distribuições dos tamanhos alélicos nos 5 grupos de populações (valor de p sem correção igual à 0,9607).

Na Figura 5.29 podemos ver a comparação das distribuições dos p-valores (sem correção e corrigidos) versus a distribuição Uniforme, em que $t = \dfrac{i}{219}$, $i = 1, \ldots, 219$. A distribuição dos p-valores corrigidos se assemelha àquela obtida pelos p-valores sem a correção. Em ambos os casos, a distribuição está praticamente abaixo da distribuição Uniforme. No entanto, a distribuição dos p-valores corrigidos está acima da distribuição

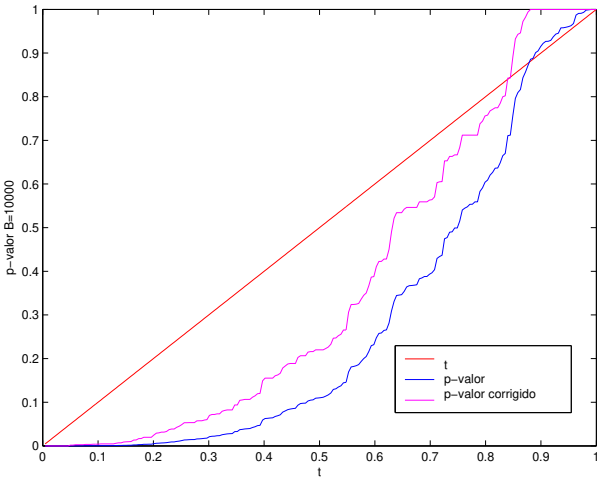

Figura 5.29: Comparação dos p-valores para 219 testes da estatística do teste baseada nos desvios absolutos relacionados à etnia com $t = \dfrac{i}{219}$ $i = 1, \ldots, 219$.

dos p-valores sem correção.

Considerando um nível de significância de 5%, temos 86 testes significativos dos 219. Com a correção, em 55 testes ou *loci* rejeitamos H_0. Comparando o resultado obtido pela estatística baseada nos desvios quadráticos, tivemos mais testes significativos, mas foram os mesmos *loci*. Da mesma maneira que fizemos na Seção anterior, considere os p-valores corrigidos. Pela Tabela 5.6, pode-se ver que na maioria dos testes encontramos diferenças nas distribuições de todos grupos e aumentou o número de *loci* com diferenças entre brancos e negros. Nota-se que em vários *loci* temos semelhanças entre o índio americano e o branco hispânico.

Agora, consideremos os grupos definidos pelo índice ALDX1. Pela Figura 5.30, vemos as distribuições da estatística do teste para amostra Bootstrap para o *locus* D5S1473, cujo p-valor, sem correção, deu significativo e para o *locus* D20S448, cujo p-valor, sem correção, não deu significativo. Para o *locus* D5S1473, o valor da estatística observado na amostra é 0,36 e a estimativa de $f^*(\theta)$ sob H_0, dada pela relação (5.2.2), é 5,65. Para esse *locus* rejeitamos H_0, ou seja, existe evidência de diferenças significativas em pelo menos dois grupos (valor de p sem correção igual à 0,0001). Nesse mesmo *locus* encontramos o

Tabela 5.6: Diferenças entre grupos de etnias para a estatística baseada nos desvios absolutos

Grupos	n^o de *loci*
Negros NH $\neq$ Negros H $\neq$ Brancos NH $\neq$ Brancos H $\neq$ IA	15
Negros NH $\sim$ Negros H $\neq$ Brancos NH $\sim$ Brancos H $\sim$ IA	6
Negros NH $\sim$ Negros H $\neq$ Brancos NH $\neq$ Brancos H $\neq$ IA	5
Negros NH $\sim$ Negros H $\sim$ AI $\neq$ Brancos NH $\sim$ Brancos H	4
Negros NH $\neq$ Negros H $\sim$ Brancos NH $\sim$ Brancos H $\sim$ IA	4
Negros NH $\neq$ Negros H $\neq$ Brancos NH $\neq$ Brancos H $\sim$ IA	3
Negros H $\neq$ Negros NH $\sim$ Brancos NH $\sim$ Brancos H $\sim$ IA	4
Negros NH $\sim$ AI $\neq$ Negros H $\sim$ Brancos NH $\sim$ Brancos H	2
Negros H $\neq$ Negros NH $\neq$ Brancos NH $\sim$ Brancos H $\sim$ IA	2
Negros H $\sim$ Brancos H $\neq$ Negros NH $\sim$ Brancos NH $\neq$ IA	3
Negros NH $\neq$ Negros H $\sim$ Brancos H $\neq$ Brancos NH $\sim$ IA	1
Negros NH $\sim$ Brancos NH $\neq$ Negros H $\sim$ Brancos H $\sim$ IA	1
Negros NH $\sim$ AI $\neq$ Negros H $\neq$ Brancos NH $\sim$ Brancos H	1
Negros NH $\sim$ Negros H $\neq$ Brancos NH $\sim$ Brancos H $\neq$ IA	3
Negros NH $\neq$ Negros H $\neq$ Brancos NH $\sim$ Brancos H $\neq$ IA	1
Negros NH $\neq$ Negros H $\sim$ AI $\neq$ Brancos NH $\sim$ Brancos H	1

teste significativo pela estatística baseada nos desvios quadráticos.

Para o *locus* D20S448, o valor da estatística observado na amostra é 0,015 e a estimativa de $f^*(\theta)$ sob H_0 é 6,7. Para esse *locus* não há evidência para rejeitar H_0, ou seja, não há evidência significativa de diferenças nas distribuições dos tamanhos alélicos nos 3 grupos (valor de p sem correção igual à 0,3227).

A comparação dos p-valores (sem correção e corrigidos) versus a distribuição Uniforme, pode ser vista na Figura 5.31. A distribuição dos p-valores corrigidos não se assemelha àquela obtida pelos p-valores sem a correção. A distribuição dos p-valores corrigidos está

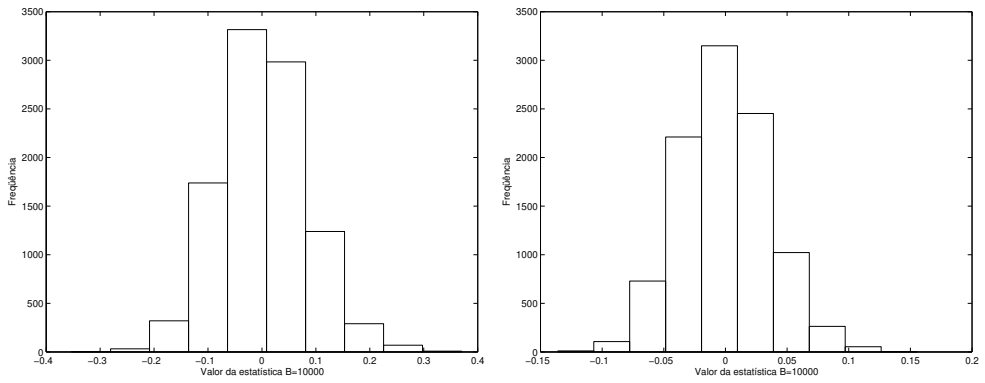

Figura 5.30: Distribuição da estatística do teste para o *locus* D5S1473 e D20S448 da análise de ALDX1, respectivamente).

bem acima da distribuição dos p-valores sem correção. Essa mesma inversão ocorreu para a estatística do teste baseada nos desvios quadráticos.

Considerando um nível de significância de 5%, temos 16 testes significativos dos 219. Com a correção, em 2 testes ou *loci* rejeitamos H_0, ou seja, encontramos evidência de diferenças significativas em pelo menos dois grupos de populações. Encontramos diferenças significativas nos *loci* D5S1473 e GATA62F03. Conforme já discutimos, no *locus* D5S1473, encontramos diferença na distribuição de todos os grupos. Através da análise da distribuição dos tamanhos alélicos, no *locus* GATA62F03, encontramos diferenças no grupo Afetado (mais detalhes na Seção 5.3). Na estatística do teste baseada em desvios quadráticos encontramos diferenças significativas somente no *locus* D5S1473.

5.3 Discussão

Nesta dissertação de mestrado, nosso interesse foi estudar as medidas de distância genética para *loci* de microsatélites baseadas nos desvios absolutos e quadráticos sob o modelo de mutação "*stepwise*". Foram também propostos dois testes de homogeneidade, um baseado na estatística do teste dos desvios quadráticos (Capítulo 3) e outro na dos devios absolutos (Capítulo 4). Aplicamos esses testes a dados reais, em que o interesse é verificar se existe

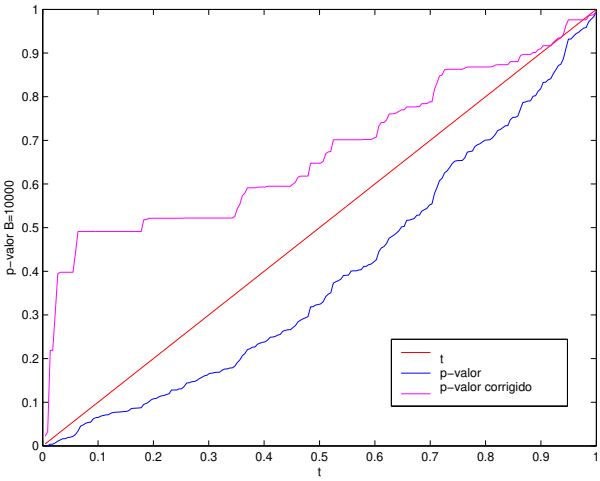

Figura 5.31: Comparação dos p-valores para 219 testes da estatística do teste baseada nos desvios absolutos relacionados à ALDX1 com $t = \dfrac{i}{219}$ $i = 1, \ldots, 219$.

ou não diferença na variação do número de repetições para os grupos definidos pela etnia e o índice de alcoolismo (ALDX1) em um determinado *locus*. Para encontrar o nível de significância do teste, aplicamos o método Bootstrap.

No Apêndice B, nas Figuras B.1, B.2 e B.3, podemos ver os gráficos das posições dos *loci* no DNA para cada cromossomo versus o p-valor corrigido obtido pela distribuição assintótica para etnia. A linha em 0,05 representa o ponto de corte, ou seja, abaixo desta linha os testes são considerados significativos. Não foi feito os gráficos dos cromossomos 17, 18 e 22, pois estes apresentavam 1 a 2 *loci* e não apresentaram *loci* significativos. Para todos cromossomos analisados vemos variação do p-valor corrigido com relação à posição do *locus* no DNA. Para os grupos do índice ALDX não foram feito esses gráficos, pois os p-valores corrigidos apresentavam pouca variação. No entanto, no cromossomo 5 essa variação ocorreu, sendo que esta é na posição do *locus* que encontramos diferenças significativas, conforme pode ser observado na Figura 5.32.

Pela Figura B.1 (Apêndice B), vemos que nos cromossomos 3 e 4 não encontramos diferenças significativas. No cromossomo 2 pode-se ver 4 *loci* que apresentam diferenças significativas entre os grupos de etnia. Nos cromossomos 1, 2, 5 e 6, os p-valores variam

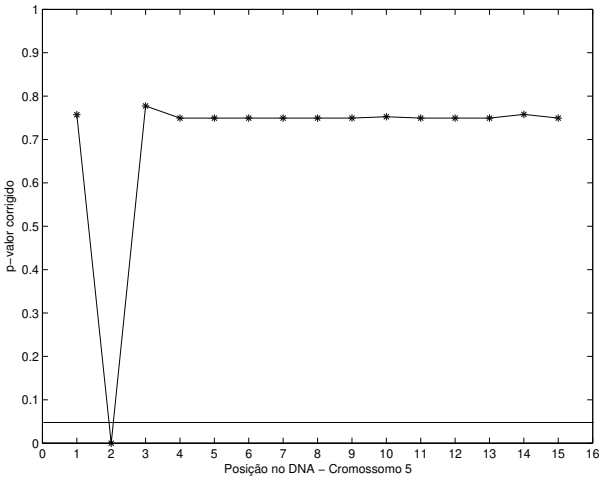

Figura 5.32: Posição no DNA versus o p-valor corrigido obtido pela distribuição assintótica para ALDX1, cromossomo 5.

bastante, sendo que no cromossomo 5 a variação é mais acentuada na posição 2. Na Figura B.2 podemos observar que os cromossomos 10, 11, 14 e 15 não apresentam *loci* estatisticamente significantes. Os cromossomos 8, 12 e 13 apresentam variações nos p-valores mais acentuadas em determinadas posições, por exemplo, no cromossomo 12 e nas posições 1 a 7 os p-valores variam pouco e na posição 8 o p-valor descresce muito. Nos cromossomos 7 e 8, as variações não são específicas a um determinado *locus*, ou seja, os p-valores aumentam e diminuem ao longo dos *loci*. Da mesma forma, podemos analisar a Figura B.3, em que os cromossomos 16 e 19 não apresentam *loci* com diferenças, entre os grupos de etnia, significativas.

Na Seção 5.2.3, considerando os p-valores corrigidos para comparações múltiplas, vimos que em 46 testes ou *loci* rejeitamos H_0 com um nível de significância 5% para o grupo etnia. Para o índice ALDX1, temos que, para a mesma situação, em 1 *locus* rejeitamos H_0.

Na Seção 5.2.4, considerando os p-valores corrigidos para comparações múltiplas, temos que em 55 testes rejeitamos H_0 ao nível de 5% para a etnia. Para o índice ALDX1, temos que, para a mesma situação, em 2 testes rejeitamos H_0.

Asssim, tanto para etnia como para o índice ALDX1, a estatística do teste baseada em desvios absolutos obteve mais testes que rejeitam H_0 do que a estatística do teste baseada em desvios quadráticos. Para os grupos da etnia, todos testes significativos para estatística do teste baseada nos desvios quadráticos foram significativos para a estatística baseada nos desvios absolutos.

Para ALDX1, a estatística do teste baseada em desvios absolutos obteve um *locus* a mais que rejeita H_0 do que a estatística do teste baseada em desvios quadráticos, ou seja, a diferença foi pequena. O *locus* D5S1473 foi encontrada evidência de diferenças significativas nas distribuições de pelos menos dois grupos, para as duas estatísticas. Analisando as distribuições dos tamanhos alélicos por grupos, a diferença está nas distribuições dos grupos Puramente não afetado/Nunca bebeu e os outros dois grupos (Afetados e Não afetado (com alguns sintomas)), conforme pode ser visto na Figura 5.33. Além disso, pode-se ver diferenças nas distribuições dos grupos Afetados e Não afetados (com alguns sintomas). Então, vemos diferenças nas distribuições dos 3 grupos, sendo que estas são maiores entre os grupos Puramente não afetado/Nunca bebeu e os outros dois (Afetados e Não afetado (com alguns sintomas)).

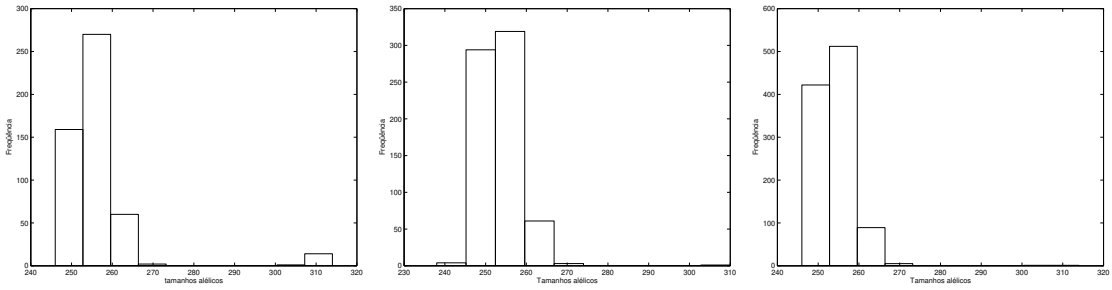

Figura 5.33: Distribuição dos tamanhos alélicos do *locus* D5S1473 para o índice ALDX1, grupos: Puramente não Afetado/Nunca Bebeu, Não afetado com alguns sintomas e Afetado, respectivamente.

Na Figura 5.34, vemos os histogramas para o *locus* GATA62F03, que deu diferença significativa na estatística do teste baseada nos desvios absolutos e não deu diferença na estatística do teste baseada nos desvios quadráticos para o índice ALDX1. Note que, no

grupo afetado temos uma pequena diferença na distribuição com relação aos outros dois grupos. Nos grupos Puramente não afetado/Nunca bebeu e Não afetado (com alguns sintomas) vemos que a distribuição é bimodal, enquanto que esta deixa de ser quando consideramos a distribuição do grupo Afetado.

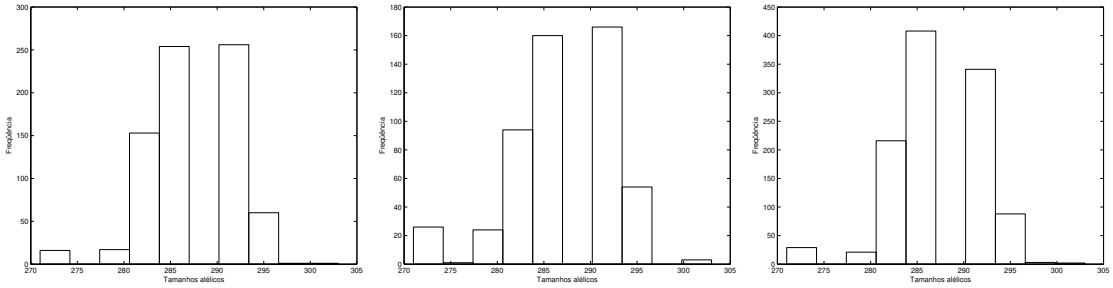

Figura 5.34: Distribuição dos tamanhos alélicos do *locus* GATA62F03 para o índice ALDX1, grupos: Puramente não Afetado/Nunca Bebeu, Não afetado com alguns sintomas e Afetado, respectivamente.

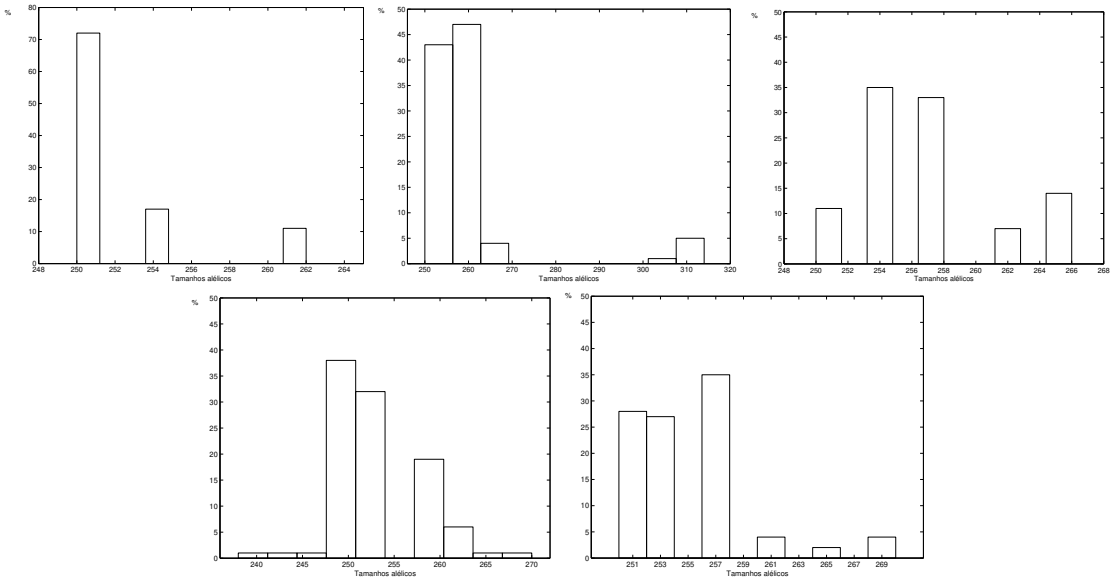

Figura 5.35: Distribuição dos tamanhos alélicos do *locus* D5S1473 para os grupos de etnia: índio americano, negros não hispânicos, negros hispânicos, brancos não hispânicos e brancos hispânicos, respectivamente.

Desta forma, considerando o método Bootstrap e o número de testes que foram significativos, o teste baseado nos desvios absolutos foi mais sensível para detectar diferenças na variação do número de repetições entre os grupos, principalmente nos grupos da etnia.

No Capítulo 3 encontramos a distribuição da estatística do teste baseada nos desvios quadráticos. Para essa estatística, aplicamos o teste de homogeneidade baseado na distribuição assintótica. Na Seção 5.2.3 vimos que, para etnia, em 18 testes rejeitamos H_0 e para ALDX1, em apenas 1 teste rejeitamos H_0. O único *locus* significativo para o índice ALDX1 foi D5S1473, que coincide com o resultado obtido pelo método Bootstrap para as estatísticas dos desvios quadráticos e absolutos. Porém, conforme já foi discutido, um *locus* a mais foi considerado significativo pela medida de desvios absolutos. Dos 18 *loci* que encontramos evidência de diferenças significativas para os grupos de etnia, nos *loci* D5S1473 e D6S1052 não encontramos evidência de diferenças, cujos p-valores foram encontrados utilizando o método Bootstrap devido à estatística baseada nos desvios quadráticos. Para a estatística baseada nos desvios absolutos, somente no *locus* D5S1473 não rejeitamos H_0.

A distribuição dos tamanhos alélicos, para o *locus* D5S1473, por grupos pode ser vista pela Figura 5.35. Todas elas diferem com relação à moda e à dispersão. Assim, as distribuições são bem distintas, então encontramos diferenças em todos os grupos de etnias. Neste *locus*, dentro de cada grupo temos bastante variação, além de grande variação entre grupos. O fato do método Bootstrap não detectar diferença significativa neste *locus* poder ser explicado por isso, pois por esse método as 5 populações distintas são misturadas com isso pode-se obter valores da estatística do teste muito grandes e ocultar diferenças significativas. A Figura 5.36 corrobora esse fato, pois podemos ver que a distribuição Bootstrap da estatística do teste baseada nos desvios absolutos e quadráticos tem valores com amplitude grande, para os grupos de etnia (o valor da estatística do teste baseada nos desvios quadráticos é 46,01 e da estatística baseada nos desvios absolutos é 1,29).

Na simulação foi visto que a estatística do teste baseada em desvios quadráticos, $QM_{(B-PW)l}(t) = QM_{Bl}(t) - QMP_{Wl}(t)$, tem um comportamento melhor com relação à

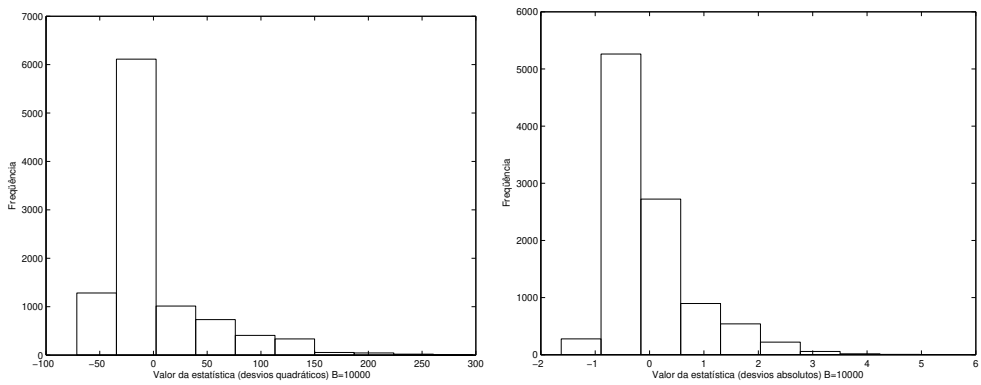

Figura 5.36: Distribuição Bootstrap para o *locus* D5S1473 para a estatística baseada nos desvios quadráticos e absolutos, respectivamente.

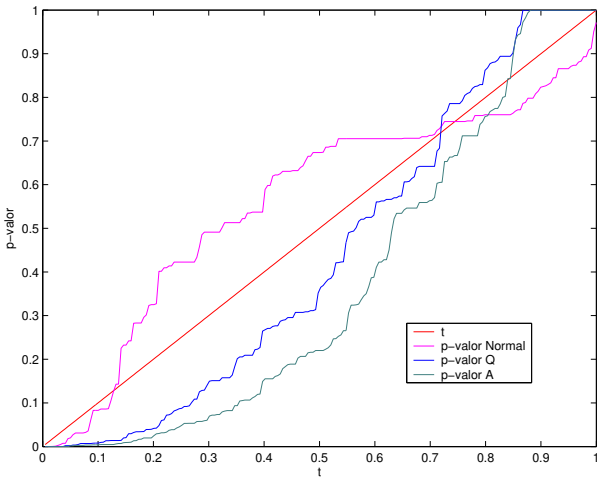

Figura 5.37: Comparação dos p-valores corrigidos obtidos pela distribuição assintótica Normal, pelo Bootstrap para a estatística do teste baseada nos desvios qudráticos-Q e absolutos-A e $t = \dfrac{i}{219}$ $i = 1, \ldots, 219$ para etnia.

normalidade que a estatística baseada em desvios absolutos, $AM_{(B-PW)l}(t) = AM_{Bl}(t) - AMP_{Wl}(t)$, quando considero amostras não balanceadas. Para amostras balanceadas, os comportamentos com respeito a normalidade são semelhantes para $QM_{(B-W)l}(t) = QM_{Bl}(t) - QM_{Wl}(t)$ e $AM_{(B-W)l}(t) = AM_{Bl}(t) - AM_{Wl}(t)$ e razoáveis para $n = 300$.

Para amostras não balanceadas, para atingir normalidade é necessário um tamanho

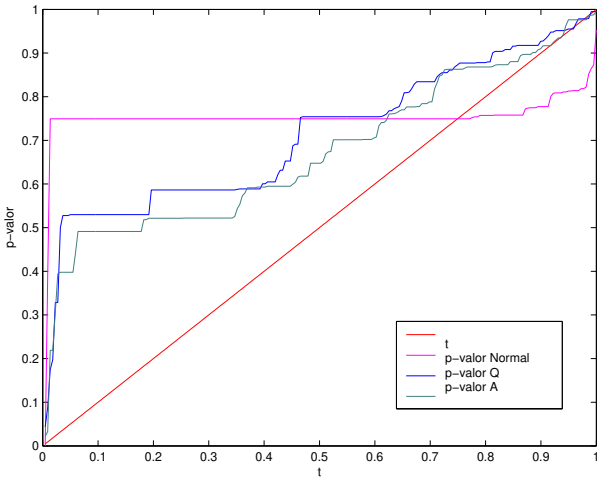

Figura 5.38: Comparação dos p-valores corrigidos obtidos pela distribuição assintótica Normal, pelo Bootstrap para a estatística do teste baseada nos desvios qudráticos-Q e absolutos-A e $t = \dfrac{i}{219}$ $i = 1, \ldots, 219$ para ALDX1.

amostral menor para $QM_{(B-PW)l}(t)$, $n > 300$, comparado com o tamanho amostral necessário para $AM_{(B-PW)l}(t)$, pois precisamos que o tamanho amostral seja maior que 400. Desta forma, os resultados obtidos pela distribuição assintótica são válidos.

Em geral, pela simulação, obervamos que o comportamento com respeito a normalidade foi melhor quando considero amostras balanceadas, pois com um tamanho amostral menor ($n = 300$) obtemos resultados melhores, tanto para a estatística do teste baseada nos desvios quadráticos com a baseada nos desvios absolutos.

Nas Figuras 5.37 e 5.38 vemos os gráficos de comparação dos p-valores obtidos pela distribuição assintótica (p-valor Normal em rosa), pelo Bootstrap para a estatística do teste baseada nos desvios quadráticos (p-valor Q em azul) e absolutos (p-valor A em verde), para etnia e o índice ALDX1, respectivamente. Observando as Figuras, vemos que os p-valores Bootstrap obtidos pelas estatísticas baseadas nos desvios absolutos e quadráticos apresentam a mesma forma de distribuição, sendo que a distribuição dos p-valores baseados na estatística dos desvios quadráticos apresentam valores maiores que os obtidos pela estatística baseada nos desvios absolutos. A distribuição dos p-valores obtidos

pela distribuição assintótica da estatística do teste baseada nos desvios quadráticos difere das outras duas distribuições. Esta apresenta valores maiores para p-valor maior que 0,7 e menores para o outro comparado com as outras duas distribuições, tanto para etnia como para ALDX1.

Não aplicamos o teste de homogeneidade utilizando a distribuição assintótica da estatística do teste baseada nos desvios absolutos e a estimativa da variância de estatística U por Jackknife, definida na equação (4.3.5), devido ao tamanho da amostra muito elevado e as dificuldades computacionais. Sendo assim, se faz necessário mais estudos para essa estatística.

Quando utilizamos o Bootstrap para fazer testes de comparações de 3 ou mais grupos podemos ter uma situação complicada. Suponhamos que temos 3 grupos, dois grupos homogêneos entre si e um outro bem heterogêneo e diferente dos outros dois. Quando fazemos as reamostras Bootstrap da mistura dos três grupos podemos ter os três grupos bem homogêneos entre si e heterogêneos dentre eles, o que aumentaria a variação dentre grupos e diminuiria a variação entre. Isso provocaria a subestimação do p-valor (mais detalhes ver referência Pinheiro et al. (2007)). Isso poderia ser uma explicação dos p-valores obtidos pelo Bootstrap serem menores quando comparados com os p-valores obtidos pela distribuição assintótica, para o p-valor$<$0,7.

Uma solução proposta pelo artigo Pinheiro et al. (2007) é aplicar o método Jackknife. No entanto, para o estudo de 219 *loci* esse método seria mais trabalhoso computacionalmente, então não foi aplicado.

Devido aos dois problemas levantados, os resultados obtidos pela distribuição assintótica da estatística do teste baseada nos desvios quadráticos são razoáveis e melhores que os obtidos pelo método Bootstrap, pois possivelmente os p-valores Bootstrap estão subestimados. Esses problemas são mais críticos quando estudamos os grupos da etnia, pois possivelmente, estes são mais heterogêneos entre si.

Para a estatística do teste baseada nos desvios absolutos a única opção é o método Bootstrap, pois não foi possível fazer o teste utilizando a distribuição assintótica devido

às dificuldades computacionais.

Apêndice A

Demonstrações

1. *Demonstração*:

$$E\left[M_1^2(t)\right] = (2N_e)^{-2}\sum_{i\neq j} ijE[S_i(t)S_j(t)] + (2N_e)^{-2}\sum_i i^2 E\left[S_i^2(t)\right].$$

Visto que $S_i(t) \sim Binomial(2N_e, \pi_i(t))$, então

$$\begin{aligned} E\left[S_i^2(t)\right] &= Var[S_i(t)] + (E[S_i(t)])^2 \\ &= 2N_e\pi_i(t)(1-\pi_i(t)) + (2N_e)^2\pi_i(t)^2 \end{aligned}$$

e o vetor $\mathbf{S}(t)$ segue uma distribuição multinomial, então $E[S_i(t)S_j(t)] = 2N_e(2N_e - 1)\pi_i(t)\pi_j(t)$, pois $\text{cov}[S_i(t), S_j(t)] = -2N_e\pi_i(t)\pi_j(t)$ e assim

$$E[S_i(t)S_j(t)] = \text{cov}[S_i(t), S_j(t)] + E[S_i(t)]E[S_j(t)],$$

desta forma

$$\begin{aligned} E\left[M_1^2(t)\right] &= \left(1-\frac{1}{2N_e}\right)\sum_{i\neq j} ij\pi_i(t)\pi_j(t) + \left(1-\frac{1}{2N_e}\right)\sum_i i^2\pi_i^2(t) + \\ &+ \frac{1}{2N_e}\sum_i i^2\pi_i(t) \\ &= \left(1-\frac{1}{2N_e}\right)\sum_{i,j} ij\pi_i(t)\pi_j(t) + \frac{1}{2N_e}\sum_i i^2\pi_i(t), \end{aligned}$$

substituindo $\pi_i(t)$ por seu estimador dado pela expressão (2.2.2) e omitindo o $(t-1)$, têm-se

$$\sum_{i,j} ij\hat{\pi}_i(t)\hat{\pi}_j(t) = \frac{1}{(2N_e)^2}\left\{(1-\beta)^2\sum_{i,j} ijn_in_j + \frac{\beta}{2}(1-\beta)\sum_{i,j} ijn_{i+1}n_j + \right.$$
$$+\frac{\beta}{2}(1-\beta)\sum_{i,j} ijn_in_{j-1} + \frac{\beta}{2}(1-\beta)\sum_{i,j} ijn_in_{j+1} + \frac{\beta}{2}(1-\beta)\sum_{i,j} ijn_{i-1}n_j +$$
$$\left.+\frac{\beta^2}{4}\sum_{i,j} ijn_{i-1}n_{j-1} + \frac{\beta^2}{4}\sum_{i,j} ijn_{i-1}n_{j+1} + \frac{\beta^2}{4}\sum_{i,j} ijn_{i+1}n_{j-1} + \frac{\beta^2}{4}\sum_{i,j} ijn_{i+1}n_{j+1}\right\}$$
$$= (1-\beta)^2(M_1(t-1))^2 + 2\beta(1-\beta)(M_1(t-1))^2 + \beta^2(M_1(t-1))^2,$$

resultado obtido com algumas manipulações algébricas e considerando que $\sum_i n_i(t-1) = 2N_e$. Assim, $\sum_{i,j} ij\hat{\pi}_i(t)\hat{\pi}_j(t) = (M_1(t-1))^2$ e,

$$\begin{aligned}
\sum_i i^2\hat{\pi}_i(t) &= \frac{1}{2N_e}\sum_i i^2\left[(1-\beta)n_i + \frac{\beta}{2}n_{i-1} + \frac{\beta}{2}n_{i+1}\right] \\
&= (1-\beta)M_2(t-1) + \frac{1}{2N_e}\left\{\frac{\beta}{2}\sum_i (i-1+1)^2 n_{i-1} + \frac{\beta}{2}\sum_i (i+1-1)^2 n_{i+1}\right\} \\
&= (1-\beta)M_2(t-1) + \frac{1}{2N_e}\left\{\frac{\beta}{2}\sum_i (i-1)^2 n_{i-1} + \beta\sum_i (i-1)n_{i-1} + \right. \\
&- \left.\frac{\beta}{2}\sum_i (i+1)^2 n_{i+1} + \beta\sum_i (i+1)n_{i+1} + \frac{\beta}{2}\sum_i n_{i+1} + \frac{\beta}{2}\sum_i n_{i-1}\right\} \\
&= (1-\beta)M_2(t-1) + \beta M_2(t-1) + \beta M_1(t-1) + \beta - \beta M_1(t-1) \\
&= M_2(t-1) + \beta.
\end{aligned}$$

Finalmente,

$$\begin{aligned}
E\left[M_1^2(t)\right] &= (1-(2N_e)^{-1})\sum_{i,j} ij\pi_i(t)\pi_j(t) + (2N_e)^{-1}\sum_i i^2\pi_i(t) \\
&= (1-(2N_e)^{-1})M_1^2(t-1) + (2N_e)^{-1}[M_2(t-1)+\beta] \\
&= (1-(2N_e)^{-1})M_1^2(t-1) + (2N_e)^{-1}M_2(t-1) + \beta(2N_e)^{-1}.
\end{aligned}$$

□

2. *Demonstração*: Temos que

$$[1+2u+2v]a\lambda^j = av\lambda^{j-2}+au\lambda^{j-1}+au\lambda^{j+1}+av\lambda^{j+2},$$

cancelando os a's, e colocando v e u em evidência, têm-se

$$[1+2u+2v\ ]\lambda^j = v\ [\lambda^{j-2}+\lambda^{j+2}\ ]+u\ [\lambda^{j-1}+\lambda^{j+1}\ ],$$

cancelando os λ^j, têm-se

$$[1+2u+2v\ ] = v\ \left[\frac{1}{\lambda^2}+\lambda^{+2}\right]+u\ \left[\frac{1}{\lambda}+\lambda\right]$$

e desta forma, completando o quadrado têm-se

$$[1+2u+2v\ ] = v\left[\frac{1}{\lambda}+\lambda\right]^2-2v+u\left[\frac{1}{\lambda}+\lambda\right].$$

□

3. *Demonstração*: Temos que

$$\frac{1}{\lambda_1} = \frac{4v}{S-u-\sqrt{(S-u)^2-(4v)^2}},$$

multiplicando o denominador por $S-u+\sqrt{(S-u)^2-(4v)^2}$, tem-se

$$\begin{aligned}\frac{1}{\lambda_1} &= \frac{4v\ [S-u+\sqrt{(S-u)^2-(4v)^2}\ ]}{(S-u)^2-(S-u)^2+(4v)^2}\\ &= \frac{S-u+\sqrt{(S-u)^2-(4v)^2}}{4v}=\lambda_3,\end{aligned}$$

o mesmo ocorre com $\lambda_2^{-1}=\lambda_4$. □

4. *Demonstração*: É fato que

$$C_0+2\sum_{i=1}^{\infty}C_j=1 \quad \text{então} \quad a_1+a_2+2\sum_{i=1}^{\infty}\left[a_1\lambda_1^j+a_2\lambda_2^j\right]=1,$$

utilizando soma infinita de PG, como $-1<\lambda_2<0<\lambda_1<1$, essa soma converge. Logo

$$\begin{aligned}a_1+a_2+2\left[a_1\sum_{i=1}^{\infty}\lambda_1^j+a_2\sum_{i=1}^{\infty}\lambda_2^j\right] &= 1,\\ a_1+a_2+2\left[a_1\frac{\lambda_1}{1-\lambda_1}+a_2\frac{\lambda_2}{1-\lambda_2}\right] &= 1,\end{aligned}$$

então o resultado $a_1\frac{1+\lambda_1}{1-\lambda_1}+a_2\frac{1+\lambda_2}{1-\lambda_2}=1$ é válido. □

5. *Demonstração (Teorema Révész (1990))*: Vamos provar que

$$E[exp(tX_\alpha)]=\left(\frac{e^t+e^{-t}}{2}\right)^\alpha.$$

Sabemos que,

$$E\left[e^{tX_\alpha}\right]=\sum_{k=-\alpha}^{\alpha}e^{tk}\left(\begin{array}{c}\alpha\\ \frac{\alpha-k}{2}\end{array}\right)2^{-\alpha}.$$

Note que, se α é ímpar então k é ímpar e se α é par então k é par. Seja $\alpha=n$, então $-n\leq k\leq n$, assim, $k\in\{-n,-n+2,-n+4,\ldots,n-4,n-2,n\}$. Por exemplo, para $\alpha=3$, temos

$$\begin{aligned}E[exp(tX_3)] &= 2^{-3}\sum_{k=-3}^{3}e^{tk}\left(\begin{array}{c}3\\ \frac{3-k}{2}\end{array}\right)\\ &= 2^{-3}\left(\left(\begin{array}{c}3\\3\end{array}\right)e^{-3t}+\left(\begin{array}{c}3\\2\end{array}\right)e^{-t}+\left(\begin{array}{c}3\\1\end{array}\right)e^{t}+\left(\begin{array}{c}3\\0\end{array}\right)e^{3t}\right)\\ &= 2^{-3}\sum_{k=0}^{3}\left(\begin{array}{c}3\\k\end{array}\right)e^{(3-2k)t}.\end{aligned}$$

Para $\alpha=4$, temos

$$\begin{aligned}E[exp(tX_4)] &= 2^{-4}\sum_{k=-4}^{4}e^{tk}\left(\begin{array}{c}4\\ \frac{4-k}{2}\end{array}\right)\\ &= 2^{-4}\left(\left(\begin{array}{c}4\\4\end{array}\right)e^{-4t}+\left(\begin{array}{c}4\\3\end{array}\right)e^{-2t}+\left(\begin{array}{c}4\\2\end{array}\right)+\left(\begin{array}{c}4\\1\end{array}\right)e^{2t}+\left(\begin{array}{c}4\\0\end{array}\right)e^{4t}\right)\\ &= 2^{-4}\sum_{k=0}^{4}\left(\begin{array}{c}4\\k\end{array}\right)e^{(4-2k)t}.\end{aligned}$$

Assim,

$$E[exp(tX_n)] = 2^{-n}\sum_{k=-n}^{n}e^{tk}\left(\begin{array}{c}n\\ \frac{n-k}{2}\end{array}\right)$$

$$
\begin{aligned}
&= 2^{-n}\left(\left(\begin{array}{c} n \\ n \end{array}\right)e^{-nt} + \left(\begin{array}{c} n \\ n-1 \end{array}\right)e^{(-n+2)t} + \ldots + \right. \\
&+ \left.\left(\begin{array}{c} n \\ 1 \end{array}\right)e^{(n-2)t} + \left(\begin{array}{c} n \\ 0 \end{array}\right)e^{nt}\right) \\
&= 2^{-n}\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)e^{(n-2k)t} = 2^{-n}\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)e^{(n-2k)t} \\
&= 2^{-n}e^{nt}\sum_{k=0}^{n}\left(\begin{array}{c} n \\ k \end{array}\right)e^{(-2t)k} = 2^{-n}e^{nt}(e^{-2t}+1)^n,
\end{aligned}
$$

em que a última igualdade se deve ao binômio de Newton, ou seja,

$$\sum_{m=0}^{n}\left(\begin{array}{c} n \\ m \end{array}\right)x^m = (1+x)^n.$$

Assim, $M_{X_n}(t) = E[exp(tX_n)] = e^{nt}\left(\dfrac{e^{-2t}+1}{2}\right)^n = \left(\dfrac{e^t+e^{-t}}{2}\right)^n$.

Para obtermos $E[X_\alpha] = 0$ e $E[X_\alpha^2] = \alpha$, basta encontrar a primeira e segunda derivadas de $M_{X_\alpha}(t)$ para $t = 0$, respectivamente. Ou seja,

$$E[X_\alpha] = \left.\frac{\partial M_{X_\alpha}(t)}{\partial t}\right|_{t=0} \quad \text{e} \quad E[X_\alpha^2] = \left.\frac{\partial^2 M_{X_\alpha}(t)}{\partial t^2}\right|_{t=0}.$$

Assim,

$$
\begin{aligned}
\frac{\partial M_{X_\alpha}(t)}{\partial t} &= \frac{\alpha}{2}\left(\frac{e^t+e^{-t}}{2}\right)^{\alpha-1}(e^t - e^{-t}); \\
\frac{\partial^2 M_{X_\alpha}(t)}{\partial t^2} &= \frac{\alpha(\alpha-1)}{4}\left(\frac{e^t+e^{-t}}{2}\right)^{\alpha-2}(e^t-e^{-t})^2 + \alpha\left(\frac{e^t+e^{-t}}{2}\right)^{\alpha}, \quad \text{com isso}
\end{aligned}
$$

$$\left.\frac{\partial M_{X_\alpha}(t)}{\partial t}\right|_{t=0} = 0 \quad \text{e} \quad \left.\frac{\partial^2 M_{X_\alpha}(t)}{\partial t^2}\right|_{t=0} = \alpha.$$

□

6. *Demonstração:* Por propriedade temos que

$$E[\Delta_{12l}^m(t) \mid N(t) = \alpha, T = t] = \left[\frac{\partial^m M(x)}{\partial x^m}\right]_{x=0}.$$

Segue do resultado do Teorema 3.1 que

$$E[\Delta_{12l}(t) \mid N(t)=\alpha, T=t]=0 \quad \text{e} \quad E[\Delta^2_{12l}(t) \mid N(t)=\alpha, T=t]=\alpha.$$

Derivando $M(x)$ com respeito a x, para obter os terceiro e quarto momentos, temos

$$\begin{aligned}
\frac{\partial M(x)}{\partial x} &= \frac{\alpha}{2}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-1}(e^x-e^{-x});\\
\frac{\partial^2 M(x)}{\partial x^2} &= \frac{\alpha(\alpha-1)}{4}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-2}(e^x-e^{-x})^2+\alpha\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha};\\
\frac{\partial^3 M(x)}{\partial x^3} &= \frac{\alpha(\alpha-1)(\alpha-2)}{8}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-3}(e^x-e^{-x})^3+\\
&+ \alpha(\alpha-1)\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-1}(e^x-e^{-x})+\frac{\alpha^2}{2}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-1}(e^x-e^{-x}) \quad \text{e}\\
\frac{\partial^4 M(x)}{\partial x^4} &= \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)}{16}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-4}(e^x-e^{-x})^4+\\
&+ 3\frac{\alpha(\alpha-1)(\alpha-2)}{4}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-2}(e^x-e^{-x})^2+\\
&+ \frac{\alpha(\alpha-1)^2}{2}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-2}(e^x-e^{-x})^2+\\
&+ 2\alpha(\alpha-1)\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha}+\frac{\alpha^2(\alpha-1)}{4}\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha-2}(e^x-e^{-x})+\\
&+ \alpha^2\left(\frac{e^x+e^{-x}}{2}\right)^{\alpha}.
\end{aligned}$$

Avaliando em $x=0$ segue o resultado de (3.2.3). □

7. *Demonstração*: Temos que,

$$\begin{aligned}
QM_{Wl}(t) &= \frac{1}{G}\sum_{g=1}^{G}\frac{2}{2n(2n-1)}\sum_{i=1}^{2n}\sum_{i'>i}(Y_{igl}(t)-Y_{i'gl}(t))^2\\
&= \frac{1}{G}\sum_{g=1}^{G}\frac{1}{2n(2n-1)}\sum_{i=1}^{2n}\sum_{i'=1}^{2n}(Y^2_{igl}(t)-2Y_{igl}(t)Y_{i'gl}(t)+Y^2_{i'gl}(t))\\
&= \frac{1}{G}\sum_{g=1}^{G}\frac{1}{2n(2n-1)}\left[2n\sum_{i=1}^{2n}Y^2_{igl}(t)-2\sum_{i=1}^{2n}Y_{igl}(t)\sum_{i'=1}^{2n}Y_{i'gl}(t)+2n\sum_{i'=1}^{2n}Y^2_{i'gl}(t)\right]\\
&= \frac{1}{G}\sum_{g=1}^{G}\frac{1}{2n(2n-1)}\left[2(2n)\sum_{i=1}^{2n}Y^2_{igl}(t)-2(2n)^2\bar{Y}^2_{igl}(t)\right]
\end{aligned}$$

$$= \frac{2}{G(2n-1)} \sum_{g=1}^{G} \left[\sum_{i=1}^{2n} Y_{igl}^2(t) - 2n\bar{Y}_{igl}^2(t) \right] = \frac{2}{G} \sum_{g=1}^{G} S_{gl}^2(t).$$

□

8. *Demonstração*: Sabemos que,

$$QM_{Totl}(t) = \frac{2}{2nG(2nG-1)} \left(S_{Wl}(t) + S_{Bl}(t) \right).$$

Utilizando o resultado de (3.3.1), temos que

$$\begin{aligned} QM_{Totl}(t) &= \frac{1}{2nG(2nG-1)} \left[\sum_{g=1}^{G} \sum_{g'=1}^{G} \sum_{i=1}^{2n} \sum_{i'=1}^{2n} (Y_{igl}(t) - Y_{i'g'l}(t))^2 \right] \\ &= \frac{1}{2nG(2nG-1)} \left[2nG \sum_{g=1}^{G} \sum_{i=1}^{2n} Y_{igl}(t)^2 - 2 \sum_{g=1}^{G} \sum_{i=1}^{2n} Y_{igl}(t) \sum_{g'=1}^{G} \sum_{i'=1}^{2n} Y_{i'g'l}(t) + \right. \\ &+ \left. 2nG \sum_{g'=1}^{g} \sum_{i'=1}^{2n} Y_{i'g'l}^2(t) \right] \\ &= \frac{2}{2nG(2nG-1)} \left[2nG \sum_{g=1}^{G} \sum_{i=1}^{2n} Y_{igl}(t)^2 - (2nG)^2 \bar{Y}_l(t)^2 \right] \\ &= \frac{2}{2nG-1} \left[\sum_{g=1}^{G} \sum_{i=1}^{2n} Y_{igl}(t)^2 - (2nG) \bar{Y}_l(t)^2 \right] = 2S_l^2(t). \end{aligned}$$

□

9. Para a prova do Teorema 3.2 precisamos da definição de Desigualdade de Jensen, dada a seguir.

Definição A.1. *Desigualdade de Jensen. Seja* Φ *uma função convexa. Então*

$$\mathrm{E}\left[\Phi(\boldsymbol{X}) \mid \boldsymbol{Y}\right] \geq \Phi\left(\mathrm{E}[\boldsymbol{X} \mid \boldsymbol{Y}]\right).$$

Demonstração: O kernel associado a S é $\frac{1}{n!} \sum_{\pi} S(x_{i1}, \ldots, x_{in})$ no qual, neste caso, quando $n = m$ é a estatística U associada com ela mesma. Isso significa que a estatística de ordem pode ser expressa por $U = E\left[S(\mathbf{X}) \mid \mathbf{X}_{(.)}\right]$, pois consideramos

que queremos encontrar a distribuição condicional de $\mathbf{X} = (X_1, \ldots, X_n)$ dadas as estatística de ordem $\mathbf{X}_{(\cdot)} = \big(X_{(1)}, \ldots, X_{(n)}\big)$. Sabemos que, por construção $X_1, \ldots, X_n$ são independentes e identicamente distribuídas. Suponha que F seja contínua. Seja $\mathbf{x}_{(\cdot)} = (x_{(1)}, \ldots, x_{(n)})$, em que $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. É evidente que $\mathbf{X}_{(\cdot)} = \mathbf{x}_{(\cdot)}$ se, e somente se, $\mathbf{X}$ é permutação de $\mathbf{x}_{(\cdot)}$. Pelo princípio de preservação de chances relativas, temos que antes da observação da estatística de ordem, toda permutação de $\mathbf{x}_{(\cdot)}$ tinha a mesma chance relativa de ser o valor de $\mathbf{X}$, pois X_i's são independentes e identicamente distribuídas. Por exemplo, suponha que $n = 2$ e que o valor observado de $\mathbf{X}_{(\cdot)}$ foi $(x_{(1)}, x_{(2)})$. Então, $\mathbf{X} = (x_1, x_2)$ ou $\mathbf{X} = (x_2, x_1)$ e os dois valores tinham *a priori*, isto é, antes da observação de X_i, a mesma chance de serem escolhidos (pois $(X_1, X_2) \sim (X_2, X_1)$, ou seja, os dois vetores possuem a mesma distribuição). Logo é natural pensar que

$$P(\mathbf{X} = (x_1, x_2) \mid \mathbf{X}_{(\cdot)} = (x_{(1)}, x_{(2)})) = P(\mathbf{X} = (x_2, x_1) \mid \mathbf{X}_{(\cdot)} = (x_{(1)}, x_{(2)})) = \frac{1}{2}.$$

Então, o nosso candidato para n geral será

$$P(\mathbf{X} = (x_{\pi_1}, \ldots, x_{\pi_n}) \mid \mathbf{X}_{(\cdot)} = (x_{(1)}, \ldots, x_{(n)})) = \frac{1}{n!}, \quad x_{(1)} < x_{(2)} < \ldots < x_{(n)},$$

em que $\{\pi_1, \ldots, \pi_n\}$ é uma permutação de $\{1, \ldots, n\}$. Mais detalhes, ver James (2004). Logo,

$$E\left[S(\mathbf{X}) \mid \mathbf{X}_{(\cdot)}\right] = \frac{1}{n!} \sum_{\pi} S(\mathbf{X}).$$

Então,

$$E_F\left[U^2\right] = E_F\left[E^2(S(\mathbf{X}) \mid \mathbf{X}_{(\cdot)})\right] \leq E_F\left[E(S^2(\mathbf{X}) \mid \mathbf{X}_{(\cdot)})\right] = E_F(S^2(\mathbf{X})),$$

em que, a desigualdade é válida por Jensen, pois a função quadrática é convexa. A igualdade $E_F\left[U^2\right] = E_F[S^2(\mathbf{X})]$ só ocorre se, e somente se $E[S(\mathbf{X}) \mid \mathbf{X}_{(\cdot)}]$ é degenerada e igual a $S(\mathbf{X})$ com probabilidade 1. Como $E_F(U) = E_F(S)$ a prova está completa. □

10. *Demonstração*: Temos que, sob H_0,

$$\vartheta^2 = \frac{4}{2nG^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}(4\theta^2+\theta) = \frac{8\theta^2+2\theta}{2nG(G-1)}$$

resultado obtido pela equação (3.5.7). Para ϱ, temos que, sob H_0, $\rho_{gg'l} = \rho_{g'g'l} = \rho_{ggl} = 0$ e $\eta_{gl} = \eta_{g'l} = \eta_l$ para $g \neq g'$, então

$$\begin{aligned}
\varrho &= \frac{8}{G^2(G-1)^2}\left\{\sum_{g=k=1}^{G}\sum_{g'>g}\sum_{k'>g'} E\left[\bar{D}_{gg'l}(t)\bar{D}_{kk'l}(t)\right] + \right. \\
&+ \left.\sum_{g=1}^{G}\sum_{k>g}\sum_{g'=k'>k} E\left[\bar{D}_{gg'l}(t)\bar{D}_{kk'l}(t)\right] + \sum_{g=1}^{G}\sum_{g'=k>g}\sum_{k'>k} E\left[\bar{D}_{gg'l}(t)\bar{D}_{kk'l}(t)\right] - 6\sum_{c}\theta^2\right\},
\end{aligned}$$

em que $\sum_c = \sum_{1\leq g<g'<k\leq G}$, ou seja, é a soma sobre a combinação $\begin{pmatrix} G \\ 3 \end{pmatrix}$, para $G \geq 3$.

$$\begin{aligned}
\varrho &= \frac{8}{G^2(G-1)^2}\left\{\sum_{1\leq g<g'<k'\leq G}\left[\frac{1}{2n}\left(\frac{3\theta^2}{2}+\frac{\theta}{2}-2\eta_l^4\right) + \frac{(n+1)}{n}\left(\frac{\theta}{2}+\eta_l^2\right)^2 + \right.\right. \\
&- \left.\frac{4\eta_l^2}{n}(n+1)\left(\frac{\theta}{2}+\eta_l^2\right) + \frac{4\eta_l^4}{n} + \frac{2}{n}\eta_l^2\left(\frac{\theta}{2}+2n\eta_l^2\right) - 4\eta_l^2\left(\frac{\theta}{2}+\eta_l^2\right) + 3\left(\frac{\theta}{2}+\eta_l^2\right)^2\right] + \\
&+ \sum_{1\leq g<k<g'\leq G}\left[3\left(\frac{\theta}{2}+\eta_l^2\right)^2 - 4\eta_l^2\left(\frac{\theta}{2}+\eta_l^2\right) + \frac{2}{n}\eta_l^2\left(\frac{\theta}{2}+2n\eta_l^2\right) + \right. \\
&- \left.\frac{4\eta_l^2}{n}(n+1)\left(\frac{\theta}{2}+\eta_l^2\right) + \frac{4\eta_l^4}{n} + \frac{1}{2n}\left(\frac{3\theta^2}{2}+\frac{\theta}{2}-2\eta_l^4\right) + \frac{(n+1)}{n}\left(\frac{\theta}{2}+\eta_l^2\right)^2\right] + \\
&+ \sum_{1\leq g<k<k'\leq G}\left[3\left(\frac{\theta}{2}+\eta_l^2\right)^2 - \frac{4\eta^2}{n}(n+1)\left(\frac{\theta}{2}+\eta_l^2\right) + \frac{4}{n}\eta_l^4 + \frac{1}{2n}\left(\frac{3\theta^2}{2}+\frac{\theta}{2}-2\eta_l^4\right) + \right. \\
&+ \left.\left.\frac{2}{n}\eta_l^2\left(\frac{\theta}{2}+2n\eta_l^2\right) + \frac{(n+1)}{n}\left(\frac{\theta}{2}+\eta_l^2\right)^2 - 4\eta_l^2\left(\frac{\theta}{2}+\eta_l^2\right)\right] - 3\sum_{c}\theta^2\right\} \\
&= \frac{8}{G^2(G-1)^2}\left\{\sum_{1\leq g<g'<k'\leq G}\left(\frac{\theta^2}{n}+\theta^2+\frac{\theta}{4n}\right) + \sum_{1\leq g<k<g'\leq G}\left(\frac{\theta^2}{n}+\theta^2+\frac{\theta}{4n}\right)\right. \\
&+ \left.\sum_{1\leq g<k<k'\leq G}\left(\frac{\theta^2}{n}+\theta^2+\frac{\theta}{4n}\right) - 3\sum_{c}\theta^2\right\}
\end{aligned}$$

$$= \frac{24G(G-1)(G-2)}{6G^2(G-1)^2}\left(\frac{\theta^2}{n}+\theta^2+\frac{\theta}{4n}-\theta^2\right) = \frac{4(G-2)}{G(G-1)}\left(\frac{\theta^2}{n}+\frac{\theta}{4n}\right).$$

Sob H_0, temos

$$\zeta_1^* = 2\theta^2+\frac{\theta}{2} \;\Rightarrow\; Var_0(QM_{Wl}(t)) = \frac{8\theta^2+2\theta}{2nG} \quad \text{e}$$

$$\begin{aligned}\text{cov}_0(QM_{Wl}(t), QM_{Bl}(t)) &= \frac{3\theta^2+\theta}{2nG} + \frac{2}{G\begin{pmatrix} G \\ 2 \end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g}\frac{\theta^2}{2n} \\ &= \frac{6\theta^2+2\theta}{2nG}+\frac{2\theta^2}{2nG} = \frac{8\theta^2+2\theta}{2nG} = Var_0(QM_{Wl}(t)).\end{aligned}$$

Com isso,

$$\Upsilon_0 = \frac{8\theta^2+2\theta}{2nG(G-1)}+\frac{4(G-2)}{G(G-1)}\left(\frac{\theta^2}{n}+\frac{\theta}{4n}\right)-\frac{8\theta^2+2\theta}{2nG}.$$

□

11. *Demonstração*: A demonstração consiste em encontrar ς sob H_0.

Sob H_0, temos que $\rho_{gg'l} = 0$ e $\eta_{gl} = \eta_l$ para $g,\ g' = 1,\ldots,G$. Para o caso, em que $g = k$, temos

$$\begin{aligned}&16n_g^2 n_{g'} n_{k'} E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right) = \\ &= 4n_{g'}n_{k'}E\left(\sum_i Y_{igl}^2(t)\right)^2 - 4n_{g'}E\left(\sum_i Y_{igl}^2(t)\sum_i\sum_{i'} Y_{igl}(t)Y_{i'k'l}(t)\right) + \\ &+4n_g n_{g'}E\left(\sum_i Y_{igl}^2(t)\sum_{i'} Y_{i'k'l}^2\right) - 4n_{k'}E\left(\sum_i Y_{igl}^2\sum_i\sum_{i'} Y_{igl}(t)Y_{i'g'l}(t)\right) + \\ &+4E\left(\sum_i\sum_{i'} Y_{igl}(t)Y_{i'g'l}(t)\sum_i\sum_{i'} Y_{igl}(t)Y_{i'k'l}(t)\right) + \\ &-4n_g E\left(\sum_i\sum_{i'} Y_{igl}(t)Y_{i'k'l}(t)\sum_{i'} Y_{i'g'l}^2(t)\right) + 4n_g n_{k'}E\left(\sum_i Y_{igl}^2(t)\sum_{i'} Y_{i'g'l}^2(t)\right) + \\ &-4n_g E\left(\sum_i\sum_{i'} Y_{igl}(t)Y_{ik'l}(t)\sum_{i'} Y_{i'g'l}^2(t)\right) + 4n_g^2 E\left(\sum_{i'} Y_{i'g'l}^2(t)\sum_{i'} Y_{i'k'l}^2(t)\right)\end{aligned}$$

$$= 8n_{g'}n_{k'}n_g\left(\frac{3\theta^2}{2}+\frac{\theta}{2}-2\eta_l^4\right)+16n_gn_{g'}n_{k'}(n_g+1)\left(\frac{\theta}{2}+\eta_l^2\right)^2+$$
$$-32n_{g'}n_{k'}n_g\eta_l^2\left[\frac{\theta}{2}+n_g\left(\frac{\theta}{2}+\eta_l^2\right)\right]+16n_g^2n_{g'}n_{k'}\left(\frac{\theta}{2}+\eta_l^2\right)^2+$$
$$-32n_{g'}n_gn_{k'}\eta_l^2\left[\frac{\theta}{2}+n_g\left(\frac{\theta}{2}+\eta_l^2\right)\right]+32n_{g'}n_gn_{k'}\eta_l^2\left(\frac{\theta}{2}+2n_g\eta_l^2\right)+$$
$$-32n_g^2n_{g'}n_{k'}\eta_l^2\left(\frac{\theta}{2}+\eta_l^2\right)+16n_g^2n_{k'}n_{g'}\left(\frac{\theta}{2}+\eta_l^2\right)^2-32n_g^2n_{g'}n_{k'}\eta_l^2\left(\frac{\theta}{2}+\eta_l^2\right)+$$
$$+16n_g^2n_{g'}n_{k'}\left(\frac{\theta}{2}+\eta_l^2\right)^2,$$

$$E\left(\bar{D}_{gg'l}(t)\bar{D}_{gk'l}(t)\right)=\frac{1}{2n_g}\left(\frac{3\theta^2}{2}+\frac{\theta}{2}-2\eta_l^4\right)+4\left(\frac{\theta}{2}+\eta_l^2\right)^2+\frac{1}{n_g}\left(\frac{\theta}{2}+\eta_l^2\right)^2+$$
$$-\frac{4\eta_l^2}{n_g}\left[\frac{\theta}{2}+n_g\left(\frac{\theta}{2}+\eta_l^2\right)\right]+\frac{2\eta_l^2}{n_g}\left(\frac{\theta}{2}+2n_g\eta_l^2\right)-4\eta_l^2\left(\frac{\theta}{2}+\eta_l^2\right)=\frac{\theta^2}{n_g}+\frac{\theta}{4n_g}+\theta^2$$

$$\Longrightarrow \operatorname{cov}_0\left(\bar{D}_{gg'l}(t),\bar{D}_{gk'l}(t)\right)=\frac{\theta^2}{n_g}+\frac{\theta}{4n_g}.$$

Da mesma forma seguem os outros dois casos.

Assim,

$$\begin{aligned}\Upsilon_0^* &= \frac{(8\theta^2+2\theta)}{G^2(G-1)^2}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{1}{2n_g}+\frac{1}{2n_{g'}}\right)+\frac{8}{G^2(G-1)^2}\left\{\sum_{g=1}^{G}\sum_{g'>g}\sum_{k'>g'}\left(\frac{\theta^2}{n_g}+\frac{\theta}{4n_g}\right)+\right.\\ &+ \left.\sum_{g=1}^{G}\sum_{k>g}\sum_{g'>k}\left(\frac{\theta^2}{n_{g'}}+\frac{\theta}{4n_{g'}}\right)+\sum_{g=1}^{G}\sum_{g'>g}\sum_{k'>k}\left(\frac{\theta^2}{n_{g'}}+\frac{\theta}{4n_{g'}}\right)\right\}+\\ &+ \frac{(8\theta^2+2\theta)}{G^2}\sum_{g=1}^{G}\frac{1}{2n_g}-\frac{(8\theta^2+2\theta)}{G\begin{pmatrix}G\\2\end{pmatrix}}\sum_{g=1}^{G}\sum_{g'>g}\left(\frac{1}{2n_g}+\frac{1}{2n_{g'}}\right).\end{aligned}$$

□

Apêndice B

Figuras

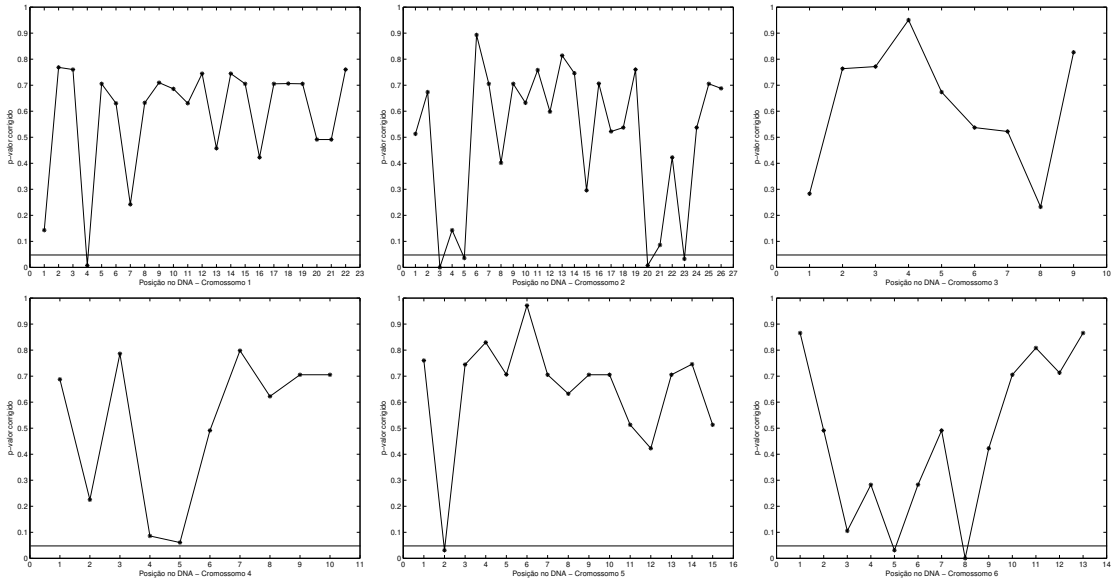

Figura B.1: Gráficos da posição no DNA versus o p-valor corrigido obtido pela distribuição assintótica para etnia, cromossomos 1 a 6.

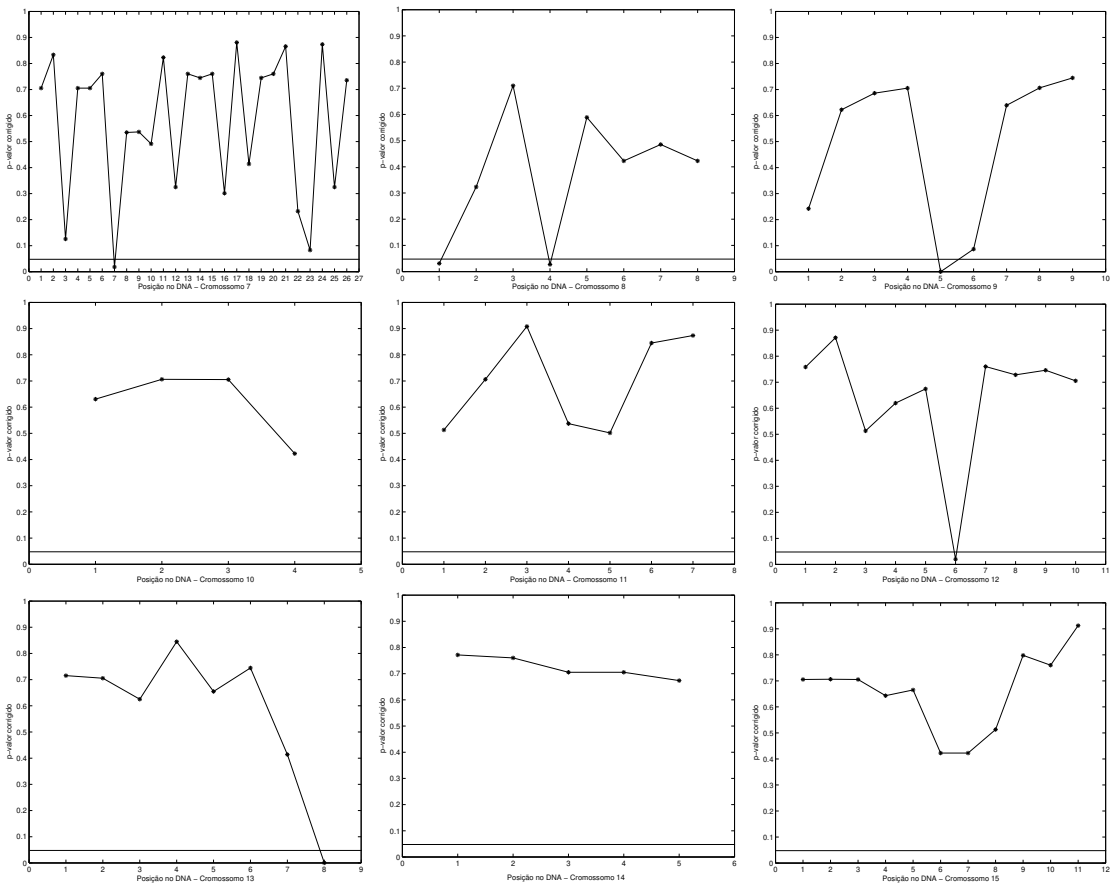

Figura B.2: Gráficos da posição no DNA versus o p-valor corrigido obtido pela distribuição assintótica para etnia, cromossomos 7 a 15.

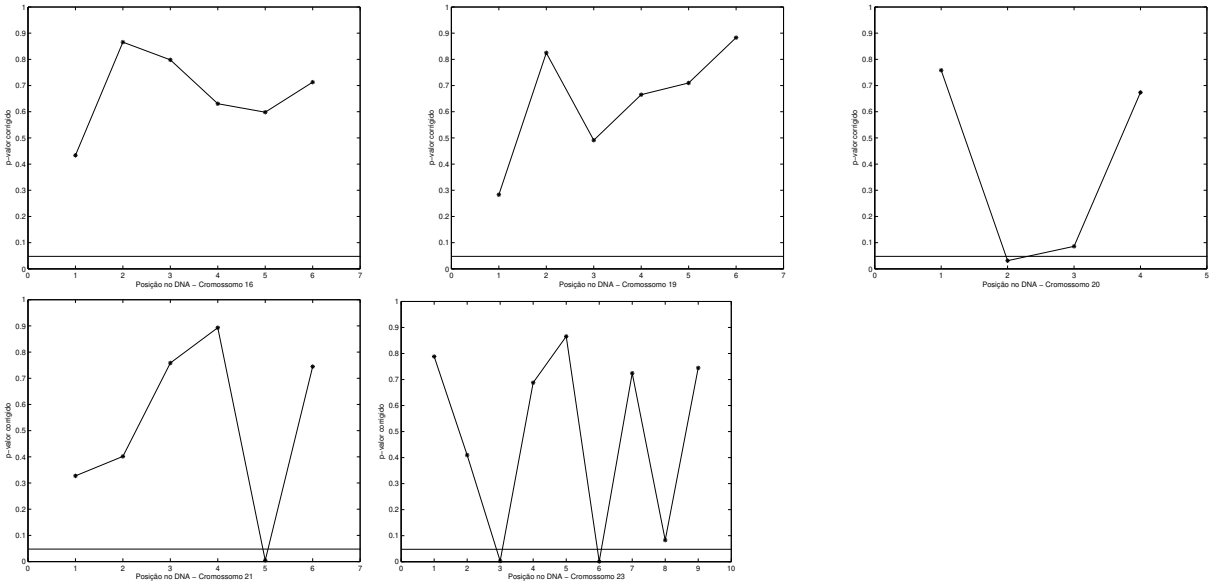

Figura B.3: Gráficos da posição no DNA versus o p-valor corrigido obtido pela distribuição assintótica para etnia, cromossomos 16, 19 a 21 e 23.

Referências Bibliográficas

Abramowitz, M. e Segun, I.A. (1972) Handbook of Mathematical Functions. Dover, New York.

Andrade, M.; Pinheiro, H.P. (2002) Métodos Estatísticos aplicados em genética humana, ABE. São Paulo-SP.

Arvesen, J.N. (1969) Jackknifing U-statistics. *Ann. Math. Statist.* **40**:2076-2100.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B* **57**:289-300.

Benjamini, Y. and Liu, W. (1999) A step-down multiple-hypothesis procedure that controls the false discovery rate under independence. *J. Statist. Planng Inf.* **82**:163-170.

Bhattacharyya, A. (1946) On a measure of divergence between two multinomial populations. *Sankhya* **7**:401-406.

Brown, A.H.D.; Marshall,D.R. e Albrecht, L. (1975) Profiles of electrophoretic alleles in natural populations. *Genet. Res.* **25**:137-143.

Bulmer, M.G.(1971) Protein polymorphism. *Nature* **234**:410-411.

Casella, G. and Berger, R. L. (2002) Statistics Inference. 2^a ed. Duxbury - Thompson Learning edition, California.

Chakraborty, R. e Rao, C.R. (1991) Measurement of genetic variation for evolutionary studies. *Handbook of statistics* Vol.**8**: 271-316.

Di Rienzo, A.; Peterson, A.C.; Garza, J.C.; Valdes, A.M.; Slatkin, M. et al. (1994) Mutational processes of simple sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* **91**:3166-3170.

Dwass, M. (1956) The large sample power of rank tests in the two samples problem. *The Annals of Mathematical statistics* **27**: 352-374.

Edenberg, H.J. , Bierut, L.J., Boyce, P. Cao, M. et al. (2005) Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single nucleodite polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genetics* **6 (Suppl 1)**: S2.

Efron, B. e Tibshirani, R.J. (1993) An introduction to the bootstrap. 1^a ed. Chapman & Hall, New York.

Ewens, W. J. (1979) Matematical population genetics. Springer-Verlag, New York.

Fisher, R.A. (1930) The genetical theory of natural selection. 1^a ed. Oxford: Clarendon Press.

Fondon, J.W. e Garner, H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *National Academy of Sciences of the USA* **101**:18058-18063.

Freimer, N.B. e Slatkin, M. (1996) Evolution and Mutational processes. *Variation in the Human Genome Ciba Foundation Symposia* **197**:51-67.

Gini, C. (1912) Variabilitá e mutabilitá, Studi Economicoaguridici della facolta di Giurisprudenza dell, Universite di Cagliari III, Part II.

Goldstein, D.B.; Linares, A.R.; Cavalli-Sforza, L.L. e Feldman, M.W. (1995) An evaluation of Genetic Distances for use with Microsatellite loci. *Genetics* **139**:463-471.

Griffiths, A.J.F., Wessler, S.R, Lewontin, R.C., Gelbart, W.M., Suzuki, D.T e Miller, J.H. (2004) An Introduction to Genetic analisys. 8^a ed. W.H.Freeman, New York.

Hartl, D.L. (2000) A primer of population genetics. 3a.ed. Massachusetts.

Hoeffding, A. (1948) A class of statistics with asymptotically normal distribution. *The Annals of Mathematical statistics* **19**: 293-325.

Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*. **7**:1-44.

James, Barry R. (2004) Probabilidade: Um curso em nível intermediário. 3^a ed. IMPA - Rio de Janeiro-RJ.

Junqueira, L.C. e Carneiro, J. (1991) Biologia Celular e Molecular. 5a. ed. S.P.

Kimura, M. (1955) Random genetic drift in multiallelic locus. *Evolution.* **9**:419-435.

Kimura, M. (1968) Evolutionary rate at molecular level. *Nature.* **217**:624-626.

Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics.* **61**:624-626.

Lehmann, El. (1951) Consistency and unbiasedness of certain nomparametric tests. *The Annals of Mathematical statistics* **22**: 165-179.

Li, C.C. (1955) Population Genetics. University of Chicago Press.

Li, W. H. (1976) Electrophoretic identy of proteins in a finite population and genetic distance between taxa. *Genet. Res.* **28**:119-127.

Mahalanobis, P.C.(1936) On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2**:49-55.

Moran, P.A.P. (1975) Wandering distribution and the electrophoretic profile. *Theor. Popul. Biol.* **8**:318-330.

Morton, N.E. (1975) Kinship, information and biological distance. *Theor. pop. Biol.* **7**:246-255.

Nei, M. (1972) Genetic Distance between populations. *Amer.Nat.* **106**:283-292.

Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Ohta, T. e Kimura, M.(1973) A model of mutation appropriate to estimate the number of eletrophoretically detectable alleles in genetic population. *Genet. Res.* **22**: 201-204.

Rao, C.R. (1982a) Gini-Simpson index of diversity: A characterization, generalization and applications. *Utilitas Mathematica* **21**:273-282.

Rao, C.R. (1982b) Diversity and dissimilarity coefficients: A unified approach. *Theor. Pop. Biol.* **21**:24-43.

Rao, C. R. (1982c) Diversity: its measurement, decomposition, apportionment and analysis. *Sankhya* **44A**:1-21.

Rao, C.R. e Boudreau, R. (1984) Diversity and cluster analyses of blood group data on some human populations. In: Chakravarti, ed., *Human Population Genetics: The Pittsburgh Sympsosium.* Van Nostrand Reinhold, New York, 331-362.

Révész, Pál. (1990) Random Walk in random and nom random environments. Singapore. Teaneck, N.J.: World Scientific.

Pinheiro, A., Pinheiro, H.P e Kiihl, S. (2006) An asymptotically normal test for Selective Neutrality Hypothesis. *in appear.*

Pinheiro, A., Pinheiro, H.P e Sen, P.K. (2007) The use of hamming distance in bioinformatics. *In Handbook of Statistics-Bioinformatics.* **26**, editor: Chakraborty Rao.

Pritchard, J.K. e Feldman, M.W. (1996) Statistics for Microsatellite Variation Based on Coalescence. *Theoretical Population Biology* **50**: 325-344.

Sen, P.K. (1960) On some convergence properties of U-statistics. *Cal.Statist.Assoc.Bull.* **10**:1-18.

Shriver, M.D.; Jin, L.; Chakraborty, R.; Boerwinkle, E. et al. (1993) VNTR Allele frequency distribuitions under the "*stepwise*" mutation model: A computer simulation approach. *Genetics* **134**:983-993.

Shriver, M.D.; Jin, L.; Chakraborty, R.; Boerwinkle, E. et al. (1995) A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Molecular Biology Evolution* **12(5)**:914-920.

Simpson, E.H. (1949) Measurement of diversity. *Nature* **163**:688.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**:457-462.

Sneath, P.H.A. e Sokal, R.P. (1973) *Numerical Taxonomy.* Freeman, San Francisco.

Valdes, A.M.; Slatkin, M. e Freimer, N.B. (1993) Allele Frequencies at Microsatellite loci. *Genetics* **133**:737-749.

Wehrhahn, C.(1975) The evolution of selectively similar electrophoretically detectable alleles in finite natural population. *Genetics* **80**:375-394.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* **16**:97-159.