Samara Flamini Kiihl

Análise Estatística de Polimorfismo Molecular em Seqüências de DNA Utilizando Informações Filogenéticas

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.

Orientadora:
Profa. Dra. Hildete Prisco Pinheiro

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

Campinas - SP

2005

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Samara Flamini Kiihl e aprovada pela comissão julgadora.

Campinas, 25 de fevereiro de 2005.

Profa. Dra. Hildete Prisco Pinheiro
Departamento de Estatística - UNICAMP
Orientador

Banca Examinadora:

1. Profa. Hildete Prisco Pinheiro (orientadora) - IMECC/UNICAMP
2. Prof. Dr. Gilberto Alvarenga Paula - IME/USP
3. Prof. Dr. Mauro Sérgio de Freitas Marques - IMECC/UNICAMP
4. Prof. Dr. Sérgio Furtado dos Reis (suplente) - IB/UNICAMP

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.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecário: Maria Júlia Milani Rodrigues – CRB8a / 2116

Kiihl, Samara Flamini

K55a Análise estatística de polimorfismo molecular em seqüências de DNA utilizando informações filogenéticas / Samara Flamini Kiihl -- Campinas, [S.P. :s.n.], 2005.

Orientadora : Hildete Prisco Pinheiro

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

1. Modelos lineares (Estatística). 2. Simulação (Computadores) – Métodos estatísticos. 3. Bioestatística. I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Statistical analysis of molecular polymorphism in DNA sequences using phylogenetic information

Palavras-chave em inglês (keywords): 1. Linear models (Statistics). 2. Simulation (Computers) – Statistical methods. 3. Biostatistics.

Área de concentração: Estatística

Titulação: Mestre em Estatística

Banca examinadora: Profa. Dra. Hildete Prisco Pinheiro (UNICAMP)
Prof. Dr. Gilberto Alvarenga Paula (USP)
Prof. Dr. Mauro Sérgio de Freitas Marques (UNICAMP)

Data da defesa: 25/02/2005

A meus pais, Rosana e José Carlos.

Agradecimentos

Aos meus pais, Rosana e José Carlos, por fornecerem todas as condições para que eu me dedicasse aos estudos com tranqüilidade.

À minha orientadora Profa. Hildete Prisco Pinheiro, pelo incentivo, dedicação e amizade.

Ao Prof. Sérgio Furtado dos Reis, pelo incentivo da aplicação da estatística nos estudos de genética, pelos esclarecimentos na teoria de biologia, pela contribuição com conjuntos de dados e pelo carinho e paciência.

Ao Prof. Aluísio Pinheiro, pelas contribuições tanto na parte teórica quanto na computacional.

Aos professores do departamento, por todo conhecimento transmitido durante estes anos no IMECC.

Aos professores Mauro Marques e Gilberto Paula, por aceitarem participar da banca examinadora, pelas correções e sugestões.

Aos colegas do IMECC, pela amizade e incentivo. Agradeço especialmente ao Benilton, pelo apoio desde os tempos da iniciação científica.

Às meninas da república, que certamente foram responsáveis por muitos momentos divertidos durante este ano.

Às amizades que fiz durante o período do meu mestrado e que desejo manter por toda a vida.

À CAPES, pelo suporte financeiro, fundamental para o desenvolvimento desse projeto.

"Statistics (...) the most important science in the whole world: for upon it depends the practical application of every other science and of every art: the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience."

Florence Nightingale - Statistician

Resumo

Variação genética no nível de nucleotídeo é uma fonte poderosa de informação para o estudo da evolução de uma população. Importantes aspectos da evolução de populações naturais têm sido investigados utilizando seqüências de nucleotídeos. A quantidade $\theta = 4N\nu$, em que N é o tamanho efetivo da população e ν é a taxa de mutação por seqüência (gene, locus) por geração, é um parâmetro essencial porque determina o grau de polimorfismo em um locus. O sucesso da inferência sobre a evolução de uma população é medido pela acurácia da estimação deste parâmetro.

Esta dissertação de mestrado apresenta diversos métodos de estimação do parâmetro θ, bem como uma comparação entre eles através de simulações e aplicações a dados reais. Utilizando informações filogenéticas de amostras de seqüências de DNA, constrói-se um modelo linear onde o coeficiente da variável independente é a estimativa do parâmetro θ. Verificou-se que utilizando informações filogenéticas dos dados obtêm-se estimadores bem mais eficientes.

Abstract

Genetic variation at the nucleotide level is a powerful source of information for studying the evolution of a population. Important aspects of the evolution of a population have been investigated by using nucleotide sequences. The quantity $\theta = 4N\nu$, where N is the effective size of the population and ν is the mutation rate per sequence (gene, locus) per generation, is an essential parameter because it determines the degree of polymorphism at the locus. The degree of success in our inference about the evolution of a population is measured to some extent by the accuracy of estimation of this essential parameter.

This work presents some methods of estimation of this parameter, comparisons between the different methods through computational simulations and applications to real data. The evolution of a species can be seen through a phylogenetic tree and a linear model can be constructed by using the phylogenetic information to estimate θ. It has been verified that the use of such information leads us to more accurate estimators of θ.

Sumário

Lista de Tabelas

Lista de Figuras

1 Introdução

1.1 Motivação

Durante os últimos dez anos um grande progresso ocorreu no estudo da evolução no nível molecular devido ao grande avanço no desenvolvimento de técnicas bioquímicas para o estudo do DNA, como a reação em cadeia de polimerase e o seqüenciamento automático de ácidos nucléicos (WATSON, 1992). Esta metodologia bioquímica tem sido utilizada recentemente em estudos de grande escala que envolvem o conhecimento do genoma completo de diferentes organismos, iniciando a era genômica da biologia molecular (GIBSON; MUSE, 2004). Esta tecnologia tem gerado uma quantidade volumosa de dados que tem, por sua vez, exigido o desenvolvimento de métodos estatísticos para análise desses dados, bem como progressos notáveis em métodos de biologia computacional (WATTERMAN, 1995).

Análise genética é possível em qualquer organismo. Por esta razão, conceitos e enfoques experimentais de genética populacional têm atraído quase todas as áreas da biologia moderna (HARTL; CLARK, 1997; NAGYLAKI, 1992). Genética populacional é o estudo de diferenças genéticas naturais entre organismos. Diferenças genéticas que são comuns entre organismos da mesma espécie são chamadas de polimorfismo genético, enquanto que diferenças genéticas acumuladas entre espécies constituem divergência genética. Desta maneira, o estudo de evolução molecular no nível de populações envolve a descrição dos padrões de polimorfismo molecular nas seqüências de DNA e a inferência das causas em termos de mecanismos e forças evolutivas (DURRETT, 2002). Os problemas estatísticos associados ao estudo da

evolução molecular são, em geral, mais complicados do que aqueles tratados na estatística tradicional e é fundamental a incorporação de modelos detalhados de mudanças evolutivas nas seqüências de DNA na dedução de métodos estatísticos eficientes para estimação e inferência.

A maioria dos problemas na análise de seqüências genômicas é essencialmente estatística. Em análise de seqüências genômicas, tipicamente encontram-se dados com um grande número de posições ou sítios e, em cada posição, tem-se uma resposta categorizada qualitativa: nucleotídeo (ou aminoácido), com quatro (ou vinte) categorias.

Há uma necessidade real de apreender os fundamentos de genética e biologia molecular da análise computacional de seqüências e, com eles, formular modelagens estatísticas e esquemas de análise. Pela natureza intrinsicamente estocástica das forças evolutivas, tais ferramentas envolveriam naturalmente a modelagem estatística de sistemas biológicos que, por sua vez, necessitariam de conceitos teóricos de probabilidade, estatística e processos estocásticos (DURRETT, 2002).

A quantidade $\theta = 4N\nu$, o número esperado de mutações na população por gene por geração, em que N é o tamanho efetivo da população e ν é a taxa de mutação por seqüência (gene, locus) por geração, é um parâmetro essencial porque determina o grau de polimorfismo em um locus (HARTL; CLARK, 1997; DURRETT, 2002). O sucesso da inferência sobre a evolução de uma população é medido pela acurácia da estimação deste parâmetro.

Apesar dos avanços nos estudos desta área, a estimação de parâmetros genéticos populacionais a partir de dados de seqüência de nucleotídeos é um campo de estudos ainda muito recente. Muitos métodos publicados ainda não foram adequadamente estudados com dados simulados ou aplicados a dados reais.

Esta dissertação de mestrado tem por objetivo apresentar estimadores mais acurados para θ através do uso de informação filogenética. No Capítulo 2 apresentam-se a teoria de coalescência e as propriedades dos estimadores para θ obtidos a partir do método dos momentos (WATTERSON, 1975; TAJIMA, 1983; FU; LI, 1993b). No

Capítulo 3 há uma breve descrição dos diferentes métodos utilizados para construção de árvores filogenéticas a partir de uma amostra. Os estimadores baseados em modelos lineares, que levam em consideração a informação da árvore genealógica, estão no Capítulo 4. Através de simulações, apresentadas no Capítulo 5, pode-se observar o comportamento dos diversos estimadores estudados. No Capítulo 5 apresenta-se também uma aplicação desses estimadores em um conjunto de dados real.

1.2 Introdução à Biologia Molecular

A informação hereditária de todos os organismos vivos, com exceção de alguns vírus, está presente nas moléculas de **DNA** (ácido desoxirribonucleico). O DNA consiste de duas cadeias complementares, cada uma formada por quatro tipos de **nucleotídeos**: adenina (A), guanina (G), timina (T) e citosina (C). Os nucleotídeos A e G são classificados como **purinas** e os demais como **piramidinas**. A localização física do nucleotídeo na seqüência de DNA é denominada **sítio** (WATSON, 1992; GRIFFITHS, 2000).

Gene é uma seqüência de DNA que é essencial para uma função específica. A totalidade de DNA numa célula é o **genoma**. Os genes são dispostos em uma ordem linear ao longo de corpúsculos filamentosos microscópicos chamados **cromossomos**. A localização física de um gene no DNA recebe o nome de **locus**. As variantes de um gene em um dado locus são chamadas de **alelos**. Assim, o gene responsável pela cor do olhos nos seres humanos possui alelos para diferentes cores. Um locus contém dois alelos, um em cada posição homóloga correspondente no cromossomo materno e paterno. Quando o alelo é dominante, ele impõe ao indivíduo a manifestação da característica por ele determinada. A **meiose** é o processo de divisão celular pelo qual as células **diplóides** de linhagem germinativa dão origem a gametas **haplóides**. A **mitose** é a divisão habitual das células somáticas (não germinativas) pelo qual o corpo cresce, se diferencia e se reconstitui. **Recombinação gênica** é um fenômeno que está intimamente ligado à meiose celular. É devido à ocorrência de recombinação que existe um aumento na variabilidade genética, conferindo igual

variação aos descendentes de uma espécie formados a partir dessas células. A recombinação baseia-se em quebras que ocorrem enquanto os cromossomos homólogos estão emparelhados, sendo que estas quebras sempre atingem duas **cromátides** irmãs em pontos correspondentes. Na recombinação, os alelos apenas trocam de posição dentro do par de cromossomos homólogos, de modo que a estrutura e a função cromossômica permanecem inalteradas. Este processo não deve ser confundido com mutação. Na formação de uma gameta, os dois homólogos são copiados de cada par de cromossomos. Na distribuição de cromossomos homólogos, a seleção de qualquer um deles proveniente do pai ou da mãe para uma célula filha é aleatória. Quando os pares de cromossomos homólogos alinham-se, pode ocorrer um processo chamado de **crossing-over**, o qual resulta na recombinação genética.

O DNA pode originar duas ou mais moléculas no futuro. Em uma avaliação retrospectiva, duas ou mais moléculas de DNA podem juntar-se em uma única cópia. Este evento de junção é denominado **coalescência**. O processo de coalescência é um modelo estocástico adequado à descrição do comportamento evolutivo de uma amostra de seqüências de DNA (DURRETT, 2002). É considerado um dos desenvolvimentos mais importantes da modelagem genética de populações. Anteriormente, a grande maioria das simulações era realizada através de procedimentos prospectivos, onde se procurava descrever a variabilidade genética no tempo $t + 1$ a partir das informações disponíveis no tempo t. Estes modelos apresentavam uma série de restrições práticas que dificultavam a sua utilização, como por exemplo, a dificuldade de se realizar previsões a longo prazo. O processo de coalescência, por outro lado, lida com o tempo retrospectivamente, utilizando as informações disponíveis no presente para testar hipóteses sobre o passado evolutivo de uma população. Uma das restrições dos métodos coalescentes é o fato de que linhagens de genes extintas ao longo do tempo não são consideradas na análise, já que do ponto de vista presente não são observáveis. Além disso, a partir do momento em que todas as linhagens de genes coalescem em uma única cópia, não é possível a obtenção de informações sobre eventos anteriores.

1.2.1 Modelos evolutivos

Mutação e Teoria de neutralidade

A mutação é o resultado de mudanças químicas na estrutura do DNA. Estas modificações podem ocorrer de forma espontânea ou através de agentes mutagênicos. Apesar de ocorrerem espontaneamente a uma taxa baixa, esta é uma das mais importantes fontes de variação molecular. A mutação tem origem quando há adição ou subtração de bases de nucleotídeos e quando há substituição de bases. A substituição de uma purina (adenina e guanina) por outra purina, ou de uma piramidina (citosina e timina) por outra piramidina é denominada de **transição**. A substituição de uma purina por uma piramidina, ou vice-versa é denominada de **transversão**. Existem também as chamadas **mutações silenciosas**, que são aquelas que não modificam o aminoácido. Por exemplo, os códons GCA e GCC codificam ambos o aminoácido arginina. Portanto, a mutação de A para G na última posição deste códon é dita silenciosa.

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 pequena. Sendo assim, as mutações deverão ocorrer apenas em **sítios selvagens**, ou seja, em sítios que não sofreram mutações. Sítios que sofreram mutações são denominados **sítios mutantes**.

Kimura (1968) sugere que parte da variação observada em nível molecular é seletivamente neutra. Desta forma, o destino de alelos neutros é, em grande parte, governado pela deriva genética aleatória.

A hipótese de que os polimorfismos observados em alelos neutros são resultantes do balanceamento entre os processos de mutação e deriva genética aleatória é conhecida como hipótese de neutralidade. Desta forma, o aumento de variabilidade ocasionado pela ocorrência de mutações é compensado pelo efeito da deriva genética aleatória, que tende a retirar a variabilidade do sistema, fixando ou extinguindo

alelos. Quando esta hipótese é verdadeira, grande parte do polimorfismo observado pode ser entendido como um mero ruído evolucionário, cujo impacto sobre a adaptação dos organismos ao ambiente não é significativo.

Deriva genética aleatória

Hardy e Weinberg propuseram um modelo para predizer as freqüências gênicas em uma população. Suposições do **Equilíbrio de Hardy-Weinberg**:

- Acasalamentos aleatórios
- Tamanho populacional muito grande
- Não há migração entre populações
- Não há mutação genética
- Não há efeito de seleção natural

Considerando apenas o efeito da aleatoridade, as freqüências gênicas são mantidas constantes ao longo das gerações.

Modelo de Wright-Fisher

Fisher (1930) e Wright (1931) desenvolveram um modelo, conhecido como Wright-Fisher. Este modelo tem muitas da suposições do equilíbrio de Hardy-Weinberg, com a exceção da suposição do tamanho da população, que aqui é constante e igual a N. Assume-se também neutralidade, ou seja, não existe diferenças seletivas entre os alelos.

Seleção Natural

O principal fator evolutivo responsável pela ocorrência de desvios a partir do equilíbrio de Hardy-Weinberg é a seleção natural, processo através do qual os genótipos com melhor ajuste ao ambiente deixam, em média, mais descendentes do

que aqueles menos adaptados. A seleção natural tende a diminuir a variabilidade genética.

Gargalo Populacional

É definido como a ocorrência de uma redução dramática no tamanho de uma população, que pode ser ocasionada por diversos fatores como mudanças ambientais. Populações submetidas a este efeito podem recuperar o seu tamanho típico após certo período, no entanto, devido ao processo de deriva genética as mudanças ocasionadas nas freqüências dos alelos tendem a afetar permanentemente a evolução de toda a espécie.

2 *Estimadores de Polimorfismo Molecular*

2.1 Introdução

O sucesso na inferência sobre a evolução de uma população é medido, de certa maneira, pela acurácia na estimação do parâmetro θ que será definido neste Capítulo. O objetivo deste Capítulo é mostrar estimadores de θ sob as condições de neutralidade do modelo de Wright-Fisher sem recombinação e subdivisão populacional.

Existem dois estimadores de θ que são mais utilizados para uma amostra de n seqüências de DNA de uma população: $\mathcal{T}_1$, baseado no número de sítios segregantes, que é discutido na Seção 2.2 e $\mathcal{T}_2$, baseado no número de diferenças em pares de nucleotídeos, discutido na Seção 2.3. Outro estimador de θ é $\mathcal{T}_3$, baseado no número de *singletones*, apresentado na Seção 2.4. Todos estes estimadores são não viesados sob a condição de sítios infinitos do modelo de neutralidade de Wright-Fisher, ou seja, sob a suposição de que a população evolui de acordo com o modelo de Wright-Fisher com um tamanho efetivo da população constante, grande número de sítios nas seqüências, de tal forma que toda mutação ocorra em um novo sítio e que todas as mutações sejam seletivamente neutras. Estimar θ usando estes estimadores é computacionalmente simples. Contudo, o preço pela simplicidade computacional é uma variância grande.

Suponha que tem-se uma amostra de 5 seqüências de DNA, das quais foram seqüenciados 500 sítios. A Tabela 2.1 apresenta apenas os 16 sítios polimórficos ou

Tabela 2.1: Sítios polimórficos em uma amostra de cinco genes

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

* : diferenças em pares

segregantes, ou seja, aqueles que apresentam nucleotídeos diferentes no sítio. Dentre os sítios polimórficos, tem-se que os sítios 3, 5, 6, 7, 8, 11, 12, 13 e 15 são *singletones*, pois apresentam apenas um nucleotídeo diferente dos demais.

2.2 Polimorfismo de Nucleotídeos

O trabalho de Watterson (1975) é fundamental para a compreensão dos modelos e métodos de estimação da quantidade e distribuição de polimorfismo de nucleotídeos em seqüências de DNA. Watterson (1975) modelou a distribuição de probabilidade para o número de sítios segregantes em uma amostra de gametas, ou seja, de indivíduos haplóides.

Considera-se um gene funcional, aquele que consiste de um grande número de sítios de nucleotídeos. As pressuposições impostas por Watterson (1975) são as seguintes:

- A recombinação de informação genética devido ao *crossing over* e quebra é rara.
- Mutações podem ocorrer em qualquer sítio de nucleotídeos.
- O número de sítios é ilimitado.
- As mutações que ocorrem em gametas diferentes são mutuamente independentes.

- Um gameta herda todos seus sítios de um gameta paterno.
- Como existem ilimitados sítios, assume-se que duas mutações nunca ocorrem no mesmo sítio (nem em gametas diferentes). Portanto, em cada sítio, existem dois tipos de nucleotídeos possíveis: o tipo selvagem e o tipo mutante.
- Para N indivíduos diplóides, o número de gametas na população é $2N$.

Como tem-se ilimitados sítios e cada sítio pode estar sujeito à mutação durante a meiose, o *número total de sítios mutantes* por gene por geração pode ser considerado como uma variável aleatória com distribuição de Poisson com esperança ν. A distribuição de Poisson é bem coerente para estudar o fenômeno de interesse: sítio mutante, pois este é um fenômeno raro.

2.2.1 Distribuição do Número de Sítios Segregantes

Considere os sítios de cada um dos $2N$ gametas na população. Se uma amostra de i gametas é escolhida aleatoriamente, denota-se por V_i *o número de sítios segregantes da amostra*, ou seja, sítios polimórficos nos quais ambos os tipos de nucleotídeos (selvagem e mutante) estão presentes. Assim, V_{2N} denota o número de sítios segregantes em toda a população e V_2 denota o número de sítios segregantes (heterozigotos) em dois gametas. Este último número pode ser interpretado como os dois gametas escolhidos aleatoriamente para formar um indivíduo diplóide.

Se o gene de um particular gameta obteve um número de sítios mutantes, distribuído de acordo com a distribuição de Poisson com esperança ν, durante a meiose, então a probabilidade de que pelo menos um sítio seja segregante, u, é a *taxa de mutação do gene*, $u = 1 - e^{-\nu} \approx \nu$, se ν for pequeno. De fato:

Seja Y uma variável aleatória representando o número de sítios segregantes

$$Y \sim \text{Poisson}(\nu) \quad P(Y = k) = e^{-\nu}\frac{\nu^k}{k!}\,.$$

$$u = P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-\nu}\frac{\nu^0}{0!} = 1 - e^{-\nu}\,.$$

Decompondo em Série de Taylor:

$$\begin{aligned} e^{-\nu} &= e^{-0} + (-e^{-0})\nu + \frac{e^{-0}(\nu^2)}{2} + \ldots \approx 1 - \nu \\ \Rightarrow 1 - e^{-\nu} &\approx 1 - 1 + \nu = \nu \,. \end{aligned}$$

Introduz-se agora o parâmetro θ, que representa o número esperado de mutações na população, por gene por geração.

$$\theta = 4Nu \approx 4N\nu \,, \tag{2.1}$$

As aproximações dos resultados são obtidas fixando θ, mas considerando grandes valores para N e pequenos valores de u e ν. Assume-se também que as populações atingiram a distribuição de regime.

2.2.2 Momentos do Número de Sítios Segregantes

Quando uma amostra de i gametas é escolhida ao acaso de uma população de $2N$ gametas da geração t, denota-se J como o *número de gametas paternos distintos na amostra*; J é uma variável aleatória restrita por $1 \leq J \leq i$. Quando $J = 1$, todos os gametas amostrados têm o mesmo ancestral, quando $J = i$, cada um dos gametas têm pais distintos. A distribuição de probabilidade de J é dada por:

$$P(J = j \mid i) = G_{i,j} \quad j = 1, 2, 3, \ldots, i \tag{2.2}$$

em que $G_{i,j}$ é a *probabilidade de existirem j gametas de pais diferentes em uma amostra de tamanho i.*

Seja $V_i^{(t)}$ o *número de sítios segregantes (heterozigotos) observados nos i gametas amostrados na geração t.* Assumindo que os J pais distintos foram escolhidos aleatoriamente da geração $t-1$, eles possuíam $V_j^{(t-1)}$ sítios segregantes. Assim:

$$V_i^{(t)} = V_j^{(t-1)} + X_i^{(t)}, \tag{2.3}$$

em que $X_i^{(t)}$ é o *número de novos sítios segregantes na prole*, devido às mutações.

Assume-se que, para $i \geq 2$, $X_i^{(t)}$ tem distribuição de Poisson($i\nu$).

Usando funções geratrizes de probabilidades, pode-se escrever

$$\begin{aligned} \mathrm{E}(s^{V_i^{(t)}}) &= \mathrm{E}(s^{V_j^{(t-1)}+X_i^{(t)}}) = \mathrm{E}(s^{V_j^{(t-1)}} s^{X_i^{(t)}}) \\ &= \sum_{j=1}^{i} G_{i,j}\mathrm{E}(s^{V_j^{(t-1)}})e^{i\nu(s-1)}, \quad i \geq 2\,. \end{aligned} \tag{2.4}$$

Deve-se observar que uma amostra de um gameta não tem sítios segregantes, pois o *crossing-over* não ocorre com apenas um gameta. Portanto: $V_1^{(t)} = 0$ e $\mathrm{E}(s^{V_1^{(t)}}) = 1$. Quando a estacionaridade do processo é atingida, pode-se ignorar os t's:

$$\mathrm{E}(s^{V_i}) = \sum_{j=1}^{i} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)} \quad i \geq 2\,. \tag{2.5}$$

Como $\mathrm{E}(s^{V_1}) = 1$, pode-se escrever a função geratriz de probabilidades para o número, V_2, de sítios heterozigotos em um diplóide formado pela união de dois gametas escolhidos ao acaso da seguinte forma:

$$\begin{aligned} \mathrm{E}(s^{V_2}) &= \sum_{j=1}^{2} G_{2,j}\mathrm{E}(s^{V_j})e^{2\nu(s-1)} = G_{2,1}\mathrm{E}(s^{V_1})e^{2\nu(s-1)} + G_{2,2}\mathrm{E}(s^{V_2})e^{2\nu(s-1)} \\ \Rightarrow \quad & \mathrm{E}(s^{V_2})(1 - G_{2,2}e^{2\nu(s-1)}) = G_{2,1}e^{2\nu(s-1)} \\ \Rightarrow \quad & \mathrm{E}(s^{V_2}) = \frac{G_{2,1}e^{2\nu(s-1)}}{1 - G_{2,2}e^{2\nu(s-1)}}\,, \end{aligned} \tag{2.6}$$

em que $G_{2,1} = 1/2N$ e $G_{2,2} = 1 - 1/2N$. A probabilidade de amostrar cada gameta é $\frac{1}{2N}$. A probabilidade de amostrar dois gametas específicos iguais é $\frac{1}{2N}\frac{1}{2N} = \frac{1}{4N^2}$. Tem-se $2N$ possíveis gametas, logo, a probabilidade de retirar dois gametas iguais quaisquer é: $\frac{1}{4N^2}2N = \frac{1}{2N}$.

A equação (2.6) mostra que V_2 tem uma distribuição Poisson-Geométrica. Para mostrar isto, utiliza-se a seguinte proposição:

Proposição 1. *Sejam N, X_1, X_2, ... variáveis aleatórias inteiras não-negativas. Suponha que N seja independente de X_1, X_2, ... e que X_1, X_2, ... seja uma seqüência*

de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), condicionalmente a N. Defina a seguinte variável aleatória:

$$S_N = X_1 + X_2 + ... + X_N \,.$$

A função geratriz de probabilidades de S_N pode ser escrita como:

$$\phi_{S_N}(t) = \phi_N(\phi_{X_1}(t))$$

em que ϕ_N e ϕ_{X_1} são as funções geratrizes de probabilidades de N e X_1, respectivamente.

Demonstração.

$$\begin{aligned}\phi_{S_N}(t) &= \mathrm{E}(t^{S_N}) = \mathrm{E}(t^{\sum_{i=1}^{N} X_i}) = \mathrm{E}[\mathrm{E}(t^{\sum_{i=1}^{N} X_i} \mid N)] = \mathrm{E}\left[\prod_{i=1}^{N} \mathrm{E}(t^{X_i} \mid N)\right] \\ &= \mathrm{E}\{\underbrace{[\mathrm{E}(t^{X_1} \mid N)]}_{s=\phi_{X_1}(t)}^{N}\} = \mathrm{E}(s^N) = \phi_N(s) = \phi_N[\phi_{X_1}(t)] \,.\end{aligned}$$

□

Esta proposição será usada para definir a distribuição exata de V_2. Sejam $X_1, X_2, \ldots$, variáveis aleatórias independentes com distribuição Poisson(2ν), $\mathrm{E}(s^X) = e^{2\nu(s-1)}$. Seja M uma variável com distribuição Geométrica$(G_{2,1})$:

$$P(M = m) = G_{2,1} G_{2,2}^{m-1} = G_{2,1}(1 - G_{2,1})^{m-1} \quad m = 1, 2, 3, \ldots \tag{2.7}$$

e a função geratriz de probabilidade é:

$$\begin{aligned}\phi_M(s) &= \mathrm{E}(s^M) = \sum_{k=1}^{\infty} s^k (1 - G_{2,1})^{k-1} G_{2,1} = s\, G_{2,1} \underbrace{\sum_{l=0}^{\infty} [s(1 - G_{2,1})]^l}_{\text{série geométrica}} \\ &= \frac{s\, G_{2,1}}{1 - (1 - G_{2,1})s} = \frac{s\, G_{2,1}}{1 - s\, G_{2,2}}, \quad \text{se } \mid s \mid < \frac{1}{1 - G_{2,1}} \,.\end{aligned}$$

Então,

$$V_2 = X_1 + X_2 + \ldots + X_M \tag{2.8}$$

tem distribuição dada por (2.6).

Interpreta-se M como *o número de gerações decorridas desde que os dois gametas pudessem ser remontados para um ancestral comum.* Assim, X_i será *o número de mutações na i-ésima geração anterior.* Desta maneira, V_2 é a soma de todas as mutações ocorridas até a M-ésima geração anterior.

A solução de (2.5) pode ser obtida teoricamente através de:

$$\mathrm{E}(s^{V_i}) = \frac{\sum_{j=1}^{i-1} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)}}{1 - G_{i,i}e^{iv(s-1)}}, \quad i \geq 2\,. \tag{2.9}$$

Demonstração. De (2.5), tem-se:

$$\begin{aligned}
\mathrm{E}(s^{V_i}) = \sum_{j=1}^{i} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)} &= \sum_{j=1}^{i-1} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)} + G_{i,i}\mathrm{E}(s^{V_i})e^{i\nu(s-1)} \\
\mathrm{E}(s^{V_i}) - G_{i,i}\mathrm{E}(s^{V_i})e^{i\nu(s-1)} &= \sum_{j=1}^{i-1} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)} \\
\mathrm{E}(s^{V_i})[1 - G_{i,i}e^{i\nu(s-1)}] &= \sum_{j=1}^{i-1} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)} \\
\mathrm{E}(s^{V_i}) &= \frac{\sum_{j=1}^{i-1} G_{i,j}\mathrm{E}(s^{V_j})e^{i\nu(s-1)}}{1 - G_{i,i}e^{i\nu(s-1)}}\,.
\end{aligned}$$

□

Para o modelo, cada um dos gametas foi produzido por uma escolha aleatória de gameta paterno (com reposição, pois a população de interesse é grande). Assim, um gameta particular tem um número de filhos distribuído segundo uma Binomial$(2N, \frac{1}{2N})$. O número com que cada gameta ocorre na amostra segue uma distribuição Binomial$(i, \frac{1}{2N})$.

A probabilidade de que j pais distintos tenham sido usados para produzir i

gametas é:

$$G_{i,j} = \frac{j}{2N} G_{i-1,j} + \frac{(2N-j+1)}{2N} G_{i-1,j-1}, \quad 2 \leq j \leq i\,, \tag{2.10}$$

em que

$$G_{1,1} = 1, \qquad G_{1,i} = \frac{(2N)!}{(2N)^i (2N-i)!}, \qquad G_{i,1} = \frac{1}{(2N)^{i-1}}\,. \tag{2.11}$$

A partir de agora, assume-se que $\theta \approx 4N\nu = O(1)$ para $2N$ grande, então, para valores moderados de i, tem-se:

$$e^{i\nu(s-1)} \approx 1 + \left(\frac{i\theta}{4N}\right)(s-1) \approx 1. \tag{2.12}$$

Definição 1. *$a_n = O(b_n)$ se existe um número K positivo e finito e um inteiro $n(K)$, tal que*

$$\mid a_n/b_n \mid \leq K, \quad n \geq n(K)\,.$$

Em particular, $a_n = O(1)$ significa que $\mid a_n \mid \leq K$, para algum K positivo e finito quando n é grande, isto é, $\{a_n\}$ é eventualmente limitada.

Definição 2. *$a_n = o(b_n)$ se para algum $\epsilon > 0$ existe um inteiro positivo $n(\epsilon)$, tal que*

$$\mid a_n/b_n \mid \leq \epsilon, \quad n \geq n(\epsilon)\,.$$

Em particular, $a_n = o(1)$ significa que $a_n \to \infty$ quando $n \to \infty$.

Se $4N\nu \approx \theta \Rightarrow i\nu(s-1) \approx \dfrac{i\theta(s-1)}{4N}$. Para demonstrar (2.12), usa-se a decomposição em Série de Taylor:

$$\begin{aligned} e^{i\nu(s-1)} &= e^0 + \frac{i(s-1)e^0\nu}{1!} + \frac{i^2(s-1)^2\nu^2 e^0}{2!} + \ldots \\ &\approx 1 + i\nu(s-1) + \frac{i^2 \overbrace{\nu^2}^{\searrow 0}(s-1)^2}{2} \approx 1 + i\nu(s-1) = 1 + \underbrace{\frac{i\theta(s-1)}{\underbrace{4N}_{\nearrow\infty}}}_{\searrow 0} \approx 1\,. \end{aligned}$$

Usando (2.6) e (2.11) para $2N$ grande, obtém-se a função geratriz de probabilidade aproximada de V_2:

$$\mathrm{E}(s^{V_2}) \approx \frac{1}{1+\theta-\theta s}\,. \tag{2.13}$$

Demonstração.

$$\begin{aligned}\mathrm{E}(s^{V_2}) &= \frac{G_{2,1}e^{2\nu(s-1)}}{1-G_{2,2}e^{2\nu(s-1)}} \approx \frac{G_{2,1}[1+\frac{\theta(s-1)}{2N}]}{1-G_{2,2}[1+\frac{\theta(s-1)}{2N}]} = \frac{\frac{1}{2N}[1+\frac{\theta(s-1)}{2N}]}{1-(1-\frac{1}{2N})[1+\frac{\theta(s-1)}{2N}]} \\ &\approx \frac{1}{1-\theta s+\theta}\,.\end{aligned}$$

□

Utilizando a função geratriz de probabilidade (2.5), obtém-se os momentos exatos para V_2:

$$\begin{aligned}\mathrm{E}(s^{V_2}) &= \frac{G_{2,1}e^{2\nu(s-1)}}{1-G_{2,2}e^{2\nu(s-1)}} = \frac{\frac{1}{2N}e^{2\nu(s-1)}}{1-(1-\frac{1}{2N})e^{2\nu(s-1)}} \\ \Rightarrow \mathrm{E}(s^{V_2}) &= \frac{e^{2\nu(s-1)}}{2N-2Ne^{2\nu(s-1)}+e^{2\nu(s-1)}} = \phi_{V_2}(s)\,.\end{aligned} \tag{2.14}$$

Derivando a função geratriz de probabilidade de V_2, dada em (2.14), tem-se:

$$\begin{aligned}\phi^{(1)}_{V_2}(s) &= \frac{2\nu e^{2\nu(s-1)}[2N-e^{2\nu(s-1)}(2N-1)]-e^{2\nu(s-1)}[-2N2\nu e^{2\nu(s-1)}+2\nu e^{2\nu(s-1)}]}{[2N-e^{2\nu(s-1)}(2N-1)]^2}\,, \\ \phi^{(1)}_{V_2}(s) &= \frac{4N\nu e^{2\nu(s-1)}}{[2N-e^{2\nu(s-1)}(2N-1)]^2}\,, \\ \phi^{(1)}_{V_2}(1) &= 4N\nu = \mathrm{E}(V_2)\,.\end{aligned}$$

Utilizando a aproximação da função geratriz de probabilidades dada em (2.13):

$$\begin{aligned}\phi_{V_2}(s) = \mathrm{E}(s^{V_2}) &\approx \frac{1}{1+\theta-\theta s}\,. \\ \mathrm{E}(V_2) = \phi^{(1)}_{V_2}(1) &\approx \left.\frac{0-1(-\theta)}{(1+\theta-\theta s)^2}\right|_{s=1} = \frac{\theta}{(1+\theta-\theta)^2} = \theta\,.\end{aligned}$$

Assim, tem-se:

$$\mathrm{E}(V_2) = 4N\nu \approx \theta\,. \tag{2.15}$$

Calcula-se agora a variância exata de V_2, utilizando (2.14). Para tanto, utiliza-se a seguinte propriedade da função geratriz de probabilidade:

$$\begin{aligned}
\mathrm{Var}(V_2) &= \phi_{V_2}^{(2)}(1) + \phi_{V_2}^{(1)}(1)(1 - \phi_{V_2}^{(1)}(1))\,. \\
\phi_{V_2}^{(2)}(s) &= \frac{8N\nu e^{2\nu(s-1)}[2N - e^{2\nu(s-1)}(2N-1)][\nu - (-4N\nu e^{2\nu(s-1)} + 2\nu e^{2\nu(s-1)})]}{[2N - e^{2\nu(s-1)}(2N-1)]^4}\,. \\
\phi_{V_2}^{(2)}(s)\Big|_{s=1} &= 8N\nu^2 - 8N\nu(2\nu - 4N\nu) \\
\mathrm{Var}(V_2) &= 8N\nu^2 - 8N\nu(2\nu - 4N\nu) + 4N\nu(1 - 4N\nu) \\
&= 4N\nu(1 + 4N\nu - 2\nu)\,.
\end{aligned}$$

Pode-se calcular também a variância aproximada de V_2, utilizando (2.13):

$$\begin{aligned}
\phi_{V_2}(s) &\approx \frac{1}{1+\theta-\theta s}\,. \\
\phi_{V_2}^{(1)}(s) &\approx \frac{0-(-\theta)}{(1+\theta-\theta s)^2} = \frac{\theta}{(1+\theta-\theta s)^2}\,. \\
\phi_{V_2}^{(2)}(s) &\approx \frac{0-\theta(-2\theta-2\theta^2+2\theta^2 s)}{(1+\theta-\theta s)^4}\,. \\
\mathrm{Var}(V_2) &= \phi_{V_2}^{(2)}(1) + \phi_{V_2}^{(1)}(1)(1-\phi_{V_2}^{(1)}(1)) \\
&\approx \frac{-\theta(-2\theta-2\theta^2+2\theta^2)}{(1+\theta-\theta)^4} + \theta(1-\theta) = \theta^2 + \theta\,.
\end{aligned}$$

Assim,

$$\mathrm{Var}(V_2) = 4N\nu(1 + 4N\nu - 2\nu) \approx \theta(1+\theta) = \theta^2 + \theta\,. \tag{2.16}$$

O conceito de coalescência foi apresentado no Capítulo 1. Quando retorna-se às gerações anteriores, as linhagens dos alelos se juntam ou passam por coalescência, onde os alelos dividem um ancestral comum.

Considere i alelos. A probabilidade de que exista coalescência na geração imedi-

atamente anterior deve ser um menos a probabilidade de que não exista coalescência na geração imediatamente anterior. Sabe-se que a probabilidade de que dois alelos tenham ancestrais distintos na geração anterior é $G_{2,2} = 1-1/2N = (2N-1)/(2N)$. A probabilidade de que um terceiro alelo tenha ancestral diferente dos outros dois alelos é $(2N-2)/(2N)$, porque uma vez que os ancestrais distintos dos dois primeiros alelos são escolhidos, restam $(2N-2)$ ancestrais distintos a serem escolhidos. Como os eventos são independentes, a probabilidade de que i alelos tenham i ancestrais distintos é:

$$\begin{aligned}\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\ldots\left(1-\frac{i-1}{2N}\right) &= \prod_{j=1}^{i-1}\left(1-\frac{j}{2N}\right)\\ &= 1-\frac{1}{2N}\underbrace{(1+2+\ldots+i-1)}_{\frac{(i-1)(i-1+1)}{2}}+o(N^{-2})\\ &\approx 1-\frac{i(i-1)}{4N}\,.\end{aligned}$$

Assim, $1-\dfrac{i(i-1)}{4N}$ é a probabilidade de que não exista coalescência na geração imediatamente anterior, ou seja, é a probabilidade de que i gametas apresentem i ancestrais distintos. Voltando à notação do modelo:

$$G_{i,\,i} \approx 1-\frac{i(i-1)}{4N}\,. \tag{2.17}$$

E a probabilidade de que exista coalescência, ou seja, de que i alelos apresentem $i-1$ ancestrais distintos é:

$$G_{i,\,i-1} = 1-G_{i,\,i} \approx \frac{i(i-1)}{4N}\,. \tag{2.18}$$

Para i alelos, a probabilidade de que não haja coalescência para as primeiras $t-1$ gerações seguida de coalescência na t-ésima geração é: $(G_{i,\,i})^{t-1}(1-G_{i,\,i-1})$, ou seja, é a probabilidade de que i alelos provenham de $i-1$ alelos, t gerações atrás e que a divergência tenha ocorrido há $t-1$ gerações.

O conceito de coalescência é extremamente importante e será utilizado no estudo

do comprimento dos ramos de uma árvore genealógica (ver Seção 2.3.1.1).

Pode-se expressar V_i, o número de sítios segregantes numa amostra de i gametas, como a soma de $i-1$ variáveis aleatórias independentes, para pequenos valores de i:

$$V_i \approx Y_1 + Y_2 + \ldots + Y_{i-1}\,, \tag{2.19}$$

em que

$$P(Y_j = n) \;=\; \left(\frac{1}{1+\frac{\theta}{j}}\right)\left(1-\frac{1}{1+\frac{\theta}{j}}\right)^n, \quad n = 0,1,2,3,\ldots \tag{2.20}$$

Y_j é interpretado como o número de novas mutações ocorrendo nas gerações anteriores quando haviam exatamente $j+1$ ancestrais distintos presentes. Veja que Y_j assume valores: 0, 1, 2, 3, ..., então:

$$\begin{aligned}
Y_j &= X_j - 1, \quad X_j \sim \text{Geométrica}\left(\frac{1}{1+\frac{\theta}{j}}\right).\\
P(X_j = n) &= \left(\frac{1}{1+\frac{\theta}{j}}\right)\left(1-\frac{1}{1+\frac{\theta}{j}}\right)^{n-1}, \quad n = 1,2,3,\ldots.\\
\mathrm{E}(X_j) &= \left(\frac{1}{1+\frac{\theta}{j}}\right) = 1+\frac{\theta}{j}\,.\\
\mathrm{E}(Y_j) &= \mathrm{E}(X_j-1) = \mathrm{E}(X_j) - 1 = \frac{\theta}{j}\,.\\
\mathrm{Var}(Y_j) = \mathrm{Var}(X_j) &= \frac{1-\frac{1}{1+\theta/j}}{\left(\frac{1}{1+\theta/j}\right)^2} = \frac{\theta/j}{1+\theta/j}(1+\theta/j)^2 = \frac{\theta}{j} + \frac{\theta^2}{j^2}\,.
\end{aligned} \tag{2.21}$$

E a função geratriz de probabilidades de Y_j é:

$$\begin{aligned}
\phi_{Y_j}(s) = \mathrm{E}(s^{Y_j}) &= \mathrm{E}(s^{X_j - 1}) = \mathrm{E}(s^{X_j} s^{-1}) = \frac{1}{s}\mathrm{E}(s^{X_j}) = \frac{1}{s}\phi_{X_j}(s) \\
&= \frac{1}{s}\sum_{k=1}^{\infty} s^k \left(\frac{\theta}{j+\theta}\right)^{k-1} \frac{j}{j+\theta} = \frac{1}{s}\left(\frac{sj}{j+\theta}\right) \underbrace{\sum_{l=0}^{\infty}\left(\frac{s\,\theta}{j+\theta}\right)^{l}}_{\text{série geométrica}} \\
&= \left(\frac{j}{j+\theta}\right)\left(\frac{1}{1-\frac{s\,\theta}{j+\theta}}\right) = \left(\frac{j}{j+\theta}\right)\left(\frac{j+\theta}{j+\theta-s\,\theta}\right) \\
&= \frac{1}{1+\left[\dfrac{\theta(s-1)}{j}\right]} \quad \text{, se} \quad \frac{s\,\theta}{j+\theta} < 1 \Leftrightarrow |s| < \frac{j+\theta}{\theta}\,.
\end{aligned}$$

Deve-se observar que V_2 em (2.8) e $Y_1 \approx V_2 = X_1 + X_2 + \ldots + X_M$, têm essencialmente a mesma interpretação e distribuição. Assim, (2.20) é aproximadamente correta. A razão de (2.19) ser apenas aproximada é que, para o modelo, pode ser que nenhuma geração na amostra tenha exatamente $j+1$ ancestrais e existe uma dependência entre o número de ramificações ocorrendo sucessivamente na árvore filogenética.

Precisa-se agora calcular os momentos da variável aleatória V_i, *número de sítios segregantes em uma amostra de i gametas*.

$$\begin{aligned}
\mathrm{E}(s^{V_i}) &\approx \mathrm{E}\left(s^{\sum_{j=1}^{i-1}}\right) = \prod_{j=1}^{i-1}\mathrm{E}(s^{Y_j}) \\
\Rightarrow \mathrm{E}(s^{V_i}) &\approx \prod_{j=1}^{i-1}[1+\theta(s-1)/j]^{-1}\,. \qquad (2.22)
\end{aligned}$$

Através da função geratriz de probabilidades, obtêm-se os momentos da variável

aleatória V_i:

$$\begin{aligned}
\mathrm{E}(V_i) = \phi_{V_i}^{(1)}(1) &\approx \frac{\partial}{\partial s}\left[\prod_{j=1}^{i-1}[1+\theta(s-1)/j]^{-1}\right]_{s=1} \\
\Rightarrow \mathrm{E}(V_i) &\approx \theta\sum_{j=1}^{i-1}\frac{1}{j}\,. \\
\mathrm{Var}(V_i) &= \phi_{V_i}^{(2)}(1) + \phi_{V_i}^{(1)}(1)(1-\phi_{V_i}^{(1)}(1)) \\
\Rightarrow \mathrm{Var}(V_i) &\approx \theta\sum_{j=1}^{i-1}\frac{1}{j} + \theta^2\sum_{j=1}^{i-1}\frac{1}{j^2}\,.
\end{aligned}$$

Pode-se encontrar mais facilmente os momentos da variável aleatória V_i usando apenas sua estrutura de soma de variáveis aleatórias independentes: $V_i \approx Y_1 + Y_2 + \ldots + Y_{i-1}$

$$\mathrm{E}(V_i) \approx \mathrm{E}\left(\sum_{j=1}^{i-1} Y_j\right) = \sum_{j=1}^{i-1}\mathrm{E}(Y_j) = \theta\sum_{j=1}^{i-1}\frac{1}{j}\,. \tag{2.23}$$

$$\mathrm{Var}(V_i) \approx \mathrm{Var}\left(\sum_{j=1}^{i-1} Y_j\right) = \sum_{j=1}^{i-1}\mathrm{Var}(Y_j)\,, \text{pela independência} \approx \sum_{j=1}^{i-1}\left(\frac{\theta}{j} + \frac{\theta^2}{j^2}\right)$$

$$\Rightarrow \mathrm{Var}(V_i) \approx \theta\sum_{j=1}^{i-1}\frac{1}{j} + \theta^2\sum_{j=1}^{i-1}\frac{1}{j^2} = a_n\theta + b_n\theta^2\,. \tag{2.24}$$

em que $a_n = \sum_{j=1}^{i-1}\frac{1}{j}$ e $b_n = \sum_{j=1}^{i-1}\frac{1}{j^2}$.

2.2.3 Estimador de Polimorfismo Molecular baseado no Número de Sítios Segregantes

O parâmetro de interesse é $\theta = 4Nu \approx 4N\nu$ que é o número esperado de mutações na população, por gene por geração. Desta forma, procuraremos estimadores de θ com boas propriedades, como por exemplo, não viciados, com variân-

cias e boas propriedades assintóticas.

Para n gametas, sabe-se que:

$$\mathrm{E}(V_n) \approx \theta a_n \,.$$

Então, pelo método dos momentos, tem-se um estimador não viciado para θ:

$$\widehat{\theta} = \mathcal{T}_1 = \frac{V_n}{a_n} \,. \tag{2.25}$$

Seguindo a notação de Tajima (1989), denota-se V_n por S. Assim, $\mathcal{T}_1 = \dfrac{S}{a_n}$. A variância deste estimador de θ é:

$$\begin{aligned} \mathrm{Var}(\mathcal{T}_1) &= \mathrm{Var}\left(\frac{S}{a_n}\right) = \frac{1}{a_n}\mathrm{Var}(S) \\ &\approx \frac{\theta}{a_n} + \frac{\theta^2 b_n}{a_n^2} \,. \end{aligned} \tag{2.26}$$

Um estudo da distribuição assintótica do estimador $\mathcal{T}_1$ de θ foi feito utilizando o teorema a seguir.

Teorema 1. ***(Teorema Central do Limite de Lindeberg):*** *Sejam X_1, X_2, ... variáveis aleatórias independentes tais que $E(X_n) = \mu_n$ e $Var(X_n) = \sigma_n^2$, onde $\sigma_n^2 < \infty$ e pelo menos um $\sigma_n^2 > 0$. Sejam $F_n = F_{X_n}$,*

$$S_n = X_1 + \ldots + X_n$$

e

$$s_n = \sqrt{Var(S_n)} = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}.$$

Então para que

$$\frac{S_n - E(S_n)}{s_n} \xrightarrow{D} N(0,1) \quad \textit{quando } n \to \infty,$$

é suficiente que a seguinte condição de Lindeberg seja satisfeita:

$$\forall \varepsilon > 0, \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{|x-\mu_k|>\varepsilon s_n} (x-\mu_k)^2 dF_k(x) = 0.$$

Em outras palavras, se a condição de Lindeberg é satisfeita, vale a convergência normal.

A condição de Lindeberg significa, basicamente, que as parcelas $\frac{X_k-\mu_k}{s_n}$ da soma $\frac{S_n-E(S_n)}{s_n}$ são uniformemente pequenas para n grande. Por exemplo, a condição de Lindeberg implica que

$$\max_{1\leq k\leq n}\frac{\sigma_k^2}{s_n^2}\to 0 \quad \textit{quando} \quad n\to\infty,$$

ou seja, para n grande, as variâncias das parcelas são uniformemente pequenas em relação à variância da soma. Nenhuma parcela tem muito peso na soma $\frac{S_n-E(S_n)}{s_n}$. Do ponto de vista intuitivo isto serve para justificar a afirmação: a soma de um grande número de pequenas quantidades independentes e de média zero tem aproximadamente a distribuição normal.

No caso do estudo de polimorfismo, têm-se: Y_1, Y_2, ... variáveis aleatórias independentes e $S = V_n \approx Y_1 + \ldots + Y_{n-1}$. Tem-se que:

$$s_n = \sqrt{\mathrm{Var}(V_n)} \approx \sqrt{\theta\sum_{j=1}^{n-1}(1/j) + \theta^2\sum_{j=1}^{n-1}(1/j^2)}\,.$$

Verificou-se, para n grande, que as variâncias das parcelas são uniformemente pequenas em relação à variância da soma. Foi visto que:

$$\max_{1\leq k\leq n-1}\frac{\sigma_k^2}{s_n^2} \approx \frac{\theta+\theta^2}{\theta\sum_{j=1}^{n-1}(1/j)+\theta^2\sum_{j=1}^{n-1}(1/j^2)} \to 0 \quad \textit{quando} \quad n\to\infty\,, \qquad (2.27)$$

pois $\sum_{j=1}^{\infty}(1/j)\to\infty$ e $\sum_{j=1}^{\infty}(1/j^2)\to\pi^2/6\,,\mathrm{Var}(V_n)\to\infty$ quando $n\to\infty\,.$

Para verificar se a condição de Lindeberg é satisfeita utiliza-se o seguinte fato:

Seja $X\sim\mathrm{Exp}(\lambda)$, a sua função densidade é $f_X(x)=\lambda e^{-\lambda x}$, $x>0$, $\lambda>0$.

Seja $Y=\lfloor X\rfloor$, o maior inteiro menor ou igual a X.

$$\begin{aligned} P(Y=n) &= \int_n^{n+1} f_X(t)dt = \int_n^{n+1} \lambda e^{-\lambda t}dt \\ &= -e^{-(n+1)\lambda} + e^{-n\lambda} \\ &= (1-e^{-\lambda})e^{-\lambda n}, \quad n = 0,1,2,... \end{aligned}$$

Portanto, Y tem distribuição Geométrica$(1-e^{-\lambda})$ e tem-se a seguinte relação: $Y_k = \lfloor X_k \rfloor \leq X_k$, então: $E(Y_k) \leq E(X_k)$, em que $Y_k \sim$ Geométrica $\left(\frac{1}{1+\frac{\theta}{k}}\right)$ e $X_k \sim \text{Exp}(\lambda_k)$, com $\lambda_k = -\ln\left(\frac{\theta}{\theta+k}\right)$.

$$\begin{aligned} \int_{|x-\mu_k|>\varepsilon s_n} (x-\mu_k)^2 dF_{Y_k}(x) &\leq \int_{|x-\mu'_k|>\varepsilon s_n} (x-\mu'_k)^2 dF_{X_k}(x) \\ &\leq \underbrace{\int_{-\infty}^{\mu'_k-\varepsilon s_n} (x-\mu'_k)^2 dF_{X_k}(x)}_{(a)} + \int_{\mu'_k+\varepsilon s_n}^{+\infty} (x-\mu'_k)^2 dF_{X_k}(x)\,, \end{aligned}$$

em que $\mu_k = E(Y_k) = \frac{\theta}{k}$ e $\mu'_k = E(X_k) = \frac{1}{\lambda_k}$.

X_k é uma variável aleatória que só assume valores positivos, portanto, a partir de um certo n, $\mu_k - \varepsilon s_n$ será negativo e, portanto, (a) não precisa ser considerada para o estudo do limite.

$$\begin{aligned} \int_{\mu'_k+\varepsilon s_n}^{+\infty} (x-\mu'_k)^2 dF_{X_k}(x) &= \int_{\mu'_k+\varepsilon s_n}^{+\infty} (x-\mu'_k)^2 \lambda_k e^{-\lambda_k x}dx = \int_{\varepsilon s_n}^{+\infty} y^2 \lambda_k e^{-\lambda_k (y+\mu'_k)}dy \\ &= e^{-\lambda_k \mu'_k} \int_{\varepsilon s_n}^{+\infty} y^2 \lambda_k e^{-\lambda_k y}dy = \frac{1}{e}\int_{\varepsilon s_n}^{+\infty} y^2 \lambda_k e^{-\lambda_k y}dy \\ &\leq \frac{1}{e}\frac{2}{\lambda_k^2} = \frac{2}{e\left[\ln\left(\frac{\theta}{\theta+k}\right)\right]^2}\,. \end{aligned}$$

Numericamente, no entanto, verificou-se que esta aproximação não satisfaz a condição de Lindeberg. A tentativa posterior foi tentar obter o limite através do cálculo exato da integral, mas o limite a ser calculado é bastante complicado.

Dessa maneira, apesar de não se obter uma definição sobre a condição de Lindeberg ser satisfeita, tem-se de (2.27) que a distribuição assintótica é infinitamente divisível.

2.3 Diversidade de Nucleotídeos

A teoria estatística sobre o número de diferenças de nucleotídeos entre genes amostrados foi desenvolvida no trabalho de Tajima (1983). O propósito do trabalho de Tajima (1983) é apresentar uma teoria estatística sobre o número de diferenças de nucleotídeos entre genes amostrados, quando a mudança evolutiva dos genes é determinada somente pela mutação e pela deriva genética aleatória.

Em alguns grupos de genes, como o DNA mitocondrial, a recombinação é muito rara e neste caso é possível construir uma árvore genealógica dos alelos.

Considera-se aqui uma população com cruzamento aleatório de N indivíduos diplóides e assume-se que não há migração de genes de populações de fora. Também assume-se que não há seleção nem recombinação entre as seqüências de DNA.

2.3.1 Relação Evolutiva Esperada de uma Amostra de Genes

Para estudar os momentos do número médio de diferenças em pares de nucleotídeos, é necessário conhecer a topologia da árvore.

Considere o caso em que os genes são amostrados aleatoriamente de uma população. Quando dois genes são amostrados, tem-se um gene ancestral comum (Figura 2.1a). Quando três genes são amostrados, tem-se dois casos possíveis. Um é quando um ancestral comum bifurca-se e uma das ramificações bifurca-se novamente (Figura 2.1b). Outro caso pode ser obtido quando um gene ancestral comum trifurca-se, mas a probabilidade deste último evento, como será visto posteriormente, é muito pequena. Conseqüentemente, assume-se que todas as ramificações são criadas por bifurcação. Quando quatro genes são amostrados, existem dois casos possíveis que

estão representados na Figura 2.2. As probabilidades de que ocorram estes casos podem ser obtidas usando a Figura 2.1b. Se a bifurcação ocorrer nos pontos C ou D, obtém-se a Figura 2.2a, se a bifurcação ocorrer em E, obtém-se a Figura 2.2b. Como estes três eventos de bifurcação têm a mesma probabilidade (1/3), as probabilidades de obtermos os casos das Figuras 2.2a e 2.2b são 2/3 e 1/3, respectivamente.

Em geral, a probabilidade de obtermos uma certa relação para n genes é:

$$P = \frac{2^{n-1-s}}{(n-1)!}\,, \tag{2.28}$$

em que s é o número de ramificações que levam a exatamente dois genes descendentes na amostra.

Demonstração. Considere casos com n genes. Assume-se que todas as ramificações são originadas por bifurcações. Começa-se de um gene ancestral comum. A bifurcação deste gene cria duas ramificações. Depois, uma destas duas ramificações bifurca-se também. Neste caso, existem dois possíveis casos de bifurcação, embora eles criem a mesma relação. Seguindo este processo, um dos $n-1$ genes finalmente bifurca. Neste caso, existem $(n-1)!$ maneiras possíveis de bifurcação para criar n genes de um ancestral comum. Nota-se, no entanto, que algumas bifurcações criam a mesma relação. Note que há $n-1$ pontos de bifurcação em uma árvore com n genes. Quando um certo ponto de ramificação é assimétrico, existem dois tipos de bifurcação que criam a mesma relação. O número de pontos de ramificação assimétrica é $n-s-1$, em que s é o número de pontos de ramificação que levam a exatamente dois genes descendentes na amostra. Assim, obtém-se (2.28). □

Muitas vezes, tem-se interesse somente nas relações topológicas. Neste caso, a probabilidade de obter-se uma certa topologia pode ser calculada através da soma das probabilidades de obter-se cada uma das relações que nos levam à relação topológica desejada. Em geral, a probabilidade de ocorrer uma relação topológica particular para genes amostrados pode ser obtida usando a seguinte probabilidade. A probabilidade de que um certo ponto de ramificação divida n genes em n_1 e n_2

genes (não importa a ordem) é:

$$P(n_1, n_2) = \begin{cases} \frac{2}{(n-1)} & \text{se} \quad n_1 \neq n_2 \\ \frac{1}{(n-1)} & \text{se} \quad n_1 = n_2 \end{cases} , \tag{2.29}$$

em que $n = n_1 + n_2$.

Demonstração. Considere o caso no qual um ponto de ramificação particular divide n genes. Denotaremos por $Q(n_1, n_2 \mid n)$ a probabilidade de que o lado esquerdo desta ramificação tenha n_1 genes e o lado direito tenha n_2 genes, onde $n = n_1+n_2$. Quando um destes n genes bifurca, a bifurcação ocorre no lado esquerdo com probabilidade n_1/n e no lado direito com probabilidade n_2/n. Assim, temos:

$$Q(n_1, n_2 \mid n) = Q(n_1 - 1, n_2 \mid n-1)\frac{(n_1-1)}{(n-1)} + Q(n_1, n_2 - 1 \mid n-1)\frac{(n_2-1)}{(n-1)} .$$

Usando $Q(1, 1 \mid 2) = 1$ como condição inicial, temos:

$$Q(n_1, n_2 \mid n) = \frac{1}{(n-1)} .$$

Se denotarmos por $P(n_1, n_2)$ a probabilidade de que um certo ponto de ramificação divida n genes em n_1 e n_2,obtém-se (2.29). □

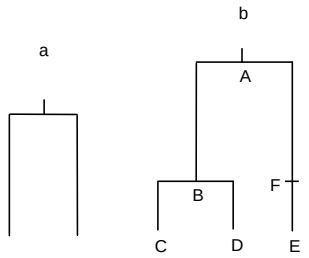


Figura 2.1: Relações evolutivas esperadas, (a) quando dois genes são amostrados e (b) quando três genes são amostrados de uma população.

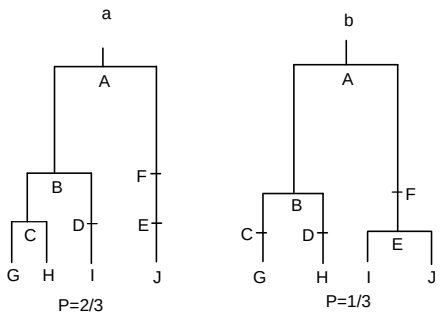


Figura 2.2: Relações evolutivas esperadas entre quatro genes amostrados.

2.3.1.1 Comprimento de Ramo

Considere agora o comprimento dos ramos de uma genealogia. É conveniente medir o comprimento do ramo em termos de números de gerações.

Seja $P(T_n = t)$ a probabilidade de que n genes amostrados aleatoriamente de uma população provenham de $n-1$ genes e que o número de gerações anteriores que apresentam bifurcação até o aparecimento da geração com coalescência é t.

Primeiramente, considere o caso de uma amostra com dois genes. A relação topológica é representada pela Figura 2.1a. A probabilidade de que dois genes provenham de um ancestral comum na geração imediatamente anterior, ou seja, a probabilidade de coalescerem, é:

$$P(T_2 = 0) = \frac{1}{2N}\,, \tag{2.30}$$

em que N é o número de indivíduos diplóides na população.

Demonstração. O primeiro gene amostrado tem $2N$ possibilidades de escolha de um ancestral. O segundo gene amostrado tem apenas uma possibilidade, que é o mesmo ancestral do gene anteriormente amostrado. Assim, a probabilidade destes dois eventos ocorrerem conjuntamente é: $\frac{2N}{2N}\frac{1}{2N} = \frac{1}{2N}$. □

Tem-se:

$$\begin{aligned} P(T_2 = t) &= P(T_2 = 0)[1 - P(T_2 = 0)]^t = \frac{1}{2N}\left[1 - \frac{1}{2N}\right]^t \\ &\approx \left[\frac{1}{2N}\right] e^{-\frac{t}{2N}}, \quad t = 0,\, 1,\, \ldots \end{aligned} \tag{2.31}$$

que é a distribuição de probabilidade do comprimento do ramo em termos do número de gerações t (ver (2.17) e (2.18) com $i = 2$).

Assim, quando considera-se um amostra de dois genes, o comprimento dos ramos em termos do número de gerações segue uma distribuição Geométrica com parâmetro $1/2N$, que pode ser aproximada pela distribuição Exponencial.

Demonstração. $P(T_2 = t)$ é a probabilidade de que dois genes amostrados provenham de 1 gene ancestral e que t gerações apresentaram bifurcação. Sabe-se que a probabilidade de que dois genes provenham de um ancestral comum na geração imediatamente anterior é $P(T_2 = 0)$, ou seja, é a probabilidade do evento de coalescência na geração imediatamente anterior. Tem-se que $P(T_2 = t)$ depende de dois eventos que ocorrem simultaneamente: a coalescência na geração imediatamente anterior e a bifurcação nas outras t gerações anteriores.

A forma aproximada da expressão é obtida da seguinte maneira. Tem-se que:

$$\ln(1 + x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}, \quad |\, x \,| < 1\,.$$

Assim:

$$\begin{aligned} \ln\left(1 - \frac{1}{2N}\right) &= \left(-\frac{1}{2N}\right) - \frac{1}{2}\left(-\frac{1}{2N}\right)^2 + \frac{1}{3}\left(-\frac{1}{2N}\right)^3 - \ldots \\ &= -\frac{1}{2N} + O(N^{-2})\,. \end{aligned}$$

Então:

$$\begin{aligned}\left(1-\frac{1}{2N}\right)^t &= \exp\left[t\ln\left(1-\frac{1}{2N}\right)\right] \\ &\approx \exp\left[t\left(-\frac{1}{2N}\right)\right] = \exp\left(-\frac{t}{2N}\right)\,.\end{aligned}$$

A aproximação da distribuição Geométrica para a distribuição Exponencial facilitará os cálculos. □

Sabendo a distribuição de probabilidades do comprimento do ramo em termos do número de gerações, T_2, calculam-se sua esperança e variância.

$$\begin{aligned}\mathrm{E}(T_2) &= \frac{1}{P(T_2=0)} - 1 = 2N-1\,. \\ \mathrm{Var}(T_2) &= \frac{1-P(T_2=0)}{[P(T_2=0)]^2} = 4N^2-2N\,.\end{aligned}$$

Quando três genes são amostrados, tem-se a relação da Figura 2.1b. O comprimento do ramo entre os pontos A e B é T_2. O comprimento do ramo entre os pontos B e C é T_3 por definição. A probabilidade de que três genes provenham de um ancestral comum na geração imediatamente anterior é:

$$\frac{1}{(2N)^2} \approx 0\,.$$

Demonstração.

$$\frac{2N}{2N}\frac{1}{2N}\frac{1}{2N} = \frac{1}{(2N)^2} \approx 0\,.$$

O primeiro gene amostrado tem $2N$ possibilidades de ancestrais distintos. O segundo gene amostrado, deve ter o mesmo ancestral que o gene anterior, portanto, tem apenas uma possibilidade, o mesmo ocorre com o terceiro gene da amostra. □

Como a probabilidade de que três genes provenham de um único ancestral comum é baixa, a trifurcação de genes é rara. Por esta razão, não se considera a

trifurcação.

A probabilidade de que três genes sejam originados a partir de dois genes na geração imediatamente anterior é:

$$P(T_3 = 0) = \binom{3}{2}\frac{1}{2N}\left(1 - \frac{1}{2N}\right) = \frac{3}{2N} - \underbrace{\frac{3}{4N^2}}_{\searrow 0} = \frac{3}{2N} + O(N^{-2})\,. \tag{2.32}$$

E a probabilidade de que três genes provenham de três genes diferentes na geração imediatamente anterior é:

$$\left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right) = 1 - \frac{3}{2N} + \underbrace{\frac{2}{4N^2}}_{\searrow 0} = 1 - \frac{3}{2N} + O(N^{-2})\,. \tag{2.33}$$

Então,

$$\begin{aligned} P(T_3 = t) &= P(T_3 = 0)[1 - P(T_3 = 0)]^t = \frac{3}{2N}\left[1 - \frac{3}{2N}\right]^t \\ &\approx \frac{3}{2N}\, e^{-\frac{3t}{2N}}, \quad t = 0,\, 1,\, \ldots \end{aligned} \tag{2.34}$$

A expressão acima dá a distribuição de probabilidades do comprimento do ramo entre os pontos B e C da Figura 2.1b. Essa é a probabilidade de que três genes provenham de dois ancestrais e que t gerações apresentaram bifurcação.

A esperança e a variância do comprimento dos ramos em termos de número de gerações, quando amostramos três genes é:

$$\mathrm{E}(T_3) = \frac{1}{P(T_3 = 0)} - 1 = \frac{2N}{3} - 1\,. \tag{2.35}$$

$$\mathrm{Var}(T_3) = \frac{1 - P(T_3 = 0)}{[P(T_3 = 1)]^2} = \frac{4N^2}{9} - \frac{2N}{3}\,. \tag{2.36}$$

Isto indica que o comprimento esperado do ramo entre os pontos B e C é três vezes menor do que entre os pontos A e B.

De forma similar,obtém-se $P(T_n = t)$:

$$P(T_n = t) = \frac{\binom{n}{2}}{2N}\left[1 - \frac{\binom{n}{2}}{2N}\right]^t \approx \frac{\binom{n}{2}}{2N}\exp\left[-\frac{\binom{n}{2}t}{2N}\right], \quad t = 0,\, 1,\, \ldots \tag{2.37}$$

que é a probabilidade de que n genes provenham de $n-1$ ancestrais há $t+1$ gerações e que a divergência tenha ocorrido há t gerações.

Então, a probabilidade de que n genes provenham de $n-1$ ancestrais na geração imediatamente anterior é:

$$P(T_n = 0) = \frac{\binom{n}{2}}{2N}\,. \tag{2.38}$$

Desta maneira, o tempo (em gerações) durante o qual existem exatamente n genes na genealogia segue uma distribuição Geométrica. A esperança e a variância de T_n, comprimento do ramo quando tem-se uma amostra de n genes, são dadas por:

$$\mathrm{E}(T_n) = \frac{1}{P(T_n = 0)} - 1 = \frac{2N}{\binom{n}{2}} - 1\,. \tag{2.39}$$

$$\mathrm{Var}(T_n) = \frac{1 - P(T_n = 0)}{[P(T_n = 0)^2]} = \frac{4N^2}{\binom{n}{2}^2} - \frac{2N}{\binom{n}{2}}\,. \tag{2.40}$$

$$\tag{2.41}$$

Repare que a esperança do comprimento do ramo (tempo de coalescência) decresce com o aumento de n. Quando tem-se uma amostra maior, tem-se mais pares de seqüências, o que significa que existe uma chance maior de que um dos pares de seqüências coalesça em uma geração, resultando em uma esperança menor para o tempo de coalescência.

Quando considera-se a aproximação Exponencial, tem-se:

$$\mathrm{E}(T_n) \approx \frac{2N}{\binom{n}{2}}. \tag{2.42}$$

$$\mathrm{Var}(T_n) \approx \frac{4N^2}{\binom{n}{2}^2}. \tag{2.43}$$

$$\mathrm{E}(T_n^2) \approx 2[\mathrm{E}(T_n)]^2. \tag{2.44}$$

2.3.2 Número de Diferenças em Pares de Nucleotídeos entre Genes Amostrados Aleatoriamente

Voltando ao exemplo apresentado na Tabela 2.1, tem-se que o número médio de diferenças em pares de nucleotídeos, $\bar{K}$, que será definido nesta Seção, é 7,9.

2.3.2.1 Distribuição de Probabilidade

Assume-se aqui que a taxa de mutação é a mesma para todos os nucleotídeos.

Considere dois genes escolhidos ao acaso de uma população. Se um gene consiste de m sítios e cada sítio assume um dos R estados ($R = 4$ no caso dos quatro nucleotídeos), a probabilidade de que o *número de diferenças de nucleotídeos entre dois genes* seja k, dado que estes dois genes são originados de um ancestral comum há $t+1$ gerações e que a divergência ocorreu t gerações atrás é dada por:

$$P(K = k \mid T_2 = t) = \binom{m}{k}[g(R,m,t)]^k[1-g(R,m,t)]^{m-k} \quad t,\ k = 0,\ 1, \ldots \tag{2.45}$$

em que

$$g(R,m,t) = \frac{R-1}{R}\left[1 - \exp\left(-\frac{2R\nu}{(R-1)m}t\right)\right]$$

e ν é a taxa de mutação por gene por geração. $g(R,m,t)$ é a probabilidade de que um sítio em particular seja polimórfico (TAKAHATA, 1982). Como o número de diferenças de nucleotídeos para um certo valor de t segue uma distribuição Binomial com parâmetros m e $g(R,m,t)$, chega-se à expressão (2.45).

Pelo Teorema das Probabilidades Totais, obtém-se a probabilidade de que o

número de diferenças de nucleotídeos entre dois genes amostrados aleatoriamente de uma população seja igual a k. Assim,

$$P(K=k)=\sum_{t=0}^{\infty}P(K=k\mid T_2=t)P(T_2=t)\,. \tag{2.46}$$

A esperança de K é dada por:

$$\mathrm{E}(K)\approx\frac{\theta}{1+\frac{R}{(R-1)m}\theta}\,, \tag{2.47}$$

em que $\theta=4N\nu$.

Demonstração.

$$\begin{aligned}
\mathrm{E}(K) &= \mathrm{E}[\mathrm{E}(K\mid T_2=t)]=\sum_{t=0}^{\infty}\mathrm{E}(K\mid T_2=t)P(T_2=t)\\
&= \sum_{t=0}^{\infty}mg(R,m,t)P(T_2=t)\\
&= \sum_{t=0}^{\infty}m\left(\frac{R-1}{R}\right)\left[1-\exp\left(-\frac{2R\nu}{(R-1)m}t\right)\right]P(T_2=t)\\
&= m\left(\frac{R-1}{R}\right)\sum_{t=0}^{\infty}P(T_2=t)-m\left(\frac{R-1}{R}\right)\sum_{t=0}^{\infty}\exp\left(-\frac{2R\nu}{(R-1)m}t\right)P(T_2=t)\\
&= m\left(\frac{R-1}{R}\right)-m\left(\frac{R-1}{R}\right)\mathrm{E}\left[\exp\left(-\frac{2R\nu}{(R-1)m}T_2\right)\right]\\
&\approx \frac{m(R-1)}{R}-\frac{m(R-1)}{R}\left[\frac{1}{2N}\int_0^{\infty}\exp\left[-\left(\frac{2R\nu}{(R-1)m}+\frac{1}{2N}\right)t\right]dt\right]\\
&\approx \frac{m(R-1)}{R}-\frac{m(R-1)}{R}\left[\frac{(R-1)m}{4N\nu R+(R-1)m}\right]\\
&\approx \frac{\theta}{1+\frac{R}{(R-1)m}\theta}\,.
\end{aligned}$$

□

$$\mathrm{Var}(K) \approx \frac{\theta}{\left[1+\frac{R\theta}{(R-1)m}\right]} + \frac{\left\{1-\frac{2}{m}\left[1+\frac{R\theta}{(R-1)m}\right]\right\}\theta^2}{\left\{\left[1+\frac{R\theta}{(R-1)m}\right]^2\left[1+\frac{2R\theta}{(R-1)m}\right]\right\}}, \qquad (2.48)$$

como foi mostrado por Tajima (1983).

Como R é igual a quatro (no caso de seqüências de nucleotídeos) e m é geralmente muito grande, será usado o modelo de infinitos sítios (ver Capítulo 1) a partir de agora. Neste modelo, assume-se as mesmas suposições que no trabalho de Watterson (1975), apresentadas na Seção 2.2. Assim, $P(K = k \mid T_2 = t)$, que considera dois genes, é dada pela seguinte expressão:

$$P(K = k \mid T_2 = t) = \frac{e^{-2\nu t}(2\nu t)^k}{k!}, \quad k = 0,\, 1, \ldots \qquad (2.49)$$

De (2.46) tem-se que:

$$P(K = k) \approx \left(\frac{1}{1+\theta}\right)\left(\frac{\theta}{1+\theta}\right)^k, \quad k = 0,\, 1,\, \ldots \qquad (2.50)$$

que é equivalente a (2.20) com $j = 1$.

Demonstração. Para $K = 4$. De (2.46), tem-se:

$$\begin{aligned} P(K=k) &= \sum_{t=1}^{\infty} P(K=k \mid T_2=t)P(T_2=t) \approx \int_0^{\infty} \frac{e^{-2\nu t}}{k!}(2\nu t)^k \frac{1}{2N}\, e^{-t/2N} dt \\ &\approx \frac{(2\nu)^k}{2N}\int_0^{\infty} \exp\left[-\left(\frac{4N\nu+1}{2N}\right)t\right]\frac{t^k}{k!}\, dt \\ &\approx \frac{\theta^k}{(2N)^{k+1}}\int_0^{\infty} \exp\left[-\left(\frac{\theta+1}{2N}\right)t\right]\frac{t^k}{k!}\, dt \end{aligned}$$

Para $k = 4$,

$$P(K=k) \approx \frac{\theta^4}{(2N)^5}\underbrace{\int_0^{\infty} \exp\left[-\left(\frac{\theta+1}{2N}\right)t\right]\frac{t^4}{4!}\, dt}_{(a)} \quad (1)$$

Integrando (a) por partes:

$$u = t^4 \quad du = 4t^3 dt \quad dv = \frac{\exp[-(\frac{\theta+1}{2N})t]}{4!}dt \quad v = -\frac{2N}{(\theta+1)}\frac{\exp[-(\frac{\theta+1}{2N})t]}{4!}$$

$$\begin{aligned}(a) &= \left[-\frac{2N}{(\theta+1)}\frac{\exp[-(\frac{\theta+1}{2N})t]}{4!}\frac{t^4}{4!}\right]_0^\infty + \int_1^\infty \frac{2N}{(\theta+1)}\frac{\exp[-(\frac{\theta+1}{2N})t]}{4!}\,4t^3\,dt \\ &= \frac{2N}{3!(\theta+1)}\underbrace{\int_0^\infty \exp\left[-\left(\frac{\theta+1}{2N}\right)t\right]t^3\,dt}_{(b)} \quad (2)\end{aligned}$$

Integrando (b) por partes: $(b) = \frac{6N}{(\theta+1)}\underbrace{\int_0^\infty \exp\left[-\left(\frac{\theta+1}{2N}\right)t\right]t^2\,dt}_{(c)}$ (3)

Integrando (c) por partes: $(c) = \frac{4N}{(\theta+1)}\underbrace{\int_0^\infty \exp\left[-\left(\frac{\theta+1}{2N}\right)t\right]t\,dt}_{(d)}$ (4)

Integrando (d) por partes: $(d) = \left(\frac{2N}{\theta+1}\right)^2$ (5)

Substituindo (5) em (4): $(c) = \frac{16N^3}{(\theta+1)^3}$ (6)

Substituindo (6) em (3): $(b) = \frac{96n^4}{(\theta+1)^4}$ (7)

Dessa forma, $(a) = \frac{32N^5}{(\theta+1)^5}$.

Então, $P(K=4) \approx \frac{\theta^4}{(\theta+1)^5}$. □

2.3.3 Estudo dos Momentos das Diferenças de Nucleotídeos para uma Amostra de Genes

O objetivo desta Seção é estudar o número médio de diferenças em pares de nucleotídeos através de um modelo de infinitos sítios, ou seja, quer-se obter a média e a variância do número médio de diferenças em pares de nucleotídeos.

Quando dois genes são amostrados de uma população, a esperança e a variância

de K quando $m \to \infty$ em (2.47) e (2.48) são dadas por:

$$\begin{aligned} \mathrm{E}(K) &\approx \theta \qquad (2.51) \\ \mathrm{Var}(K) &\approx \theta + \theta^2 \,, \qquad (2.52) \end{aligned}$$

que são os mesmos momentos obtidos para $\mathcal{T}_1$ por Watterson (1975) (equações 2.23 e 2.24 para V_2).

Quando três genes são amostrados, a relação evolutiva entre eles é dada pela Figura 2.1b. Neste caso, temos três estimativas: o número de diferenças de nucleotídeos entre C e D, entre C e E e entre D e E (Figura 2.1b). Ao denotar-se o número de diferenças de nucleotídeos entre os genes i e j por K_{ij} , as seguintes relações são obtidas (observe a Figura 2.1b):

$$\begin{aligned} K_{CD} &= K_{BC} + K_{BD} \\ K_{CE} &= K_{BF} + K_{BC} + K_{EF} \qquad (2.53) \\ K_{DE} &= K_{BF} + K_{BD} + K_{EF} \,. \end{aligned}$$

De (2.51) e (2.52), tem-se:

$$\begin{aligned} \mathrm{E}(K_{BF}) &\approx \theta \\ \mathrm{Var}(K_{BF}) &\approx \theta + \theta^2 \,. \end{aligned}$$

As esperanças, variâncias e covariâncias de K_{BC}, K_{BD}, K_{EF} são dadas por:

$$\mathrm{E}(K_{BC}) = \mathrm{E}(K_{BD}) = \mathrm{E}(K_{EF}) = \sum_{t=0}^{\infty} \sum_{k=0}^{\infty} k P(T_3 = t) Q(K \mid T_3 = t) \approx \theta/6 \,.$$

$$\begin{aligned} \mathrm{Var}(K_{BC}) = \mathrm{Var}(K_{BD}) = \mathrm{Var}(K_{EF}) &= \sum_{t=0}^{\infty} \sum_{k=0}^{\infty} k^2 P(T_3 = t) Q(K \mid T_3 = t) - (\theta/6)^2 \\ &\approx \theta/6 + (\theta/6)^2 \,. \end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(K_{BC}, K_{BD}) &= \mathrm{Cov}(K_{BC}, K_{EF}) = \mathrm{Cov}(K_{BD}, K_{EF}) \\ &= \sum_{t=0}^{\infty}\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty} k_1 k_2 P(T_3 = t) Q(K_1 \mid T_3 = t) Q(K_2 \mid T_3 = t) - (\theta/6)^2 \\ &\approx (\theta/6)^2 \,,\end{aligned}$$

em que $Q(K \mid T_n = t) \approx \exp(-\nu t)(\nu t)^k/k!$ é a probabilidade do número de mutações em um segmento do ramo de comprimento T_n quando $m \to \infty$. Portanto, tem-se:

$$\mathrm{E}(K_{CD}) = \mathrm{E}(K_{BC}) + \mathrm{E}(K_{BD}) \approx \theta/3 \,.$$

$$\mathrm{Var}(K_{CD}) = \mathrm{Var}(K_{BC}) + \mathrm{Var}(K_{BD}) + 2\,\mathrm{Cov}(K_{BC}, K_{BD}) \approx \theta/3 + (\theta/3)^2 \,.$$

$$\mathrm{E}(K_{CE}) = \mathrm{E}(K_{BF}) + \mathrm{E}(K_{BC}) + \mathrm{E}(K_{EF}) \approx (4/3)\theta \,.$$

$$\mathrm{Var}(K_{CE}) = \mathrm{Var}(K_{BF}) + \mathrm{Var}(K_{BC}) + \mathrm{Var}(K_{EF}) + 2\,\mathrm{Cov}(K_{BC}, K_{EF}) \approx (4/3)\theta + (10/9)\theta^2 \,.$$

$$\mathrm{E}(K_{DE}) = \mathrm{E}(K_{CE}), \quad \mathrm{Var}(K_{DE}) = \mathrm{Var}(K_{CE}) \,.$$

Para generalizar o exemplo, seja $K_{ij}^{(n)}$ o número de diferenças de nucleotídeos entre os genes i e j em uma amostra de n genes. Então:

$$K_{ij}^{(n)} = \begin{cases} A_i^{(n)} + A_j^{(n)}, & Pr = \frac{2}{n(n-1)} \\ K_{ij}^{(n-1)} + A_i^{(n)} + A_j^{(n)}, & Pr = 1 - \frac{2}{n(n-1)}, \end{cases} \tag{2.54}$$

em que $A_i^{(n)}$ (ou $A_j^{(n)}$) é o número de mutações em um segmento do ramo de comprimento T_n.

$$\begin{aligned}
\mathrm{E}(A_i^{(n)}) &= \sum_{t=0}^{\infty}\sum_{k=0}^{\infty} kP(T_n = t)Q(K \mid T_n = t) \\
&\approx \frac{\theta}{n(n-1)}\,. \qquad (2.55)\\
\mathrm{Var}(A_i^{(n)}) &= \sum_{t=0}^{\infty}\sum_{k=0}^{\infty} k^2 P(T_n = t)Q(K \mid T_n = t) - [\mathrm{E}(A_i^{(n)})]^2 \\
&\approx \frac{\theta}{n(n-1)} + \left[\frac{\theta}{n(n-1)}\right]^2\,. \qquad (2.56)\\
\mathrm{Cov}(A_i^{(n)}, A_j^{(n)}) &= \sum_{t=0}^{\infty}\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty} k_1 k_2 P(T_n = t)Q(K_1 \mid T_n = t)Q(K_2 \mid T_n = t) \\
&- [\mathrm{E}(A_i^{(n)})]^2 \\
&\approx \left[\frac{\theta}{n(n-1)}\right]^2\,. \qquad (2.57)
\end{aligned}$$

Usando as fórmulas gerais, pode-se obter as esperanças e as variâncias do número médio de diferenças de nucleotídeos ($\bar{K}$):

$$\mathrm{E}(\bar{K}) = \mathrm{E}[(K_{CD} + K_{CE} + K_{DE})/3] \approx \theta\,. \qquad (2.58)$$

$$\mathrm{Var}(\bar{K}) = \mathrm{Var}[(K_{CD} + K_{CE} + K_{DE})/3] \approx (2/3)\theta + (5/9)\theta^2\,. \qquad (2.59)$$

A esperança e a variância do número médio de diferenças de nucleotídeos para um dado tamanho de amostra podem ser obtidas da mesma forma. Por exemplo, quando quatro genes são amostrados, existem dois tipos de relações topológicas, como foi mostrado na Figura 2.2. Usando o mesmo método, obtém-se:

$$\mathrm{E}_a(\bar{K}) \approx (17/18)\theta, \quad \mathrm{Var}_a(\bar{K}) \approx (53/108)\theta + (115/324)\theta^2$$

para o tipo a e

$$\mathrm{E}_b(\bar{K}) \approx (10/9)\theta, \quad \mathrm{Var}_b(\bar{K}) \approx (10/9)\theta + (89/324)\theta^2$$

para o tipo b. A esperança e a variância dependem do tipo da relação evolutiva.

Assim, esperança e a variância quando considera-se uma amostra de genes deve levar em conta a probabilidade de obter-se um certo tipo de relação. Neste caso, as probabilidades de obter-se as relações de tipo a e b são $2/3$ e $1/2$, respectivamente. Assim, tem-se:

$$\begin{aligned} \mathrm{E}(\bar{K}) &= (2/3)\mathrm{E}_a(\bar{K}) + (1/3)\mathrm{E}_b(\bar{K}) \approx \theta\,. \qquad (2.60) \\ \mathrm{Var}(\bar{K}) &= (2/3)\mathrm{Var}_a(\bar{K}) + (1/3)\mathrm{Var}_b(\bar{K}) + (2/3)[\mathrm{E}_a(\bar{K}) - \mathrm{E}(\bar{K})]^2 \\ &+ (1/3)[\mathrm{E}_b(\bar{K}) - \mathrm{E}(\bar{K})]^2 \\ &\approx (5/9)\theta + (23/54)\theta^2\,. \qquad (2.61) \end{aligned}$$

Em geral, a esperança e a variância do número médio de diferenças de nucleotídeos entre dois genes quando n genes são amostrados de uma população são dadas por:

$$\begin{aligned} \mathrm{E}(\bar{K}) &\approx \theta\,. \qquad (2.62) \\ \mathrm{Var}(\bar{K}) \approx \frac{n+1}{3(n-1)}\,\theta &+ \frac{2(n^2+n+3)}{9n(n-1)}\,\theta^2\,. \qquad (2.63) \end{aligned}$$

Demonstração. Define-se:

$$\bar{K} = \left(\sum_{i<j} K_{ij}\right) \Big/ \binom{n}{2} = \left(\sum_{i<j}^{n}\sum_{k=1}^{m} I\!I(X_{ik} \neq X_{jk})\right) \Big/ \binom{n}{2}\,,$$

em que

$$I\!I(X_{ik} \neq X_{jk}) = \begin{cases} 1 & \text{se} \quad X_{ik} \neq X_{jk} \\ 0 & \text{caso contrário} \end{cases}$$

e X_{ik} é uma variável aleatória que representa o nucleotídeo presente na i-ésima seqüência e no k-ésimo sítio do DNA. Como $\mathrm{E}(K_{ij}) \approx \theta$, por (2.51), tem-se:

$$\mathrm{E}(\bar{K}) = \left(\sum_{i<j} \mathrm{E}(K_{ij})\right) \Big/ \binom{n}{2} \approx \theta \quad (1)$$

$$\mathrm{Var}(\bar{K}) = \mathrm{E}(\bar{K}^2) - [\mathrm{E}(\bar{K})]^2 \quad (2)$$

$\bar{K}^2$ pode ser escrito como:

$$\bar{K}^2 = \frac{\left(\sum_{i<j} K_{ij}\right)^2}{\binom{n}{2}^2} = \frac{\left(\sum_{i<j} K_{ij}^2 + \sum_{i\neq j\neq r} K_{ij}K_{ir} + \sum_{\substack{i<j \\ i\neq j\neq r\neq s}} \sum_{r<s} K_{ij}K_{rs}\right)}{\binom{n}{2}^2} \quad (3)$$

Define-se:

$$U_2 = \mathrm{E}(K_{ij}^2) - \theta^2 \quad U_3 = \mathrm{E}(K_{ij}K_{ir}) - \theta^2 \quad U_4 = \mathrm{E}(K_{ij}K_{rs}) - \theta^2 \,.$$

Como $\mathrm{E}(\bar{K}) \approx \theta$, tem-se,

$$\begin{aligned} \mathrm{Var}(\bar{K}) &= \left(\sum_{i<j} U_2 + \sum_{i\neq j\neq r} U_3 + \sum_{\substack{i<j \\ i\neq j\neq r\neq s}} \sum_{r<s} U_4\right) \Big/ \binom{n}{2}^2 \\ &= \left\{U_2 + 2(n-2)U_3 + \binom{n-2}{2} U_4\right\} \Big/ \binom{n}{2} \quad (4) \end{aligned}$$

Quando $n = 2$, $\mathrm{Var}(\bar{K}) \approx \theta + \theta^2$, por (2.52). Portanto, de (4) tem-se:

$$\theta + \theta^2 \approx U_2 \,. \quad (5)$$

Quando $n = 3$, $\mathrm{Var}(\bar{K}) \approx (2/3)\theta + (5/9)\theta^2$, por (2.59). Assim, tem-se:

$$(2/3)\theta + (5/9)\theta^2 \approx (U_2 + 2U_3)/3 \,. \quad (6)$$

Quando $n = 4$, de (2.61):

$$(5/9)\theta + (23/54)\theta^2 \approx (U_2 + 4U_3 + U_4)/6 \,. \quad (7)$$

De (5), (6) e (7):

$$U_2 \approx \theta + \theta^2 \quad U_3 \approx (1/2)\theta + (1/3)\theta^2 \quad U_4 \approx (1/3)\theta + (2/9)\theta^2 \,. \quad (8)$$

Substituindo as fórmulas de (8) em (4),obtém-se:

$$\mathrm{Var}(\bar{K}) \approx \frac{n+1}{3(n-1)}\,\theta + \frac{2(n^2+n+3)}{9n(n-1)}\,\theta^2\,.$$

□

2.4 Número de *Singletones*

Parte da teoria sobre o número de *singletones*, denotado aqui por $\mathcal{S}^*$ está apresentada no trabalho de Fu e Li (1993b). Nesta Seção apresentam-se os resultados de Fu e Li (1993b) e a distribuição de probabilidade desta estatística.

2.4.1 Propriedades Estatísticas dos Ramos Internos e Externos

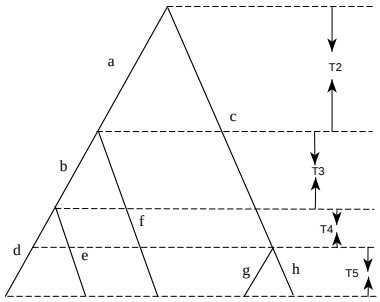


Figura 2.3: Um exemplo de genealogia de 5 genes.

Considere uma amostra aleatória de n seqüências de uma região do DNA de uma população diplóide de tamanho N com cruzamento aleatório. Suponha que todas as mutações na região sejam seletivamente neutras. Além disso, suponha que a região do DNA é de tal forma que não ocorra recombinação entre as seqüências. Então as n

seqüências na amostra são associadas a uma única árvore genealógica (Figura 2.3). Em outras palavras, as n seqüências podem ser remontadas (ou coalescem) para $n-1$ seqüências ancestrais, depois para $n-2$, etc, até atingirem o único ancestral comum.

Definiu-se T_n como o tempo necessário, em número de gerações, para a coalescência de n seqüências em $n-1$ e $T_1 = 0$ (ver equações 2.37, 2.42, 2.43 e 2.44).

A genealogia de n genes tem $2(n-1)$ ramos. Um ramo é dito externo se estiver diretamente associado a um nó externo, caso contrário é dito interno. Assim, n de $2(n-1)$ ramos são externos e os outros $n-2$ são internos. Na Figura 2.3, os ramos d, e, f, g, h são externos e os ramos a, b, c são internos. Enumeramos os n ramos externos arbitrariamente de 1 até n. A numeração é apenas para conveniência operacional e não indica a posição do ramos na árvore.

2.4.1.1 Comprimento dos Ramos Internos e Externos

Definem-se $\mathcal{J}_n$, $\mathcal{I}_n$ e $\mathcal{L}_n$, respectivamente, o comprimento total de todos os ramos, comprimento total dos ramos internos e comprimento total dos ramos externos. Note que:

$$\mathcal{J}_n = \mathcal{I}_n + \mathcal{L}_n.$$

Seja $L_i^{(n)}$ o comprimento do i-ésimo ramo externo. Então: $\mathcal{L}_n = L_1^{(n)} + L_2^{(n)} + \ldots + L_n^{(n)}$. Seja L_n o comprimento de um ramo externo escolhido aleatoriamente de uma genealogia de n genes. Tem-se:

$$\mathrm{E}(\mathcal{L}_n) = \mathrm{E}(L_1^{(n)}) + \mathrm{E}(L_2^{(n)}) + \ldots + \mathrm{E}(L_n^{(n)}) = n\mathrm{E}(L_n),$$

usando o fato de que todo ramo externo tem a mesma distribuição, pois a numeração dos ramos é apenas operacionalmente conveniente, não indicando nenhuma característica especial.

A genealogia de uma amostra aleatória de n genes é gerada adicionando dois ramos externos de comprimento T_n ao fim de um ramo externo escolhido ao acaso da

genealogia de $n-1$ genes, enquanto que os $n-2$ ramos externos restantes crescem um comprimento T_n. Portanto, tem-se a seguinte relação recorrente:

$$L_n = \begin{cases} L_{n-1} + T_n, & Pr = \frac{n-2}{n} \\ T_n, & Pr = \frac{2}{n} \quad . \end{cases} \tag{2.64}$$

Por exemplo, na Figura 2.3, o comprimento de cada um dos ramos g e h é T_5 e dos ramos d, e e f é $L_4 + T_5$.

Assim,

$$\begin{aligned} \mathrm{E}(L_n) &= \frac{\mathrm{E}(G_n)}{n(n-1)} = \frac{(n-1)4N}{n(n-1)} = \frac{4N}{n}\,. & (2.65) \\ \mathrm{E}(\mathcal{L}_n) &= n\mathrm{E}(L_n) = 4N\,. & (2.66) \end{aligned}$$

Demonstração.

$$\begin{aligned} \mathrm{E}(L_n) &= \frac{n-2}{n}\mathrm{E}(L_{n-1} + T_n) + \frac{2}{n}\mathrm{E}(T_n) \\ &= \frac{n-2}{n}\mathrm{E}(L_{n-1}) + \mathrm{E}(T_n)\,. \end{aligned}$$

Seja $G_n = n(n-1)L_n$:

$$\begin{aligned} \mathrm{E}(G_n) &= n(n-1)\mathrm{E}(L_n) \\ &= n(n-1)\left[\frac{n-2}{n}\mathrm{E}(L_{n-1}) + \mathrm{E}(T_n)\right] \\ &= (n-1)(n-2)\mathrm{E}(L_{n-1}) + n(n-1)\mathrm{E}(T_n) \\ &= \mathrm{E}(G_{n-1}) + n(n-1)\frac{2N}{\binom{n}{2}} = \mathrm{E}(G_{n-1}) + 4N \\ &= 4N + 4N + \mathrm{E}(G_{n-2}) \\ &\vdots \\ &= \underbrace{4N + \ldots + 4N}_{n-1\,vezes} + \underbrace{\mathrm{E}(G_{n-(n-1)})}_{\mathrm{E}(G_1)=0} \\ &= (n-1)4N\,. \end{aligned}$$

□

Observe que (2.66) independe do tamanho amostral n. O número de seqüências amostradas não altera o comprimento total esperado dos ramos externos, que é sempre $4N$ gerações.

A relação recorrente para $\mathcal{J}_n$ (comprimento total dos ramos) é obtida adicionando um gene à genealogia de $n-1$ genes e desta forma o comprimento total da genealogia cresce nT_n. Tem-se que:

$$\mathrm{E}(\mathcal{J}_n) = 4Na_n\,. \tag{2.67}$$

Demonstração.

$$\begin{aligned}
\mathrm{E}(\mathcal{J}_n) &= \mathrm{E}(\mathcal{J}_{n-1}) + \mathrm{E}(nT_n) = \mathrm{E}(\mathcal{J}_{n-1}) + n\mathrm{E}(T_n) \\
&= \mathrm{E}(\mathcal{J}_{n-1}) + \frac{4N}{n-1} = \frac{4N}{n-1} + \frac{4N}{n-2} + \mathrm{E}(\mathcal{J}_{n-2}) \\
&= \frac{4N}{n-1} + \frac{4N}{n-2} + \ldots + \frac{4N}{n-(n-1)} + \mathrm{E}(\mathcal{J}_{n-(n-1)}) \\
&= 4N\sum_{j=1}^{n-1}\frac{1}{j} = 4Na_n\,.
\end{aligned}$$

Uma outra maneira de se obter este resultado é definir $\mathcal{J}_n$ da seguinte maneira:

$$\mathcal{J}_n = \sum_{k=1}^{n} kT_k, \quad T_1 = 0\,. \tag{2.68}$$

Desta maneira:

$$\begin{aligned}
\mathrm{E}(\mathcal{J}_n) &= \mathrm{E}\left(\sum_{k=1}^{n} kT_k\right) = \mathrm{E}\left(\sum_{k=2}^{n} kT_k\right) = \sum_{k=2}^{n} k\mathrm{E}(T_k) \\
&= \sum_{k=2}^{n} k\frac{4N}{k(k-1)} = 4N\sum_{k=1}^{n-1}\frac{1}{k} = 4Na_n\,.
\end{aligned} \tag{2.69}$$

□

Para o comprimento total dos ramos internos, tem-se:

$$\mathrm{E}(\mathcal{I}_n) = \mathrm{E}(\mathcal{J}_n - \mathcal{L}_n) = \mathrm{E}(\mathcal{J}_n) - \mathrm{E}(\mathcal{L}_n) = 4N(a_n - 1). \tag{2.70}$$

As relações recursivas são úteis na demonstração de momentos de ordens maiores também.

$$\mathrm{Var}(\mathcal{J}_n) = \mathrm{E}(\mathcal{J}_n^2) - [\mathrm{E}(\mathcal{J}_n)]^2 = b_n(4N)^2\,. \tag{2.71}$$

Demonstração. Para obter-se a variância de $\mathcal{J}_n$, calcula-se:

$$\begin{aligned}
\mathrm{E}(\mathcal{J}_n^2) &= E\left[\left(\sum_k kT_k\right)^2\right] \\
&= \sum_{i\neq j} ij\mathrm{E}(T_iT_j) + \sum_k k^2\mathrm{E}(T_k^2) \qquad (2.72)\\
&= \sum_{i\neq j} ij\mathrm{E}(T_iT_j) + 2\sum_k [k\mathrm{E}(T_k)]^2 \quad \text{(por 2.44)} \\
&= \underbrace{\sum_{i\neq j} ij\mathrm{E}(T_iT_j) + \sum_k [k\mathrm{E}(T_k)]^2}_{\left[\sum_k k\mathrm{E}(T_k)\right]^2} + \sum_k [k\mathrm{E}(T_k)]^2 \\
&= \left[\sum_k k\frac{4N}{k(k-1)}\right]^2 + \sum_k k^2\frac{(4N)^2}{k^2(k-1)^2} \\
&= (4N)^2\left(\sum_k \frac{1}{k-1}\right)^2 + (4N)^2\sum_k \frac{1}{(k-1)^2} \\
&= (4N)^2a_n^2 + (4N)^2b_n \\
&= (a_n^2 + b_n)(4N)^2\,, \qquad (2.73)
\end{aligned}$$

em que $b_n = \displaystyle\sum_{k=1}^{n-1} \frac{1}{k^2}$. □

A variância da comprimento total dos ramos externos é dada por:

$$\mathrm{Var}(\mathcal{L}_n) = \mathrm{E}(\mathcal{L}_n^2) - [\mathrm{E}(\mathcal{L}_n)]^2 = c_n(4N)^2, \tag{2.74}$$

em que $c_2 = 1$ e

$$c_n = 2\frac{na_n - 2(n-1)}{(n-1)(n-2)} \quad \text{quando} \quad n > 2.$$

Demonstração.

$$\begin{aligned} \mathrm{E}(\mathcal{L}_n^2) &= E\left[\left(\sum_k L_k^{(n)}\right)^2\right] = \sum_{i\neq j} ij\mathrm{E}(L_iL_j) + \sum_k \mathrm{E}(L_k^{(n)2}) \\ &= n(n-1)\mathrm{E}(L_nL_n') + n\mathrm{E}(L_n^2), \end{aligned} \tag{2.75}$$

em que L_n e L_n' são os comprimentos de dois ramos externos distintos escolhidos aleatoriamente.

Para obter a esperança de uma variável, muitas vezes é mais conveniente definir uma função da variável aleatória e então obter a esperança através de uma transformação. Define-se: $G_n = n(n-1)L_n^2$.

$$\begin{aligned} \mathrm{E}(G_n) &= n(n-1)\mathrm{E}(L_n^2) \\ &= n(n-1)\left[\frac{n-2}{n}\mathrm{E}(L_{n-1}^2 + T_n^2 + 2L_{n-1}T_n) + \frac{2}{n}\mathrm{E}(T_n^2)\right] \\ &= n(n-1)\mathrm{E}(T_n^2) + 2(n-1)(n-2)\mathrm{E}(T_n)\mathrm{E}(L_{n-1}) \\ &+ \underbrace{(n-1)(n-2)\mathrm{E}(L_{n-1}^2)}_{\mathrm{E}(G_{n-1})} \\ &= n(n-1)\frac{2(4N)^2}{n^2(n-1)^2} + 2(n-1)(n-2)\frac{4N}{n(n-1)}\frac{4N}{n-1} + \mathrm{E}(G_{n-1}) \\ &= \frac{2(4N)^2}{n(n-1)} + \frac{2(n-2)(4N)^2}{n(n-1)} + \mathrm{E}(G_{n-1}) \\ &= \frac{2(4N)^2}{n} + \mathrm{E}(G_{n-1}) \\ &\vdots \\ &= 2(4N)^2\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-(n-2)}\right) + \underbrace{\mathrm{E}(G_{n-(n-1)})}_{\mathrm{E}(G_1)=0} \\ &= 2(4N)^2\left(\sum_{j=1}^{n}\frac{1}{j} - 1\right) = 2(4N)^2(a_{n+1} - 1)\,. \end{aligned}$$

Assim:

$$\mathrm{E}(L_n^2) = \frac{2(4N)^2(a_{n+1}-1)}{n(n-1)}\,. \tag{2.76}$$

Agora, calcula-se $\mathrm{E}(L_nL'_n)$. Como foi definido anteriormente, L_n e L'_n são os comprimentos de dois ramos externos diferentes escolhidos aleatoriamente. Então:

$$L_nL'_n = \begin{cases} T_n^2, & Pr = \frac{2}{n(n-1)} \\ T_n(L_{n-1}+T_n), & Pr = \frac{2(n-2)}{n(n-1)} \\ (L'_{n-1}+T_n)T_n, & Pr = \frac{2(n-2)}{n(n-1)} \\ (L_{n-1}+T_n)(L'_{n-1}+T_n), & Pr = \frac{(n-2)(n-3)}{n(n-1)}\,. \end{cases} \tag{2.77}$$

Para entender mais facilmente estas expressões, pode-se observar a Figura 2.3. Se são escolhidos os ramos g e h ($L_nL'_n = T_n^2$), a probabilidade deste evento ocorrer é:

$$\begin{aligned} P(\{g,h\}) &= P(\{g,h\} \mid \{g\})P(\{g\}) + P(\{g,h\} \mid \{h\})P(\{h\}) \\ &= \frac{1}{n-1}\frac{1}{n} + \frac{1}{n-1}\frac{1}{n} = \frac{2}{n(n-1)}\,, \quad n=5\,. \end{aligned}$$

Se são escolhidos os ramos g e f (ou e, ou d) ou os ramos h e f (ou e, ou d), tem-se $L_nL'_n = T_n(L_{n-1}+T_n)$ (ou $L_nL'_n = (L'_{n-1}+T_n)T_n$) com probabilidade:

$$\begin{aligned} &P\left([\{g,f\}\cup\{g,e\}\cup\{g,d\}]\cup[\{h,f\}\cup\{h,e\}\cup\{h,d\}]\right) = \\ &= P(\{g,f\}\cup\{g,e\}\cup\{g,d\}) + P(\{h,f\}\cup\{h,e\}\cup\{h,d\}) \\ &= P([\{f\}\cup\{e\}\cup\{d\}]\cap\{g\}) + P([\{f\}\cup\{e\}\cup\{d\}]\cap\{h\}) \\ &= P(\{f\}\cup\{e\}\cup\{d\} \mid \{g\})P(\{g\}) + P(\{f\}\cup\{e\}\cup\{d\} \mid \{h\})P(\{g\}) \\ &= \frac{1}{n}\frac{n-2}{n-1} + \frac{1}{n}\frac{n-2}{n-1} \\ &= \frac{2(n-2)}{n(n-1)}\,, \quad n=5\,. \end{aligned}$$

Se são escolhidos dois ramos dentre f, e e d, tem-se $(L_{n-1}+T_n)(L'_{n-1}+T_n)$ com

probabilidade:

$$P(\{d,e\}\cup\{e,f\}\cup\{d,f\}) = \frac{\binom{n-2}{2}}{\binom{n}{2}} = \frac{(n-2)(n-3)}{n(n-1)}\,.$$

Portanto,

$$\begin{aligned}
\mathrm{E}(L_nL_n') &= \mathrm{E}(T_n^2) + \frac{2(n-2)}{n(n-1)}(2+n-3)\mathrm{E}(L_{n-1})\mathrm{E}(T_n) \\
&+ \frac{(n-2)(n-3)}{n(n-1)}\mathrm{E}(L_{n-1}L_{n-1}') \\
&= \frac{2(4N)^2}{n^2(n-1)^2} + \frac{2(n-2)}{n}\left(\frac{4N}{n-1}\right)\left(\frac{4N}{n(n-1)}\right) \\
&+ \frac{(n-2)(n-3)}{n(n-1)}\mathrm{E}(L_{n-1}L_{n-1}') \\
&= \frac{2(4N)^2}{n^2(n-1)} + \frac{(n-2)(n-3)}{n(n-1)}\mathrm{E}(L_{n-1}L_{n-1}')\,.
\end{aligned}$$

Para se obter a esperança através da relação recorrente acima, define-se:

$$G_n = n(n-1)L_nL_n'$$

e calcula-se sua esperança:

$$\begin{aligned}
\mathrm{E}(G_n) &= n(n-1)\mathrm{E}(L_nL'_n) = \frac{2(4N)^2}{n} + (n-2)(n-3)\mathrm{E}(L_{n-1}L'_{n-1}) \\
&= \frac{2(4N)^2}{n} + \frac{n-3}{n-1}\mathrm{E}(G_{n-1}) \\
&= \frac{2(4N)^2}{n} + \frac{n-3}{n-1}\left[\frac{2(4N)^2}{n-1} + \frac{n-4}{n-2}\mathrm{E}(G_{n-2})\right] \\
&= \frac{2(4N)^2}{n} + \frac{2(4N)^2(n-3)}{(n-1)^2} + \frac{(n-3)(n-4)}{(n-1)(n-2)}\frac{2(4N)^2}{n-2} + \frac{n-3}{n-1}\mathrm{E}(G_{n-3}) \\
&= \frac{2(4N)^2}{n} + \frac{2(4N)^2(n-3)}{(n-1)^2} \\
&+ \frac{2(4N)^2}{(n-1)(n-2)}\left[\frac{(n-3)(n-4)}{n-2} + \ldots + \frac{3\times 2}{4} + \mathrm{E}(G_3)\right] \\
&= \frac{2(4N)^2}{(n-1)(n-2)}\sum_{k=1}^{n}\frac{(k-1)(k-2)}{k} \\
&= \frac{2(4N)^2}{(n-1)(n-2)}\left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right).
\end{aligned}$$

Desta maneira,

$$n(n-1)\mathrm{E}(L_nL'_n) = \frac{2(4N)^2}{(n-1)(n-2)}\left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right)$$

e, usando (2.66), (2.75) e (2.76), tem-se:

$$\begin{aligned}
\mathrm{Var}(\mathcal{L}_n) &= \mathrm{E}(\mathcal{L}_n^2) - [\mathrm{E}(\mathcal{L}_n)]^2 \\
&= \frac{2(4N)^2(a_{n+1}-1)}{(n-1)} + \frac{2(4N)^2}{(n-1)(n-2)}\left(\frac{n(n+1)}{2} - 3n + 2a_{n+1}\right) - (4N)^2 \\
&= c_n(4N)^2,
\end{aligned}$$

em que $c_n = 1$, quando $n = 2$ e

$$c_n = 2\frac{na_n - 2(n-1)}{(n-1)(n-2)} \quad \text{quando} \quad n > 2.$$

□

Note que

$$\mathrm{E}(\mathcal{I}_n\mathcal{L}_n) = \mathrm{E}(\mathcal{J}_n\mathcal{L}_n) - \mathrm{E}(\mathcal{L}_n^2),$$
$$\mathrm{E}(\mathcal{I}_n^2) = \mathrm{E}(\mathcal{J}_n^2) - 2\mathrm{E}(\mathcal{J}_n\mathcal{L}_n) + \mathrm{E}(\mathcal{L}_n^2).$$

Assim, para calcular a variância de $\mathcal{I}_n$ e a covariância entre $\mathcal{I}_n$ e $\mathcal{L}_n$, precisa-se da $\mathrm{E}(\mathcal{J}_n\mathcal{L}_n)$. Fu e Li (1993b) mostraram, através de relações recorrentes, que:

$$\mathrm{E}(\mathcal{J}_n\mathcal{L}_n) = \frac{n}{n-1}\, a_n(4N)^2.$$

Portanto, tem-se:

$$\begin{aligned}
\mathrm{Cov}(\mathcal{I}_n, \mathcal{L}_n) &= \mathrm{E}(\mathcal{I}_n\mathcal{L}_n) - \mathrm{E}(\mathcal{I}_n)\mathrm{E}(\mathcal{L}_n), \\
&= \left(\frac{1}{n-1}a_n - c_n\right)(4N)^2, \qquad (2.78) \\
\mathrm{Var}(\mathcal{I}_n) &= \mathrm{E}(\mathcal{I}^2) - [\mathrm{E}(\mathcal{I}_n)]^2 \\
&= \left[a_n^2 + b_n - 2\left(\frac{n}{n-1}a_n - c_n\right) + (c_n+1) - (a_n-1)^2\right](4N)^2 \\
&= \left(b_n - 2\frac{a_n}{n-1} + c_n\right)(4N)^2. \qquad (2.79)
\end{aligned}$$

2.4.1.2 Número de Mutações em Ramos Internos e Externos

Sejam R_e e R_i o número total de mutações nos ramos externos e internos, respectivamente e seja $S = R_i + R_e$ o número de mutações (número de sítios polimórficos) que ocorreram na genealogia de n genes. Como foi visto na Seção 2.2.1, a taxa de mutação do gene é ν. Assume-se, em geral, que o número de mutações que ocorre em uma seqüência em um período de l gerações segue a distribuição Poisson com esperança νl. Desta maneira, o número total de mutações nos ramos externos, dado $\mathcal{L}_n$ segue a distribuição Poisson:

$$P(R_e = k \mid \mathcal{L}_n) = \frac{\exp(-\nu\mathcal{L}_n)(\nu\mathcal{L}_n)^k}{k!}$$

e o número total de mutações, dado $\mathcal{J}_n$:

$$P(S = k \mid \mathcal{J}_n) = \frac{\exp(-\nu\mathcal{J}_n)(\nu\mathcal{J}_n)^k}{k!}.$$

A esperança e a variância de R_e são:

$$\begin{aligned} \mathrm{E}(R_e) &= \mathrm{E}[\mathrm{E}(R_e \mid \mathcal{L}_n)] = \mathrm{E}(\nu\mathcal{L}_n) = \theta \qquad (2.80) \\ \mathrm{Var}(R_e) &= \mathrm{E}[\mathrm{Var}(R_e \mid \mathcal{L}_n)] + \mathrm{Var}[\mathrm{E}(R_e \mid \mathcal{L}_n)] \\ &= \mathrm{E}(\nu\mathcal{L}_n) + \mathrm{Var}(\nu\mathcal{L}_n) = \nu\mathrm{E}(\mathcal{L}_n) + \nu^2\mathrm{Var}(\mathcal{L}_n) \\ &= \theta + c_n\theta^2, \qquad (2.81) \end{aligned}$$

usando (2.66) e (2.74).

É interessante notar que

$$\lim_{n\to\infty} \mathrm{Var}(\mathcal{L}_n) = (4N)^2 \lim_{n\to\infty} c_n = 0$$

o que indica que R_e segue assintoticamente a distribuição de Poisson com parâmetro θ.

Similarmente, a esperança e a variância do número total de mutações da genealogia de n genes são:

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}[\mathrm{E}(S \mid \mathcal{J}_n)] = \mathrm{E}(\nu\mathcal{J}_n) = a_n\theta\,, \qquad (2.82) \\ \mathrm{Var}(S) &= \mathrm{E}[\mathrm{Var}(S \mid \mathcal{J}_n)] + \mathrm{Var}[\mathrm{E}(S \mid \mathcal{J}_n)] = \mathrm{E}(\nu\mathcal{J}_n) + \mathrm{Var}(\nu\mathcal{J}_n) \\ &= a_n\theta + b_n\theta^2, \qquad (2.83) \end{aligned}$$

usando (2.69) e (2.73). As equações (2.82) e (2.83) são as mesmas encontradas por Watterson (1975) (ver equações 2.23 e 2.24).

Usando o fato de que $S = R_e + R_i$, podemos calcular a esperança e a variância

do número total de mutações nos ramos internos:

$$\begin{aligned} \mathrm{E}(R_i) &= \mathrm{E}(S) - \mathrm{E}(R_e) = (a_n - 1)\theta\,, && (2.84)\\ \mathrm{Cov}(R_i, R_e) &= \mathrm{E}[\mathrm{E}(R_i R_e \mid \mathcal{L}_n, \mathcal{I}_n)] - \mathrm{E}(R_i)\mathrm{E}(R_e)\,, && (2.85)\\ \mathrm{Var}(R_i) &= \mathrm{Var}(S) + \mathrm{Var}(R_e) - 2\mathrm{Cov}(R_i, R_e)\\ &= a_n\theta + b_n\theta^2 + \theta + c_n\theta^2 - 2\left(\frac{a_n}{n-1} - c_n\right)\theta^2\\ &= (a_n + 1)\theta + \left(b_n - 2\frac{a_n}{n-1} + 3a_n\right)\theta^2\,. && (2.86) \end{aligned}$$

Tem-se

$$\begin{aligned} \mathrm{Cov}(S, R_e) &= \mathrm{Cov}(R_i + R_e, R_e) = \mathrm{Cov}(R_i, R_e) + \mathrm{Var}(R_e)\\ &= \left(\frac{a_n}{n-1} - c_n\right)\theta^2 + \theta + c_n\theta^2 = \theta + \left(\frac{a_n}{n-1}\right)\theta^2\,. && (2.87) \end{aligned}$$

2.4.2 Propriedades Estatísticas de $\mathcal{S}^*$

Como considera-se apenas árvores com bifurcações, existem exatamente dois ramos ligados à raiz da árvore genealógica. Seja Z o número de ramos externos ligados à raiz da árvore. Então, para $n > 2$, Z pode assumir os valores 0 ou 1. Por exemplo, $Z = 1$ para a Figura 2.4a e 0 para a Figura 2.4b. De acordo com (2.29), $P(Z = 1) = 2/(n-1)$. Considere o caso $Z = 1$. Seja M_i o número de mutações no ramo interno (ramo a) e M_e o número de mutações no ramo externo (ramo b) ocorrendo durante o período de coalescência T_2. Então, $\mathrm{E}(M_i) = \mathrm{E}[\mathrm{E}(M_i \mid T_2)] = \mathrm{E}(M_e) = \mathrm{E}[\mathrm{E}(M_e \mid T_2)] = \theta/2 = \mathrm{E}(\nu T_2)$, pois $M_i \mid T_2 \sim \mathrm{Poisson}(\nu T_2)$ e $M_e \mid T_2 \sim \mathrm{Poisson}(\nu T_2)$.

Fu e Li (1993b) definem $\mathcal{S}^*$, o número de *singletones*, que representa a quantidade de sítios em que aparece um único indivíduo (seqüência) com nucleotídeo diferente dos demais, da seguinte maneira:

$$\mathcal{S}^* = R_e + W,$$

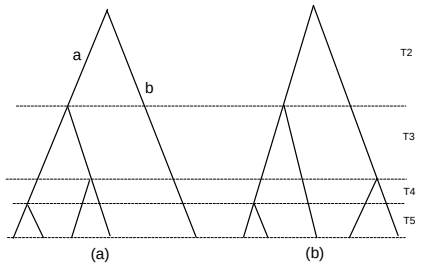


Figura 2.4: (a) Um dos dois ramos ligados à raiz é externo ($Z = 1$) e (b) Os dois ramos ligados à raiz são internos ($Z = 0$).

em que

$$W = \begin{cases} 0, & \text{se } Z = 0 \\ M_i, & \text{se } Z = 1\,. \end{cases} \tag{2.88}$$

$$\mathrm{E}(\mathcal{S}^*) = \mathrm{E}(R_e) + \mathrm{E}(W) = \theta + \frac{\theta}{n-1} = \frac{n}{n-1}\,\theta\,. \tag{2.89}$$

Portanto, pelo método dos momentos, tem-se outro estimador de θ:

$$\widehat{\theta} \;=\; \mathcal{T}_3 = \frac{(n-1)\mathcal{S}^*}{n}\,. \tag{2.90}$$

Demonstração. Como $P(Z = 1) = 2/(n-1)$, tem-se:

$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(W \mid Z = 0)P(Z = 0) + \mathrm{E}(W \mid Z = 1)P(Z = 1) \\ &= \mathrm{E}(0)P(Z = 0) + \mathrm{E}(M_i)P(Z = 1) \\ &= \mathrm{E}(M_i)\frac{2}{n-1} = \frac{\theta}{n-1} \end{aligned}$$

□

Fu e Li (1993b) mostraram que:

$$\begin{aligned} \mathrm{Var}(\mathcal{S}^*) &= \mathrm{Var}(R_e) + \mathrm{Var}(W) + 2\,\mathrm{Cov}(R_e, W) \\ &= \frac{n}{n-1}\,\theta + d_n\theta^2\,, \end{aligned}$$

em que

$$d_n = c_n + \frac{n-2}{(n-1)^2} + \frac{2}{n-1}\left(\frac{3}{2} - \frac{2a_{n+1}-3}{n-2} - \frac{1}{n}\right).$$

Portanto,

$$\mathrm{Var}(\mathcal{T}_3) = \left(\frac{n-1}{n}\right)\theta + \left(\frac{n-1}{n}\right)^2 d_n\theta^2\,. \tag{2.91}$$

2.4.2.1 Distribuição de $\mathcal{S}^*$

O número de *singletones*, $\mathcal{S}^*$, pode ser escrito da seguinte forma:

$$\mathcal{S}^* = \sum_{l=1}^{m} I\!\!I\left\{\sum_{1\le i<j\le n} I\!\!I(X_{il} \neq X_{jl}) = n-1\right\},$$

em que m é o número de sítios seqüenciados.

$$E(\mathcal{S}^*) = \sum_{l=1}^{m} \underbrace{\mathrm{P}\left(\sum_{i<j} I\!\!I(X_{il} \neq X_{jl}) = n-1\right)}_{\mathcal{P}} = m\mathcal{P}$$

e de (2.89),

$$E(\mathcal{S}^*) = \frac{n}{n-1}\,\theta\,.$$

$$\begin{aligned} \mathcal{P} &= n\{\pi_A[(\pi_C)^{n-1} + (\pi_G)^{n-1} + (\pi_T)^{n-1}] + \pi_C[(\pi_A)^{n-1} + (\pi_G)^{n-1} + (\pi_T)^{n-1}] \\ &+ \pi_G[(\pi_A)^{n-1} + (\pi_C)^{n-1} + (\pi_T)^{n-1}] + \pi_T[(\pi_A)^{n-1} + (\pi_G)^{n-1} + (\pi_C)^{n-1}]\}, \end{aligned}$$

que é a probabilidade de um sítio ser *singletone*.

$$\begin{aligned} \mathrm{E}(\mathcal{S}^*) &= m\mathcal{P} = \frac{n}{n-1}\,\theta \Rightarrow \mathcal{P} = \frac{n}{n-1}\frac{1}{m}\,\theta\,. \\ \mathcal{P} &\to \frac{\theta}{m} \quad \text{quando} \quad n \to \infty\,. \end{aligned}$$

$$\begin{aligned} P(\mathcal{S}^*=0) &= (1-\mathcal{P})^m = \left(1-\frac{n}{n-1}\frac{1}{m}\,\theta\right)^m \to \left(1-\frac{\theta}{m}\right)^m \quad \text{quando} \quad n\to\infty \\ P(\mathcal{S}^*=1) &= m\mathcal{P}(1-\mathcal{P})^{m-1} \to \theta\left(1-\frac{\theta}{m}\right)^{m-1} \quad \text{quando} \quad n\to\infty \\ &\vdots \\ P(\mathcal{S}^*=k) &= \binom{m}{k}\mathcal{P}^k(1-\mathcal{P})^{m-k} \to \binom{m}{k}\left(\frac{\theta}{m}\right)^k\left(1-\frac{\theta}{m}\right)^{m-k} \quad \text{quando} \quad n\to\infty\,. \end{aligned}$$

Sabe-se que:

$$\begin{aligned} \left(1-\frac{\theta}{m}\right)^m &= \exp\left[m\ln\left(1-\frac{\theta}{m}\right)\right] \\ &\approx \exp\left[m\left(\frac{-\theta}{m}\right)\right] = e^{-\theta} \quad \text{quando} \quad m\to\infty \end{aligned}$$

$$\begin{aligned} \left(1-\frac{\theta}{m}\right)^{m-k} &= \exp\left[(m-k)\ln\left(1-\frac{\theta}{m}\right)\right] \\ &\approx \exp\left[(m-k)\left(-\frac{\theta}{m}\right)\right] = \exp\left(-\theta+\frac{k\theta}{m}\right) \approx e^{-\theta} \quad \text{quando} \quad m\to\infty \end{aligned}$$

$$\binom{m}{k} \approx \frac{m^k}{k!} \quad \text{quando} \quad m\to\infty\,.$$

Assim, quando $m \to \infty$, tem-se:

$$\begin{aligned} P(\mathcal{S}^*=0) &= e^{-\theta} \\ P(\mathcal{S}^*=1) &= \theta e^{-\theta} \\ &\vdots \\ P(\mathcal{S}^*=k) &= \frac{\theta^k}{k!}e^{-\theta}\,. \end{aligned}$$

Resumindo:

- A distribuição de $\mathcal{S}^*$ é Binomial$(m, \mathcal{P})$.
- Quando a amostra é muito grande ($n \to \infty$), e o número de sítios (m) é bem menor do que o tamanho da amostra, a distribuição de $\mathcal{S}^*$ é Binomial(m, θ/m).
- Quando tanto o tamanho amostral quanto o número de sítios são grandes, mas de tal forma que o número de indivíduos ainda seja bem maior do que o número de sítios ($n >> m$), a distribuição de $\mathcal{S}^*$ é Poisson(θ).

2.5 Teste de Neutralidade Seletiva

Vimos nas Seções 2.2, 2.3 e 2.4 que θ pode ser estimado por $\mathcal{T}_1$, $\mathcal{T}_2$ ou $\mathcal{T}_3$.

A principal diferença entre o número de sítios segregantes (S) e o número médio de diferenças (em pares) de nucleotídeos ($\bar{K}$) é o efeito da seleção. Nucleotídeos mutantes são mantidos na população com baixa freqüência. Como o número de sítios segregantes ignora a freqüência de nucleotídeos mutantes, este valor pode ser fortemente afetado pela existência destes, mesmo que apareçam com baixa freqüência. Por outro lado, a existência de nucleotídeos mutantes com baixa freqüência não afeta o número médio de diferenças de nucleotídeos, pois neste caso a freqüência de mutações é considerada. Em outras palavras, se algumas das mutações observadas têm efeitos seletivos, então o estimador $\mathcal{T}_1$ de θ, que usa o número de sítios segregantes pode não ser o mesmo que $\mathcal{T}_2$, número de diferenças de nucleotídeos.

Fu e Li (1993b) propuseram uma nova abordagem. Considere agora a distribuição das mutações em uma genealogia de uma amostra de genes de uma população. Mutações mais antigas tenderão a ser encontradas na parte mais antiga da genealogia, enquanto as jovens mutações serão encontradas mais provavelmente na parte mais jovem da genealogia. A parte mais antiga da genealogia consiste principalmente dos ramos internos, enquanto que a parte mais jovem, dos ramos mais externos. Na presença de seleção negativa ou purificante, existirá uma tendência

de excesso de mutações nos ramos externos , pois alelos deletérios estão presentes com baixa freqüência. Também é mais provável a existência de mutações em ramos externos se um alelo *advantageous* (alelo que favorece a espécie) tiver se fixado recentemente na população, pois neste caso a maioria das mutações na população são esperadas ser novas. Por outro lado, se uma seleção balanceadora (*overdominant*) estiver atuando no locus, então alguns alelos podem ser antigos e portanto pode haver deficiência de mutações em ramos externos. Portanto, comparar o número de mutações em ramos internos e externos com suas respectivas esperanças sob neutralidade seletiva pode ser uma maneira poderosa de detectar seleção, segundo Fu e Li (1993b). Vimos que o número de *singletones* é uma estatística que leva em consideração o número de mutações nos ramos externos, portanto, é uma estatística interessante para estudar as considerações acima expostas. Assim, Fu e Li (1993b) propuseram estatísticas de teste que consideram a diferença entre os estimadores $\mathcal{T}_1$ e $\mathcal{T}_3$ e entre $\mathcal{T}_2$ e $\mathcal{T}_3$.

Nesta Seção, foi estudada a relação entre o número de sítios segregantes e o número médio de diferenças de nucleotídeos sob a hipótese do modelo de mutação neutra. A relação entre o número de *singletones* e o número de sítios segregantes e entre o número médio de diferenças de nucleotídeos foram estudadas durante os trabalhos de iniciação científica, mas não serão apresentados aqui.

2.5.1 Relação Entre os Dois Estimadores

A covariância entre o número de sítios segregantes e o número médio de diferenças de nucleotídeos é dada por:

$$\mathrm{Cov}(S, \bar{K}) = \mathrm{Cov}(S, K_{ij})\,. \tag{2.92}$$

Esta covariância pode ser obtida pela relação genealógica das seqüências de DNA.

Quando $n = 2$, S é igual a K_{ij} (Figura 2.1a), portanto: $\mathrm{Cov}(S, K_{ij}) = \mathrm{Var}(K_{ij}) =$

$\mathrm{Var}(S)$. De (2.16), temos que $\mathrm{Var}(S) \approx \theta + \theta^2$. Então:

$$\mathrm{Cov}(S, \bar{K}) \approx \theta + \theta^2 \,. \tag{2.93}$$

A relação genealógica quando $n = 3$ está representada na Figura 2.1b. Neste caso, existem dois ancestrais comuns possíveis (A e B) entre as duas seqüências de DNA escolhidas aleatoriamente entre as seqüências. Como B é o ancestral comum quando C e D são escolhidos, e A é o ancestral comum quando C e E ou D e E são escolhidos, a probabilidade de que B seja o ancestral comum é $1/3$, e que A seja o ancestral comum é $2/3$. Portanto, a covariância é dada por:

$$\mathrm{Cov}(S, \bar{K}) = (1/3)\mathrm{Cov}(S, K_{CD}) + (2/3)\mathrm{Cov}(S, K_{CE}),$$

em que $S = K_{BF} + K_{BC} + K_{BD} + K_{EF}$. Usando os resultados da Seção 2.3.3, pode-se escrever:

$$\begin{aligned} \mathrm{Cov}(S, K_{CE}) &= \mathrm{Cov}(K_{BF} + K_{BC} + K_{BD} + K_{EF}, K_{BF} + K_{BC} + K_{EF}) \\ &\approx (4/3)\theta + (7/6)\theta^2 \,. \end{aligned}$$

$$\begin{aligned} \mathrm{Cov}(S, K_{CD}) &= \mathrm{Cov}(K_{BF} + K_{BC} + K_{BD} + K_{EF}, K_{BC} + K_{BD}) \\ &\approx \theta/3 + \theta^2/6 \,. \end{aligned}$$

Usando estas equações:

$$\mathrm{Cov}(S, \bar{K}) = (1/3)\mathrm{Cov}(S, K_{CD}) + (2/3)\mathrm{Cov}(S, K_{CE}) \approx \theta + (5/6)\theta^2 \,. \tag{2.94}$$

Quando $n > 3$, Tajima (1989) mostrou que:

$$\mathrm{Cov}(S, \bar{K}) \approx \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2 \,. \tag{2.95}$$

2.5.1.1 Diferença Entre os Dois Estimadores

Define-se a diferença entre os estimadores da seguinte forma:

$$D_1 = \bar{K} - \frac{S}{a_n} = \mathcal{T}_2 - \mathcal{T}_1 \,. \tag{2.96}$$

Então, a esperança de D_1 é zero e a variância é dada por:

$$\mathrm{Var}(D_1) = \mathrm{Var}(\bar{K}) - \frac{2}{a_n}\mathrm{Cov}(S, \bar{K}) + \frac{1}{a_n^2}\mathrm{Var}(S) \,, \tag{2.97}$$

em que $\mathrm{Var}(\bar{K})$, $\mathrm{Var}(S)$ e $\mathrm{Cov}(S, \bar{K})$ são dadas por (2.63), (2.24) e (2.95), respectivamente. Substituindo os valores em (2.97):

$$\mathrm{Var}(D_1) \approx \frac{n+1}{3(n-1)}\theta + \left[\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n\, n} + \frac{b_n}{a_n^2}\right]\theta^2 \,. \tag{2.98}$$

2.5.2 Estatística para testar a Hipótese de Mutação Neutra

Usa-se a seguinte estatística do teste para testar a hipótese de mutação neutra, que é conhecida como estatística D de Tajima:

$$D = \frac{D_1}{\sqrt{\mathrm{Var}(D_1)}} \,.$$

A variância de D_1, obtida em (2.98), não pode ser usada diretamente, pois depende do parâmetro desconhecido θ. O parâmetro θ pode ser estimado por $\mathcal{T}_1$ ou por $\mathcal{T}_2$. Comparando as variâncias destes dois estimadores, percebe-se que a variância de $\mathcal{T}_1$, equação (2.26), é menor do que a variância de $\mathcal{T}_2$, equação (2.63). Portanto, $\mathcal{T}_1$ será usado como estimador de θ quando a hipótese de mutação neutra é verdadeira, esta será a hipótese nula. No entanto, $\mathcal{T}_1^2$ é um estimador viesado para θ^2, de fato:

$$\mathrm{E}(S^2) = \mathrm{Var}(S) + [\mathrm{E}(S)]^2 \approx a_n\theta + (a_n^2 + b_n)\theta^2, \tag{2.99}$$

que é diferente de $a_n^2\theta^2$ (lembrando que a_n e b_n estão definidos na expressão 2.24).

Como $\mathrm{E}(S^2) - \mathrm{E}(S) \approx (a_n^2 + b_n)\theta^2$, θ^2 pode ser estimado por:

$$\widehat{\theta^2} = \frac{S(S-1)}{a_n^2 + b_n}\,. \tag{2.100}$$

Assim, estima-se $\mathrm{Var}(D_1)$ por:

$$\begin{aligned}
\widehat{\mathrm{Var}(D_1)} &\approx \frac{n+1}{3(n-1)}\widehat{\theta} + \left[\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n\, n} + \frac{b_n}{a_n^2}\right]\widehat{\theta^2} \\
&\approx \frac{n+1}{3(n-1)}\frac{S}{a_n} \\
&+ \left[\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n\, n} + \frac{b_n}{a_n^2}\right]\frac{S(S-1)}{a_n^2 + b_n} \qquad (2.101) \\
&\approx e_1 S + e_2 S(S-1) \quad \text{(para simplificar a notação)} \qquad (2.102)
\end{aligned}$$

Então, a esperança e a variância de D são aproximadamente 0 e 1, respectivamente. Sabendo a distribuição de D, pode-se usá-la para testar a hipótese de mutação neutra.

As distribuições empíricas da estatística D do teste de Tajima para diversos tamanhos amostrais e diferentes valores para o parâmetro θ estão apresentadas no Capítulo 5.

Interpretação da Estatística *D* do Teste de Neutralidade Seletiva

Foi visto que a estatística D utiliza a diferença entre os dois estimadores: $\mathcal{T}_1$ e $\mathcal{T}_2$. Ambos são estimadores consistentes para θ, a não ser que algum processo evolutivo cause uma discrepância entre eles. As discrepâncias ocorrem quando:

- As freqüências relativas de variantes polimórficos são quase idênticas. Este padrão aumenta a proporção de diferenças em pares de nucleotídeos, portanto, $\mathcal{T}_2 - \mathcal{T}_1$ é positiva. Isto sugere algum tipo de seleção balanceada, em que genótipos heterozigotos são favorecidos, ou algum tipo de seleção diversificadora, em que genótipos que carregam alelos menos comuns são favorecidos.

- As freqüências relativas de variantes polimórficos são muito diferentes, com um

excesso do tipo mais comum e uma deficiência do tipo menos comum. Este padrão resulta num decréscimo na proporção de diferenças em pares, logo $\mathcal{T}_2 - \mathcal{T}_1$ é negativa. Uma razão típica para isto pode ser tempo insuficiente desde o gargalo populacional para restaurar o equilíbrio entre mutação e deriva genética aleatória.

3 Métodos de Construção de Árvores Filogenéticas

3.1 Introdução

Neste capítulo apresenta-se o estudo e a comparação dos diferentes métodos de construção de árvores filogenéticas, verificando similaridades e diferenças entre eles e enfatizando a interpretação dos resultados do ponto de vista estatístico.

A filogenética molecular é o estudo das relações evolucionárias entre organismos através de dados moleculares como as seqüências de DNA e proteínas. Os objetivos de estudos filogenéticos são reconstruir a genealogia correta entre entidades biológicas e estimar o tempo de divergência entre organismos (i.e., o tempo desde a última vez que estes compartilhavam um ancestral comum).

Dados moleculares, particularmente seqüências de DNA e aminoácidos, são muito mais adequados para estudos evolucionários do que dados morfológicos e fisiológicos. Primeiro, seqüências de DNA e proteínas são entidades transmitidas estritamente de maneira hereditária. Isto pode não ser verdade para muitos aspectos morfológicos que podem ser influenciados por fatores ambientais. Segundo, a descrição de caracteres moleculares não são ambíguos.

Inferir uma filogenia é um processo de estimação, no qual a estimativa de uma história evolucionária é feita com base em informações incompletas. No contexto da filogenética molecular, geralmente não se tem informações sobre o passado; tem-se acesso apenas às seqüências contemporâneas dos organismos. Como muitas árvores

filogenéticas diferentes podem ser produzidas para uma amostra, deve-se especificar um critério para a seleção de uma ou de algumas árvores que representarão a estimativa da verdadeira história evolucionária. Portanto, uma reconstrução filogenética consiste de duas etapas:

- Definição de uma critério de otimalidade ou função objetivo.
- Criar algoritmos para calcular o valor da função objetivo e identificar a árvore (ou o conjunto de árvores) que tem o melhor valor de acordo com o critério.

3.2 Terminologia de Árvores Filogenéticas

Em estudos filogenéticos, as relações evolucionárias dentre um grupo de organismos são ilustradas através de uma **árvore filogenética**. Uma árvore filogenética é um gráfico composto por **nós** e **ramos**, nos quais apenas um ramo conecta dois nós adjacentes quaisquer (Figura 3.1). Os nós representam as unidades taxonômicas, que podem representar espécies, populações, indivíduos ou genes. Os ramos definem as relações entre as unidades taxonômicas em termos de descendência e ancestralidade. O padrão de ramificação de uma árvore é chamado de **topologia**.

Será feita aqui distinção entre **nós internos** e **nós externos**, e entre **ramos externos** (ramos que terminam em um nó externo) e **ramos internos** (ramos que se estendem de um nó interno a outro). Por exemplo, na Figura 3.1, os nós A, B, C, D e E são externos enquanto todos os outros (F, G e H) são internos. Os ramos AF, BF, CG, DG e EI na Figura 3.1 são externos e os demais ramos são internos. Nós terminais representam as unidades taxonômicas ainda existentes e sob comparação, as quais são referidas como **unidades taxonômicas operacionais (UTOs)**. Os nós internos representam unidades ancestrais inferidas, e desde que não se possui dados empíricos para estas unidades taxonômicas, elas são algumas vezes referidas como **unidades taxonômicas hipotéticas (UTHs)**.

Um nó é **bifurcado** se possui apenas duas linhagens descendentes imediatas

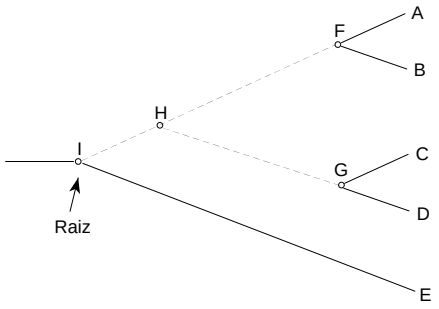
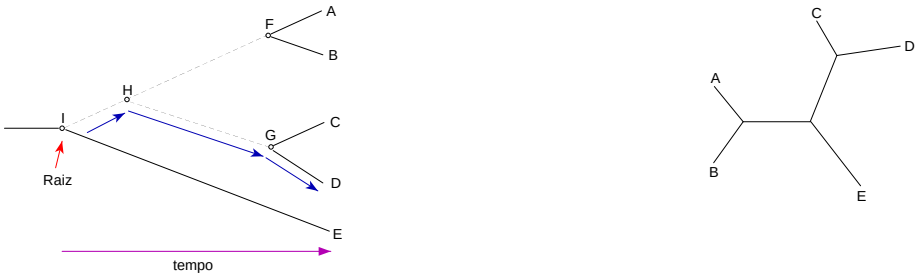


Figura 3.1: Uma árvore filogenética ilustrando relações evolucionárias entre cinco UTOs (A - E). Os círculos denotam nós internos. Os nós internos (F - H) representam as UTHs. O nó I é a raiz.

e é **multifurcado** se possui mais do que duas linhagens descendentes imediatas. Em uma árvore estritamente bifurcada, cada nó interno é ligado a exatamente três ramos, dois descendentes e um ancestral. Em estudos evolucionários assume-se que o processo é geralmente binário, i.e., cada espécie é capaz de gerar não mais que duas espécies a um único tempo. Então, a representação comum de filogenia emprega árvores bifurcadas, nas quais cada ancestral se divide em dois descendentes. Há duas interpretações possíveis para uma multifurcação (ou politomia) em uma árvore: ou representa a seqüência real dos eventos, em que um ancestral tenha dado origem a três ou mais descendentes simultaneamente, ou representa uma situação em que a ordem exata de duas ou mais bifurcações não pode ser determinada de maneira não ambígua através dos dados disponíveis.

Árvores podem ou não apresentar raiz. Em uma árvore **com raiz** existe um nó em particular chamado **raiz**, do qual um único caminho atinge qualquer outro nó (Figura 3.2a).

A direção de cada caminho corresponde ao tempo evolucionário, e a raiz é o



(a) (b)

Figura 3.2: Árvore com raiz (a) e sem raiz (b). A seta indica o único caminho da raiz à UTO D.

ancestral comum mais recente de todas as unidades taxonômicas sob estudo. Uma árvore **sem raiz** é uma árvore que apenas especifica o grau de relacionamento entre as unidades taxonômicas mas não define o caminho evolucionário (Figura 3.2b).

Então, estritamente falando, uma árvore sem raiz pode não ser considerada em si mesma uma árvore filogenética, já que o sentido do tempo não é especificado.

Para três espécies existem três tipos diferentes de árvores com raiz. Para quatro espécies, tem-se quinze possíveis árvores com raiz. O número de árvores com raiz que apresentam bifurcação (N_R) quando consideramos n UTOs é dado por:

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}, \tag{3.1}$$

quando $n \geq 2$ (CAVALLI-SFORZA; EDWARDS, 1967). O número de possíveis árvores sem raiz para n UTOs é igual ao número de possíveis árvores com raiz para $n-1$ UTOs. Para vinte UTOs existem aproximadamente 10^{22} árvores com raiz. Desde que somente uma destas representa a verdadeira relação evolucionária entre as UTOs, geralmente é muito difícil identificar a verdadeira árvore filogenética quando

n é grande.

A seqüência de eventos de especiação que leva à formação de qualquer grupo de UTOs é historicamente único. Assim, somente uma de todas as possíveis árvores que podem ser construídas dado um certo número de UTOs é a verdadeira história evolucionária. Tal árvore filogenética é chamada de **árvore verdadeira**. Uma árvore que é obtida através de um certo conjunto de dados e um método de reconstrução de árvore é chamada **árvore inferida**, esta pode ou não ser idêntica à árvore verdadeira.

3.3 Distância entre seqüências de DNA

Antes de analisar os métodos clássicos de construção de árvores filogenéticas, faz-se necessário conhecer alguns dos métodos mais utilizados para registrar distâncias entre seqüências de DNA.

3.3.1 Distância de Hamming

A distância de Hamming é muito utilizada como análise descritiva (SEILLIER-MOISEIWITSCH; MARGOLIN; SWANSTROM, 1994) e como medida de distância para desenvolvimento de métodos de análise de variância para dados categorizados (PINHEIRO; SEILLIER-MOISEIWITSCH; SEN, 2000). Seja $(X_{i1}, X_{i2}, ..., X_{iK})$' um vetor representando a seqüência i de tamanho K. X_{ik} é então o nucleotídeo ou aminoácido presente na posição k. Considere $\mathbf{X_i}$ e $\mathbf{X_{i'}}$. A distância de Hamming $H_{ii'}$ é:

$$\begin{aligned} H_{ii'} &= \frac{1}{K}\sum_{k=1}^{K}\delta(X_{ik} \neq X_{i'k}) \\ &= \frac{1}{K} \times \text{número de posições onde } \mathbf{X_i} \text{ e } \mathbf{X_{i'}} \text{ diferem,} \end{aligned} \tag{3.2}$$

em que δ representa a função indicadora ($\delta(A) = 1$ se o evento A é verdade e 0 caso contrário). Embora essa distância deva ser tratada como uma estatística descritiva, em muitas situações ela fornece uma estimativa razoável da distância real quando as seqüências estão intimamente relacionadas(i.e., são separadas por poucas replicações, de tal forma que em uma posição específica, é muito raro terem ocorrido ambas: uma mutação reversa e uma do tipo "forward").

3.3.2 Distância log determinante

Para calcular distâncias entre seqüências com diferentes composições de nucleotídeos ou aminoácidos, (LOCKHART, 1994) introduziu a *distância log determinante*, baseada na chamada matriz de divergência F_{xy}. Para seqüências x e y, seu elemento (i, j) é a proporção de sítios em que x está na categoria (i.e., nucleotídeo ou aminoácido) i enquanto y é j. A soma de todos os elementos da matriz é 1. A distância *LogDet* entre x e y é definida como:

$$d_{xy} = \ln(\det F_{xy}), \tag{3.3}$$

em que det é o determinante da matriz. Para assinalar uma distância de zero entre uma seqüência e ela mesma, esta quantidade é modificada; para seqüências de nucleotídeos torna-se

$$d'_{xy} \equiv -\frac{1}{4} \ln \left(\frac{\det F_{xy}}{\sqrt{(\det F_{xx})(\det F_{yy})}} \right) .$$

Note que quando as quatro freqüências de bases são iguais, $\det F_{xx} = \det F_{yy} = (1/4)^4$, e o valor esperado da distância *LogDet* é o número médio de substituições por sítio.

3.3.3 Distâncias baseadas em modelos

Para estudos comparativos de seqüências de DNA, métodos estatísticos têm sido utilizados para estimar o número de substituições de nucleotídeos. Estes métodos são úteis para ajustar mutações que podem ter ocorrido, mas não foi possível observar, como substituições paralelas ou reversas.

O modelo mais simples de substituição de DNA (JUKES; CANTOR, 1969) assume que as freqüências de bases são iguais ($\pi_A = \pi_C = \pi_G = \pi_T$) e que as taxas de mudanças (Figura 3.3) são iguais ($r_1 = r_2 = ... = r_{12}$).

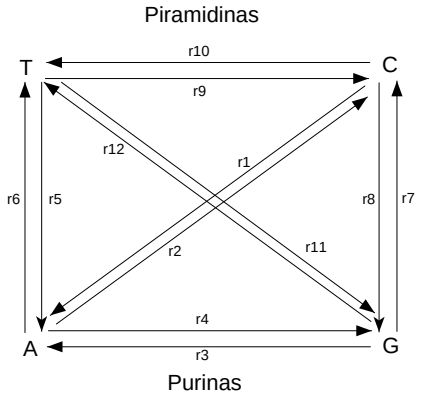


Figura 3.3: Taxas de substituição de nucleotídeos.

Muitos modelos evolucionários de substituição de DNA podem ser encontrados na literatura. A escolha do modelo depende das suposições que o biólogo quer fazer. Por exemplo, alguns biólogos dizem que a taxa de transição é mais alta do que a taxa de transversão ou pode-se assumir que para alguns organismos as taxas de mudança são sempre as mesmas.

Considere o modelo de Jukes-Cantor em que $r_1 = r_2 = ... = r_{12} = \mathsf{a}$. Um modelo geral de substituições de DNA pode ser representado pela matriz de taxas imediatas $\mathbf{Q}$ na forma:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} . & r_2\pi_C & r_4\pi_G & r_6\pi_T \\ r_1\pi_A & . & r_8\pi_G & r_{10}\pi_T \\ r_3\pi_A & r_7\pi_C & . & r_{12}\pi_T \\ r_5\pi_A & r_9\pi_C & r_{11}\pi_G & . \end{pmatrix}$$

em que q_{ij} representa a taxa de mudança do nucleotídeo i para o nucleotídeo j. Por exemplo, $r_2\pi_C$ representa a taxa de mudança de "A" para "C".

Considere duas seqüências homólogas que divergiram t unidades de tempo atrás. Denota-se por I(t) a probabilidade de que dois nucleotídeos em sítios correspondentes são idênticos no tempo t. Supõe-se que a taxa de substituição é a mesma para todos os pares de nucleotídeos e não varia com o tempo (JUKES; CANTOR, 1969). Denota-se esta taxa por a. Desta maneira, para um certo sítio, seja B o evento: dois nucleotídeos homólogos permanecem idênticos no tempo $t + \Delta t$, sendo que eram idênticos no tempo t.

$$P(B) = [(1 - 3\mathsf{a}\Delta t)^2 + 3(\mathsf{a}\Delta t)^2]I(t).$$

O evento B é a união de dois outros eventos mutuamente exclusivos:

- Evento C: ambas as bases de nucleotídeos (idênticas) mudam para outras duas bases idênticas, o que ocorre com probabilidade $3(\mathsf{a}\Delta t)^2$.

- Evento D: ambas as bases de nucleotídeos (idênticas) não mudam, o que ocorre com probabilidade $(1 - 3\mathsf{a}\Delta t)^2$.

Os eventos C e D estão representados na Figura 3.4, em que $\Delta t = 1$.

Seja E o evento: dois nucleotídeos homólogos tornam-se idênticos no tempo $t + \Delta t$, sendo que eram diferentes no tempo t.

$$P(E) = [2\mathsf{a}\Delta t(1 - 3\mathsf{a}\Delta t) + 2(\mathsf{a}\Delta t)^2][1 - I(t)].$$

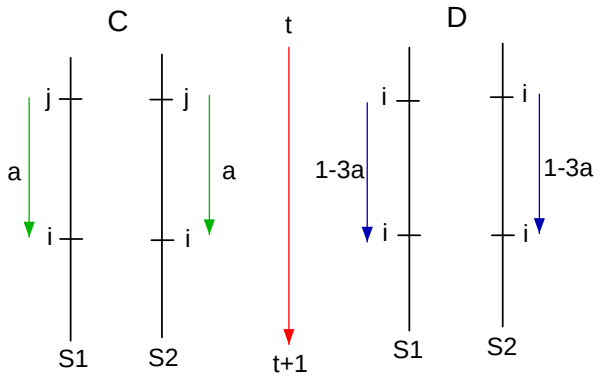


Figura 3.4: Eventos C e D, representando duas das possíveis situações de substituição de nucleotídeos em duas seqüências.

A probabilidade de que dois nucleotídeos são diferentes no tempo t é $1 - I(t)$. O evento E é a união de dois outros eventos mutuamente exclusivos:

- Evento F: uma das bases de nucleotídeo permanece igual e a outra muda para uma base igual à base do nucleotídeo homólogo, o que ocorre com probabilidade $2\mathsf{a}\Delta t(1 - 3\mathsf{a}\Delta t)$.

- Evento G: ambas as bases mudam para outras duas bases idênticas. O que ocorre com probabilidade $2(\mathsf{a}\Delta t)^2$.

Os eventos F e G estão representados na Figura 3.5, em que $\Delta t = 1$.

Desta maneira, a probabilidade de que dois nucleotídeos em sítios correspondentes são idênticos no tempo $t + \Delta t$ é:

$$\begin{aligned} I(t+\Delta t) &= P(B) + P(E) \\ &= [(1-3\mathsf{a}\Delta t)^2 + 3(\mathsf{a}\Delta t)^2]I(t) + [2\mathsf{a}\Delta t(1-3\mathsf{a}\Delta t) + 2(\mathsf{a}\Delta t)^2][1-I(t)] \\ &= [1 - 6\mathsf{a}\Delta t + 12\mathsf{a}^2(\Delta t)^2]I(t) + [2\mathsf{a}\Delta t - 4\mathsf{a}^2(\Delta t)^2[1-I(t)]. \end{aligned}$$

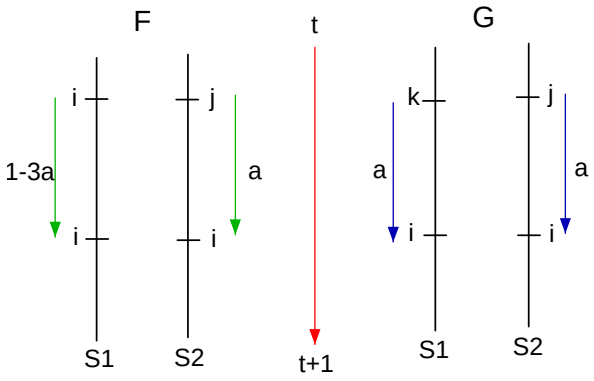


Figura 3.5: Eventos F e G, representando duas das possíveis situações de substituição de nucleotídeos em duas seqüências.

$$I(t+\Delta t) - I(t) = -8\mathsf{a}\Delta t I(t) + 16\mathsf{a}^2(\Delta t)^2 I(t) + 2v\Delta t - 4\mathsf{a}^2(\Delta t)^2 .$$

$$\lim_{\Delta t\to 0} \frac{I(t+\Delta t) - I(t)}{\Delta t} = \lim_{\Delta t\to 0}[-8\mathsf{a}I(t) + 16\mathsf{a}^2\Delta t I(t) + 2\mathsf{a} - 4\mathsf{a}^2\Delta t] .$$

$$\frac{dI(t)}{dt} = -8\mathsf{a}I(t) + 2\mathsf{a},$$

que é uma equação diferencial de primeira ordem. Reescrevendo-a, tem-se:

$$I'(t) + 8\mathsf{a}I(t) = 2\mathsf{a}. \tag{3.4}$$

O fator integrante é $M(t) = \exp\left[\int 8\mathsf{a}dt\right] = e^{8\mathsf{a}t}$. Multiplicando o fator integrante à

equação (3.4), temos:

$$\begin{aligned} e^{8\mathsf{a}t}I'(t) + 8\mathsf{a}e^{8\mathsf{a}t}I(t) &= 2\mathsf{a}e^{8\mathsf{a}t} \\ \Rightarrow \left(I(t)e^{8\mathsf{a}t}\right)' &= 2\mathsf{a}e^{8\mathsf{a}t} \\ \Rightarrow I(t)e^{8\mathsf{a}t} &= \int 2\mathsf{a}e^{8\mathsf{a}t}dt + C \\ &= \frac{2\alpha}{8\mathsf{a}}\, e^{8\mathsf{a}t} + C \\ \Rightarrow I(t) &= \frac{1}{4} + Ce^{-8\mathsf{a}t}. \end{aligned}$$

Se temos a condição inicial $I(0) = 1$:

$$I(0) = \frac{1}{4} + Ce^0 \Rightarrow C = \frac{3}{4},$$

e a solução da equação é:

$$I(t) = \frac{1}{4} + \frac{3}{4}\, e^{-8\mathsf{a}t}\,. \tag{3.5}$$

Se a condição inicial é $I(0) = 0$:

$$I(0) = \frac{1}{4} + Ce^0 \Rightarrow C = -\frac{1}{4},$$

e a solução da equação é:

$$I(t) = \frac{1}{4} - \frac{1}{4}\, e^{-8\mathsf{a}t}.$$

A solução geral é dada por:

$$I(t) = \frac{1}{4} + \left[I(0) - \frac{1}{4}\right] e^{-8\mathsf{a}t}.$$

A probabilidade de que as duas seqüências sejam diferentes em um sítio no tempo t é:

$$p = 1 - I(t).$$

Assim, usando a equação (3.5) tem-se que:

$$p = \frac{3}{4}\left(1 - e^{-8\mathsf{a}t}\right) . \tag{3.6}$$

A equação (3.6) pode ser reescrita como:

$$\begin{aligned} \frac{4}{3}p &= 1 - e^{-8\mathsf{a}t} \\ \Rightarrow e^{-8\mathsf{a}t} &= \left(1 - \frac{4}{3}p\right) \\ \Rightarrow 8\mathsf{a}t &= -\ln\left(1 - \frac{4}{3}p\right) . \end{aligned} \tag{3.7}$$

Uma vez que o tempo de divergência entre duas seqüências é geralmente desconhecido, não é possível estimar a. Calcula-se então K, que é o *número de substituições por sítio ocorridas desde o início da divergência entre as duas seqüências*. No caso do modelo Jukes-Cantor (1 parâmetro), tem-se:

$$K = 2(3\alpha t) ,$$

em que $3\alpha t$ é o número de substituições por sítio em cada uma das duas linhagens.

Usando a equação (3.7), pode-se calcular K da seguinte forma:

$$\begin{aligned} 6\alpha t &= \frac{6}{8}\left[-\ln\left(1 - \frac{4}{3}p\right)\right] \\ \Rightarrow K &= -\frac{3}{4}\left[\ln\left(1 - \frac{4}{3}p\right)\right] , \end{aligned}$$

em que p é a proporção de nucleotídeos diferentes entre as duas seqüências.

3.4 Métodos de Construção de Árvores Filogenéticas

Existem vários métodos para construção de árvores filogenéticas: método do vizinho mais próximo (SAITOU; NEI, 1987), UPGMA, que será visto nesta Seção com maiores detalhes, método da máxima parcimônia (FITCH, 1977) e método da máxima verossimilhança, o mais caro computacionalmente (LI; GRAUR, 1991). Para os modelos lineares que serão vistos no Capítulo 4, interessam aqueles métodos que fornecem uma árvore com raiz.

Dentre os métodos, o mais simples computacionalmente é o UPGMA - *"Unweighted pair-group with arithmetic means"*. Além disso, este método fornece uma árvore com raiz. Foi originalmente desenvolvido para construir fenogramas taxonômicos, i.e., árvores que refletem as similaridades fenotípicas entre UTO's (SOKAL; MICHENER, 1958), mas pode também ser usado para construir árvores filogenéticas quando as taxas de evolução são aproximadamente constantes entre as diferentes linhagens, de maneira que exista uma relação aproximadamente linear entre distância evolucionária e tempo de divergência (NEI, 1975).

O UPGMA emprega um algoritmo seqüencial de agrupamento (*clustering*), no qual relações topológicas locais são identificadas em ordem decrescente de similaridade, e a árvore filogenética é construída passo-a-passo. Em outras palavras, primeiramente são identificadas dentre as UTOs as duas mais similares entre si e estas são tratadas como uma nova UTO. Uma unidade como esta será chamada de **UTO composta**. Para o novo grupo de UTOs é construída uma nova matriz de distâncias e é identificado o par com maior similaridade. Este processo é repetido até que tenha-se apenas duas UTOs.

Para ilustrar o método, considere um caso de quatro UTOs, A, B, C e D. As distâncias evolucionárias entre os pares são dadas pela seguinte matriz:

UTO	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{BD}

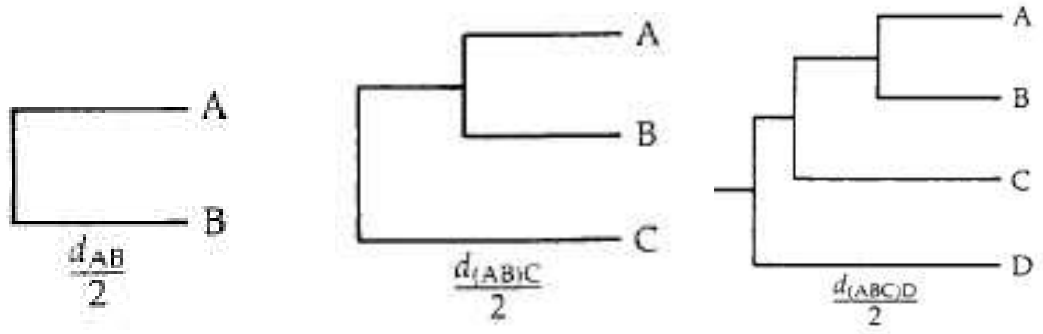
Nesta matriz, d_{ij} representa a distância entre as UTOs i e j. As duas primeiras UTOs a serem agrupadas são aquelas com menor distância. Assuma que d_{AB} é a menor. Então, as UTOs A e B serão as primeiras a serem agrupadas, e o ponto de ramificação, l_{AB}, é posicionado à distância de $d_{AB}/2$ substituições (Figura 3.6a). Após o primeiro agrupamento, A e B são consideradas como uma única UTO composta (AB), e a nova matriz de distâncias é calculada:

UTO	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

Nesta matriz, $d_{(AB)C} = (d_{AC} + d_{BC})/2$, e $d_{(AB)D} = (d_{AD} + d_{BD})/2$. Em outras palavras, a distância entre uma UTO simples e uma UTO composta é a média aritmética das distâncias entre a UTO simples e as que constituem a UTO composta. Se $d_{(AB)C}$ vier a ser a menor distância na nova matriz, então a UTO C será agrupada à UTO composta (AB) com um nó de ramificação em $l_{(AB)C} = d_{(AB)C}/2$ (Figura 3.6b).

O passo final consiste em agrupar a última UTO, D, à nova UTO composta, (ABC). A raiz da árvore toda está posicionada a $l_{(ABC)D} = d_{(ABC)D}/2 = [(d_{AD} + d_{BD} + d_{CD})/3]/2$. A árvore final construída a partir do UPGMA é mostrada na Figura 3.6c.

No UPGMA, o ponto de ramificação entre duas UTOs simples, i e j, está posicionado à metade da distância entre elas:



(a) (b) (c)

Figura 3.6: Diagrama ilustrativo da construção passo-a-passo de uma árvore filogenética para quatro UTOs utilizando UPGMA.

$$l_{ij} = \frac{d_{ij}}{2}.$$

O ponto de ramificação entre uma UTO simples, i, e uma UTO composta, (jm), está posicionado à metade da média aritmética das distâncias entre a UTO simples e as UTOs que compõem a UTO composta.

$$l_{(i)(jm)} = \frac{(d_{ij} + d_{im})/2}{2}$$

O ponto de ramificação entre duas UTOs compostas está posicionado à metade da média aritmética das distâncias entre as UTOs simples que constituem cada UTO composta. Por exemplo, a posição do ponto de ramificação entre as UTOs compostas (ij) e (mn) é

$$l_{(ij)(mn)} = \frac{(d_{im} + d_{in} + d_{jm} + d_{jn})/4}{2}.$$

No caso de uma UTO composta tripartida, (ijk), e uma UTO composta bipartida, (mn), a posição do ponto de ramificação é

$$l_{(ijk)(mn)} = \frac{(d_{im} + d_{in} + d_{jm} + d_{jn} + d_{km} + d_{kn})/6}{2}.$$

O UPGMA é um dos poucos métodos de reconstrução filogenética que produz uma árvore com raiz. Note que utilizando UPGMA pode-se obter simultaneamente a topologia da árvore e o comprimento dos ramos. É um método que funciona bem quando a suposição de que a taxa de mutação permanece razoavelmente constante é válida.

4 *Estimadores de Polimorfismo Utilizando Informações Filogenéticas*

Estimar θ utilizando estimadores apresentados no capítulo 2 é simples do ponto de vista computacional. Contudo, o preço por essa simplicidade é uma variância grande. Neste capítulo apresentam-se os estimadores propostos por Fu (1994), mais eficientes por fazerem uso da informação filogenética em uma amostra de seqüências de DNA.

4.1 Estimação de θ por máxima verossimilhança

Fu e Li (1993a) obtiveram a variância mínima do estimador de θ quando observa-se o número de mutações em cada ramo da genealogia.

4.1.1 Densidade conjunta de eventos evolucionários

Suponha que uma amostra de n genes é retirada de uma população com cruzamento aleatório. Sob a suposição de mutações neutras, o processo que governa a evolução das seqüências amostradas é inteiramente determinado pelo valor de θ. Este processo é conhecido como processo de coalescência.

Utiliza-se o modelo de Wright-Fisher para a população e assume-se as suposições descritas na Seção 2.2. Por conveniência, o tempo em que a amostra foi tomada (nós

externos) é considerado o n-ésimo evento de ramificação (Figura 4.1). Como visto na Seção 2.3.1.1, $T_i \sim \text{Exp}\left(\frac{i(i-1)}{4N}\right)$ representa o tempo de coalescência entre o $(i-1)$-ésimo e o i-ésimo nó.

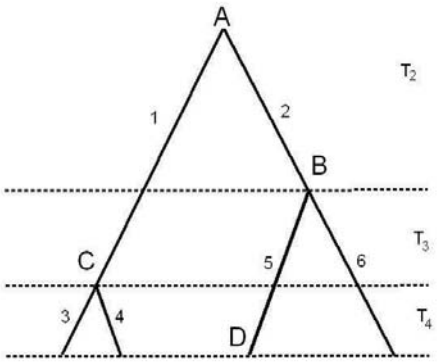


Figura 4.1: Exemplo de topologia para quatro seqüências. O primeiro, segundo, terceiro e quarto eventos de ramificação são A, B, C e D, respectivamente.

$$f_{T_i}(t_i) = \frac{i(i-1)}{4N}\, e^{-\frac{i(i-1)}{4N}t_i}, \quad t_i = 0,\, 1, \ldots, \quad i = 2,\, 3, \ldots$$

Considere a genealogia de uma amostra aleatória de uma população com cruzamento ao acaso. Existem $n-1$ nós internos na árvore enumerados de 2 a $n-1$ de acordo com suas ordens de ocorrência (B e C na Figura 4.1). Portanto, entre o $(i-1)$-ésimo e o i-ésimo nó existem exatamente i ramos, enumerados de 1 a i. O tempo desde o $(i-1)$-ésimo nó até o i-ésimo é o tempo de coalescência T_i. Define-se Y_{ij} como o número de mutações ocorrendo no j-ésimo ramo dentre os i ramos existentes durante o tempo de coalescência T_i. Assume-se que $Y_{ij} \sim \text{Poisson}(\nu t_i)$:

$$P(Y_{ij} = y_{ij} \mid t_i) = \frac{e^{-\nu t_i}(\nu t_i)^{y_{ij}}}{y_{ij}!}, \quad y_{ij} = 0,\, 1, \ldots$$

Seja $Y_{i1}, Y_{i2}, \ldots Y_{ii}$ condicionadas em t_i uma amostra aleatória com distribuição

Poisson(νt_i). Então, sabe-se que $Y_i = \sum_{j=1}^{i} Y_{ij}$ é estatística suficiente e $Y_i \mid t_i \sim$ Poisson($\nu i t_i$). Y_i é o número total de mutações que ocorreu entre o $(i-1)$-ésimo e o i-ésimo nó, ou seja, durante o tempo de coalescência T_i.

É razoável assumir que a distribuição espacial de um certo número de mutações entre os sítios de um gene é independente do parâmetro θ, embora o número de mutações dependa de θ. Por exemplo, pode-se assumir que a distribuição espacial é uniforme entre todos os sítios, ou outra distribuição qualquer que não dependa de θ. Portanto, toda a informação relevante sobre o valor de θ em uma amostra de n genes está contida no vetor

$$\mathbf{\Psi} = \{T_i,\, Y_i :\, i = 2, ..., n\}\,,$$

cujos elementos assumem valores não negativos.

A função de densidade conjunta das quantidades em $\mathbf{\Psi}$ é

$$\begin{aligned} f_{\mathbf{\Psi}}(\boldsymbol{\psi}) &= \prod_{i=2}^{n} P(Y_i = y_i \mid \nu i t_i) f_{T_i}(t_i) \\ &= \prod_{i=2}^{n} \frac{e^{-\nu i t_i}(\nu i t_i)^{y_i}}{y_i!} \frac{i(i-1)}{4N} e^{-\frac{i(i-1)}{4N} t_i} \\ &= \exp\left[-\nu \sum_{i=2}^{n} i t_i - \frac{\nu}{\theta} \sum_{i=2}^{n} i(i-1) t_i - (n-1)\log\theta\right] \\ &\times \exp\left[\left(\sum_{i=2}^{n} y_i + n - 1\right)\log\nu + \sum_{i=2}^{n} \log\left(\frac{i-1}{y_i!}\right)\right]. \end{aligned} \tag{4.1}$$

As funções densidade dos Y_i's já foram apresentadas na Seção 2.2.2.

A função de densidade conjunta de $\mathbf{Y} = (Y_2, ..., Y_n)$ é

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=2}^{n} \left(\frac{i-1}{\theta + i - 1}\right)\left(\frac{\theta}{\theta + i - 1}\right)^{y_i}. \tag{4.2}$$

Uma situação ideal seria aquela em que pode-se observar todo o processo de evolução de uma amostra de genes, da raiz da árvore até o presente. Nesta situação

idealizada, pode-se contar não apenas o número de mutações em cada ramo ($\{Y_i, i = 2, ...n\}$), mas também o número de gerações ($\{T_i, i = 2, ...n\}$) entre os nós sucessivos. Um estimador ótimo de θ com estes dois conjuntos de dados terá uma variância menor do que os estimadores que contenham menos informações do processo. O estimador ótimo pode ser obtido através do método de máxima verossimilhança, pois este tem boas propriedades assintóticas.

Considera-se aqui uma situação menos ideal, mas mais interessante do ponto de vista prático, na qual observa-se somente o número de mutações em cada ramo da genealogia, ou seja, tem-se informação sobre ($\{Y_i, i = 2, ...n\}$), mas não sobre o tempo de coalescência ($\{T_i, i = 2, ...n\}$).

Definição 3. *Uma família de distribuições $\{P_\theta : \theta \in \Theta\}$ é dita ser uma família exponencial k paramétrica, se existirem funções reais $c_1, \ldots, c_k$, d de θ, funções reais $T_1, \ldots, T_k$, S em $\mathbb{R}^n$ e um conjunto $A \subset \mathbb{R}^n$ tais que as funções de densidade de P_θ podem ser escritas como:*

$$f(\boldsymbol{X}, \theta) = \left\{ \exp \left[\sum_{i=1}^{k} c_i(\theta) T_i(\boldsymbol{x}) + d(\theta) + S(\boldsymbol{x}) \right] \right\} I\!I_a(\boldsymbol{x}) \, .$$

Proposição 2. *Se a distribuição de uma variável aleatória pertence à família exponencial e $c(\theta)$ tem derivada contínua em Θ, então as condições de regularidade são satisfeitas.*

Teorema 2. ***(Desigualdade da Informação):*** *Seja $T(\boldsymbol{X})$ um estimador não viesado de θ com variância finita para todo θ. Suponha que as condições de regularidade sejam satisfeitas e que*

$$E \left\{ \left[\frac{\partial}{\partial \theta} \log f(\boldsymbol{x}, \theta) \right]^2 \right\} ,$$

seja positiva e finita. Então, para todo θ,

$$Var(T(\boldsymbol{X})) \geq \frac{1}{E \left\{ \left[\frac{\partial}{\partial \theta} \log f(\boldsymbol{x}, \theta) \right]^2 \right\}} \, .$$

A quantidade

$$\frac{1}{E\left\{\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{x},\theta)\right]^2\right\}}$$

é denominada limite inferior de Cramer-Rao (LICR).

Reescrevendo (4.2), tem-se:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \exp\left\{\sum_{i=1}^{n} y_i \log\left(\frac{\theta}{\theta+i-1}\right) + \sum_{i=1}^{n}\log\left(\frac{i-1}{\theta+i-1}\right)\right\}, \quad y_i \geq 0, \quad i=2,\ldots,n \\ c_i(\theta) &= \log\left(\frac{\theta}{\theta+i-1}\right) \qquad T_i(\mathbf{y}) = y_i \\ d(\theta) &= \sum_{i=2}^{n}\log\left(\frac{i-1}{\theta+i-1}\right) \qquad S(\mathbf{y}) = 0\,. \end{aligned}$$

A função de verossimilhança neste caso é dada por (4.2) e sua função de log-verossimilhança é

$$\log L(\theta,\mathbf{y}) = \sum_{i=2}^{n}\left\{y_i\log\left(\frac{\theta}{i-1}\right) - (y_i+1)\log\left(\frac{\theta+i-1}{i-1}\right)\right\}.$$

$$\frac{d\log L(\theta,\mathbf{y})}{d\theta} = \sum_{i=2}^{n}\left(\frac{y_i}{\theta} - \frac{y_i+1}{\theta+i-1}\right).$$

Portanto, o estimador de máxima verossimilhança para θ é a solução da equação:

$$\sum_{i=2}^{n}\frac{y_i+1}{\widehat{\theta}_m+i-1} = \frac{V_n}{\widehat{\theta}_m}.$$

Segunda derivada:

$$\frac{d^2\log L(\theta,\mathbf{y})}{d\theta} = \sum_{i=2}^{n}\left(-\frac{y_i}{\theta^2} + \frac{y_i+1}{(\theta+i-1)^2}\right) = -\frac{v_n}{\theta^2} + \sum_{i=2}^{n}\frac{y_i+1}{(\theta+i-1)^2}.$$

Para calcular o limite inferior da variância de $\widehat{\theta}_m$, calcula-se a informação de Fisher:

$$\mathrm{E}\left(-\frac{d^2 \log L(\theta, \mathbf{Y})}{d\theta^2}\right) = \frac{a_n}{\theta} - \sum_{i=1}^{n} \frac{1}{i(\theta+i)}\,.$$

$$\begin{aligned} \mathrm{Var}(\widehat{\theta}_m) &= 1/\mathrm{E}\left(-\frac{d^2 \log L(\theta, \mathbf{Y})}{d\theta^2}\right) \\ &= \frac{\theta}{a_n}\left(1 + \frac{\alpha}{1-\alpha}\right) \\ &= \frac{\theta}{\sum_{k=1}^{n-1} \frac{1}{\theta+k}}\,, \end{aligned} \tag{4.3}$$

em que $\alpha = 1 - \frac{1}{a_n}\sum_{k=1}^{n-1} \frac{1}{\theta+k}$.

Assim, o LICR dado por (4.3) pode ser usado para comparar a eficiência dos estimadores.

4.2 Estimação de θ quando a genealogia de uma amostra é conhecida

Assume-se nesta Seção que a genealogia dos genes em uma amostra aleatória de uma população é conhecida. Por genealogia, entende-se a topologia de ligação dos genes em uma amostra até seu ancestral comum mais recente, a ordem das ramificações na topologia e o número de mutações em cada ramo da topologia.

4.2.1 Modelo Linear para o Número de Mutações em cada Ramo

A fonte de informação mais completa que pode-se obter na prática é o número de mutações em cada ramo da árvore genealógica.

Considere que as ramificações sejam enumeradas sucessivamente de tal maneira que o primeiro evento de ramificação é a raiz e que o $(n-1)$-ésimo é a ramificação mais recente. Por conveniência, o tempo em que a amostra foi tomada (nós externos)

é considerado como a n-ésima ramificação (Figura 4.1). Como visto na Seção 2.3.1.1, T_k é uma variável aleatória com distribuição aproximadamente Exponencial com parâmetro $k(k-1)/4N$ que representa o tempo de coalescência entre o $(k-1)$-ésimo e o k-ésimo evento de ramificação.

A genealogia de de n genes tem exatamente $2(n-1)$ ramos. Para o ramo i, são definidas variáveis indicadoras c_{ik} da seguinte maneira:

$$c_{ik} = \begin{cases} 1 & \text{se o ramo } i \text{ tem um segmento no tempo de coalescência } T_k \\ 0 & \text{caso contrário.} \end{cases} \tag{4.4}$$

$i = 1, 2, ..., 2(n-1)$ e $k = 2, ..., n$. A topologia de uma genealogia de n genes é completamente caracterizada por essas $2(n-1)^2$ variáveis indicadoras.

Tabela 4.1: c_{ik} para a topologia da Figura 4.1.

	k		
i	2	3	4
1	1	1	0
2	1	0	0
3	0	0	1
4	0	0	1
5	0	1	1
6	0	1	1

Seja l_i o comprimento do ramo i (em termos de número de gerações) e m_i o número de mutações que ocorreram no ramo i. Como visto no início da Seção 2.2, a taxa de mutação do gene é ν. Assume-se, em geral, que o número de mutações que ocorre em uma seqüência durante o período de l gerações segue a distribuição de Poisson com esperança νl. Desta maneira, o número de mutações que ocorre no ramo i, m_i, dado o comprimento deste ramo, l_i, segue a distribuição de Poisson:

$$P(m_i = a \mid l_i) = \frac{e^{-\nu l_i}(\nu l_i)^a}{a!}, \quad a = 0, 1, \ldots$$

A variável aleatória l_i é definida da seguinte maneira:

$$l_i = \sum_{k=2}^{n} c_{ik} T_k \, .$$

Então,

$$\begin{aligned} \mathrm{E}(m_i) &= \theta x_i \, , \\ \mathrm{Var}(m_i) &= \theta x_i + \theta^2 z_{ii} \, , \\ \mathrm{Cov}(m_i, m_j) &= \theta^2 z_{ij} \, . \end{aligned} \tag{4.5}$$

em que

$$x_i = \sum_{k} \frac{c_{ik}}{k(k-1)} \, , \tag{4.6}$$

$$z_{ij} = \sum_{k} \frac{c_{ik} c_{jk}}{k^2 (k-1)^2} \, . \tag{4.7}$$

De fato:

$$\begin{aligned} \mathrm{E}(l_i) &= \mathrm{E}\left(\sum_{k=2}^{n} c_{ik} T_k \right) = \sum_{k=2}^{n} c_{ik} \mathrm{E}(T_k) \\ &= \sum_{k=2}^{n} c_{ik} \frac{4N}{k(k-1)} \, . \end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}(l_i) &= \mathrm{E}(l_i^2) - [\mathrm{E}(l_i)]^2\,, \\
E(l_i^2) &= \mathrm{E}\left[\left(\sum_k c_{ik}T_k\right)^2\right] = \sum_{m\neq n} c_{im}c_{in}\mathrm{E}(T_mT_n) + \sum_k c_{ik}^2\mathrm{E}(T_k^2) \\
&= \text{de (2.44)} = \sum_{m\neq n} c_{im}c_{in}\mathrm{E}(T_mT_n) + 2\sum_k c_{ik}^2[\mathrm{E}(T_k)]^2 \\
&= \underbrace{\sum_{m\neq n} c_{im}c_{in}\mathrm{E}(T_mT_n) + \sum_k c_{ik}^2[\mathrm{E}(T_k)]^2}_{[\sum_k c_{ik}\mathrm{E}(T_k)]^2} + \sum_k c_{ik}^2[\mathrm{E}(T_k)]^2 \\
&= \left(\sum_k c_{ik}\frac{4N}{k(k-1)}\right)^2 + \sum_k c_{ik}^2\frac{(4N)^2}{k^2(k-1)^2}\,, \\
\Rightarrow \mathrm{Var}(l_i) &= \sum_k c_{ik}^2\frac{(4N)^2}{k^2(k-1)^2} = \sum_k c_{ik}\frac{(4N)^2}{k^2(k-1)^2}\,.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(m_i) &= \mathrm{E}(\mathrm{E}(m_i \mid l_i)) = \mathrm{E}(\nu l_i) = \nu\mathrm{E}(l_i) = \underbrace{4N\nu}_{\theta}\sum_k \frac{c_{ik}}{k(k-1)} = \theta x_i\,, \\
\mathrm{Var}(m_i) &= \mathrm{E}(\mathrm{Var}(m_i \mid l_i)) + \mathrm{Var}(\mathrm{E}(m_i \mid l_i)) \\
&= \mathrm{E}(\nu l_i) + \mathrm{Var}(\nu l_i) = \theta x_i + \nu^2\mathrm{Var}(l_i) = \theta x_i + \theta^2\sum_k \frac{c_{ik}}{k^2(k-1)^2} \\
&= \theta x_i + \theta^2 z_{ii}\,, \\
\mathrm{Cov}(m_i, m_j) &= \mathrm{E}(m_im_j) - \mathrm{E}(m_i)\mathrm{E}(m_j) \\
&= \mathrm{E}(\mathrm{E}(m_im_j \mid l_i, l_j)) - \mathrm{E}(m_i)\mathrm{E}(m_j) \\
&= \nu^2\mathrm{E}(l_il_j) - \nu^2\mathrm{E}(l_i)\mathrm{E}(l_j) \\
&= \nu^2\mathrm{Cov}(l_i, l_j) = \nu^2\mathrm{Cov}\left(\sum_k c_{ik}T_k, \sum_k c_{jk}T_k\right) \\
&= \nu^2\sum_k c_{ik}c_{jk}\mathrm{Cov}(T_k, T_k) = \nu^2\sum_k c_{ik}c_{jk}\mathrm{Var}(T_k) \\
&= \nu^2\sum_k c_{ik}c_{jk}[\mathrm{E}(T_k^2) - [\mathrm{E}(T_k)]^2] = \nu^2\sum_k c_{ik}c_{jk}[\mathrm{E}(T_k)]^2 \quad \text{de (2.44)} \\
&= \nu^2\sum_k c_{ik}c_{jk}\frac{(4N)^2}{k^2(k-1)^2} = \theta^2 z_{ij}\,.
\end{aligned}
$$

Aplica-se agora a teoria de modelos lineares para a obtenção de um estimador

linear de θ com variância mínima dentre os não viesados.

Seja o vetor aleatório $\mathbf{m} = (m_1, \ldots, m_{2(n-1)})'$ em que m_i é o número de mutações que ocorreu no ramo i. Foi visto que:

$$\mathrm{E}(m_i) = x_i\theta\,,$$

$$\mathrm{Cov}(m_i, m_j) = \sigma_{ij}\,\theta\,,$$

em que

$$\sigma_{ij} = \begin{cases} x_i + z_{ij}\,\theta & \text{se} \quad i = j \\ z_{ij}\,\theta & \text{caso contrário} \end{cases} \tag{4.8}$$

e x_i e z_{ij} são constantes definidas em (4.6) e (4.7), respectivamente.

Procura-se uma função linear do vetor aleatório $\mathbf{m}$ que seja um estimador não viesado de θ e que sua variância seja mínima entre todos os estimadores de θ que são funções lineares do vetor aleatório $\mathbf{m}$, ou seja, o estimador linear de θ com mínima variância dentre os não viesados. A seguir apresentam-se alguns resultados da teoria de modelos lineares que serão utilizados.

Modelo Linear Clássico

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0} \\ \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I} \\ \mathbf{X} \quad \text{com posto completo p} \end{cases} \tag{4.9}$$

Seja $\mathcal{B}$ o conjunto de todos os possíveis vetores $\boldsymbol{\beta}$. $\mathcal{B} = \mathbb{R}^p$.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in B} Q(\boldsymbol{\beta})\,.$$

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,. \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,, \\ \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} &= 2\mathbf{X}'\mathbf{X} \quad \text{(não negativa definida)}\,. \end{aligned}$$

Assim, o sistema de equações normais pode ser descrito por:

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

E o BLUE (Best Linear Unbiased Estimator) de $\boldsymbol{\beta}$ é:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,,$$

isto é, dentre todos os estimadores lineares de $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$ é o de mínima variância.

Modelo Linear Generalizado

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0} \\ \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{W} \\ \mathbf{X} \quad \text{com posto completo p} \\ \mathbf{W} \quad \text{positiva definida} \end{cases} \tag{4.10}$$

O BLUE neste caso é o estimador de Aitken:

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}\,,$$

ou seja, $\widetilde{\boldsymbol{\beta}}$ é o estimador de mínimos quadrados ponderados de $\boldsymbol{\beta}$.

Teorema 3. ***(Teorema de Gauss-Markov-Aitken):*** *Se* $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ *em que* $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2\mathbf{W})$*, o estimador generalizado de mínimos quadrados:*

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\mathbf{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{-1}\boldsymbol{y}$$

é não viesado e é o melhor estimador linear não viesado (BLUE) para $\boldsymbol{\beta}$, ou seja, dentre todos os estimadores lineares não viesados para $\boldsymbol{\beta}$, $\widetilde{\boldsymbol{\beta}}$ é aquele com menor variância. Sua matriz de covariância é dada por:

$$Var(\widetilde{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\mathbf{W}^{-1}\boldsymbol{X})^{-1}\,.$$

Demonstração. Desde que $\mathbf{W}$ e $\mathbf{W}^{-1}$ são simétricas e positivas definidas, existem matrizes $\mathbf{M}$ e $\mathbf{N}$ tais que: $\mathbf{W} = \mathbf{MM}$ e $\mathbf{W}^{-1} = \mathbf{NN}$, em que $\mathbf{M} = \mathbf{W}^{1/2}$ e $\mathbf{N} = \mathbf{W}^{-1/2}$ são regulares e simétricas.

Multiplicando (4.10) por $\mathbf{N}$, tem-se:

$$\underbrace{\mathbf{Ny}}_{\widetilde{\mathbf{y}}} = \underbrace{\mathbf{NX}}_{\widetilde{\mathbf{X}}}\boldsymbol{\beta} + \underbrace{\mathbf{N}\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{\epsilon}}}\,.$$

Portanto,

$$\mathrm{E}(\widetilde{\boldsymbol{\epsilon}}) = \mathrm{E}(\mathbf{N}\boldsymbol{\epsilon}) = \mathbf{0}$$

e

$$\mathrm{Var}(\widetilde{\boldsymbol{\epsilon}}) = \mathrm{Var}(\mathbf{N}\boldsymbol{\epsilon})\mathbf{N}\mathrm{Var}(\boldsymbol{\epsilon})\mathbf{N}' = \sigma^2\mathbf{NWN} = \sigma^2\mathbf{NMMN} = \sigma^2\mathbf{I}\,.$$

Assim, após a transformação linear, o modelo $\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\epsilon}}$ satisfaz as suposições do modelo linear clássico. O BLUE de $\boldsymbol{\beta}$ é, portanto:

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{y}} \\ &= (\mathbf{X}'\mathbf{NN}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{NN}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}\,.\end{aligned}$$

□

Considerando agora o caso de interesse, tem-se o vetor $\mathbf{m}$, cujos elementos representam o número de mutações que ocorreu em cada ramo da árvore. Quer-se obter um estimador linear não viesado para θ com variância mínima dentre todos os estimadores lineares não viesados que utilizam a informação sobre o número de mutações que ocorreu em cada ramo.

Aplicando a teoria de modelos lineares generalizados ao caso de interesse, tem-se, de (4.5) o seguinte modelo:

$$\begin{cases} \mathbf{m} = \mathbf{x}\theta + \boldsymbol{\epsilon} \\ \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0} \\ \mathrm{Var}(\boldsymbol{\epsilon}) = \mathrm{Var}(\mathbf{m}) = \theta\mathbf{D}_x + \theta^2\mathbf{Z} = \theta\mathbf{V}_\theta \\ \mathbf{x} \quad \text{com posto completo } 2(n-1) \\ \mathbf{V}_\theta \quad \text{positiva definida} \end{cases} \tag{4.11}$$

em que $\mathbf{D}_x = \mathrm{diag}(\mathbf{x})$ e $\mathbf{Z} = \{z_{ij}\}$, i, $j = 1, 2, \ldots, 2(n-1)$.

O BLUE de θ, pelo Teorema 3, é dado por:

$$\widehat{\theta}_m = \mathbf{u}'\mathbf{m}\,, \tag{4.12}$$

em que $\mathbf{u}' = (u_1, \ldots, u_{2(n-1)})$ é

$$\mathbf{u}' = (\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}_\theta^{-1}. \tag{4.13}$$

A esperança do estimador é:

$$\mathrm{E}(\widehat{\theta}_m) = \mathrm{E}(\mathbf{u}'\mathbf{m}) = (\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1}(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})\theta = \theta\,,$$

e sua variância é dada por:

$$\mathrm{Var}(\widehat{\theta}_m) = \theta(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1}\,. \tag{4.14}$$

Demonstração.

$$\begin{aligned} \mathrm{Var}(\widehat{\theta}_m) &= \mathrm{Var}(\mathbf{u}'\mathbf{m}) = \mathbf{u}'\mathrm{Var}(\mathbf{m})\mathbf{u} \\ &= \mathbf{u}'\theta\mathbf{V}_\theta\mathbf{u} = \theta(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{V}_\theta\mathbf{V}_\theta^{-1}\mathbf{x}(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1} \\ &= \theta(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1}\,. \end{aligned}$$

□

No entanto, a equação (4.12) não pode ser usada para estimar θ diretamente, pois exige o valor do próprio parâmetro desconhecido θ. Por esta razão, é preciso utilizar um procedimento iterativo. Suponha que tem-se uma estimativa inicial de θ, denotada por $\theta_{(0)}$. Então as equações (4.12) e (4.13) sugerem que pode-se obter uma série de $\mathbf{u}_{(k)}$ e uma série de $\theta_{(k)}$ através de:

$$\mathbf{u}'_{(k)} = (\mathbf{x}'\mathbf{V}^{-1}_{\theta_{k-1}}\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}^{-1}_{\theta_{k-1}}\,,$$

$$\theta_{(k)} = \mathbf{u}'_{(k)}\mathbf{m}\,.$$

Se a série $\mathbf{u}_{(k)}(k = 1, 2, ...)$ converge, seu valor limite $\mathbf{u}_{(\infty)}$ pode ser usado como uma estimativa de $\mathbf{u}$ e $\widetilde{\theta}_m = \theta_{(\infty)}$ pode ser usado como estimativa de θ. Refere-se a $\widetilde{\theta}_m$ como o melhor estimador linear não viesado (BLUE) de θ. Quando a estimativa de Watterson (1975), $\mathcal{T}_1$, é usada como valor inicial $\theta_{(0)}$, o processo iterativo não precisa, geralmente, de mais de quatro ciclos.

Deve-se notar que embora pretenda-se estimar θ, que é uma função linear de $\mathbf{m}$, estritamente falando $\widetilde{\theta}_m$ não é uma função linear destas variáveis aleatórias, pois $\mathbf{u}_{(\infty)}$ depende de $\theta_{(\infty)}$, que é função de $\mathbf{m}$. Isto dificulta a obtenção de uma variância amostral exata de $\widetilde{\theta}_m$, além de implicar que $\widetilde{\theta}_m$ pode apresentar viés. Contudo, resultados numéricos apresentados mais adiante mostram que tratar $\mathbf{u}$ como um vetor de constantes é apropriado para estudar as propriedades amostrais de $\widetilde{\theta}_m$. Assim, tem-se:

$$\begin{aligned} \mathrm{Var}(\widetilde{\theta}_m) &= \mathbf{u}'(\mathbf{D}_x\theta + \theta^2\mathbf{Z})\mathbf{u} \\ &= a\theta + b\theta^2, \end{aligned}$$

em que $a = \mathbf{u}'\mathbf{D}_x\mathbf{u}$ e $b = \mathbf{u}'\mathbf{Z}\mathbf{u}$ e

$$\begin{aligned} \mathrm{E}(\widetilde{\theta}^2_m) &= \mathrm{Var}(\widetilde{\theta}_m) + [\mathrm{E}(\widetilde{\theta}_m)]^2 \\ &= a\theta + b\theta^2 + \theta^2 \\ &= a\theta + \theta^2(1 + b)\,. \end{aligned}$$

Como uma estimativa não viesada de θ^2 (pelo método dos momentos) é:

$$\frac{\widetilde{\theta}_m(\widetilde{\theta}_m - a)}{1+b}\,, \tag{4.15}$$

uma estimativa aproximadamente não viesada da variância de $\widetilde{\theta}_m$ é:

$$V_c = \widehat{\mathrm{Var}(\widetilde{\theta}_m)} = a\widetilde{\theta}_m + \frac{\widetilde{\theta}_m(\widetilde{\theta}_m - a)b}{1+b}\,. \tag{4.16}$$

Outra maneira é estimar θ^2 diretamente de $(\widetilde{\theta}_m)^2$ e substituir em 4.15. Assim, obtém-se a seguinte estimativa da variância de $\widehat{\theta}_m$:

$$V_{nc} = \widehat{\mathrm{Var}(\widetilde{\theta}_m)} = \widetilde{\theta}_m(\mathbf{x}'\mathbf{V}^{-1}_{\widetilde{\theta}_m}\mathbf{x})^{-1}\,.$$

4.2.2 Outros Modelos Lineares

Uma genealogia de n genes consiste de $2(n-1)$ ramos. Um ramo é dito ser de *tamanho* i se exatamente i seqüências na amostra descendem deste ramo. Portanto, mutações em uma amostra podem ser classificadas em $n-1$ tamanhos. No exemplo apresentado na Figura 4.1, os ramos 1 e 2 são de tamanho 2 e os demais de tamanho 1.

Seja r_i *a soma das mutações que ocorreram nos ramos de tamanho* i e $\mathbf{r}$ o vetor:

$$\mathbf{r} = (r_1, \ldots, r_{n-1})'\,.$$

O vetor $\mathbf{r}$ é uma fonte primária de informação utilizada para estimar θ. Na Seção 2.4.1.2 apresentam-se os momentos de r_1 (R_e na notação daquela Seção).

No caso da topologia de 3 genes, temos que $r_2 = R_i$, pois $R_e = r_1$ e $R_i = r_2 + \ldots + r_{n-1}$ no caso geral. Assim, de (4.17) a (2.86), tem-se que:

$$\mathrm{E}(r_1) = \theta \qquad \mathrm{E}(r_2) = \frac{1}{2}\,\theta$$

$$\mathrm{Var}(r_1) = \theta + \frac{1}{2}\,\theta^2 \qquad \mathrm{Var}(r_2) = \frac{1}{2}\,\theta + \frac{1}{4}\,\theta^2$$

$$\mathrm{Cov}(r_1, r_2) = \frac{1}{4}\,\theta^2\,.$$

Ao considerarmos uma amostra de n seqüências, o intervalo de tempo entre o momento no qual a amostra é tomada e o ancestral comum mais recente pode ser dividido em um número de períodos de acordo com os eventos ocorridos. Considera-se evento, nesta Seção, uma coalescência (um nó). Por conveniência, trata-se o momento no qual a amostra é tomada como um evento. O tempo (em número de gerações) de coalescência entre o $(i-1)$-ésimo e o i-ésimo nó é, como já visto, T_i. Sob o modelo de neutralidade de Wright-Fisher, só existem eventos de coalescência. Assim, T_i representa o tempo no período em que a amostra tem i ancestrais distintos.

Suponha que existam no total M sítios em cada seqüência. Cada sítio pode ser considerado como um locus e existe uma genealogia para cada sítio que liga os n nucleotídeos de um sítio ao seu ancestral comum mais recente. Considere a genealogia do l-ésimo sítio. O comprimento dos ramos de tamanho k para o l-ésimo sítio, $L_k^{(l)}$, é dado por:

$$L_k^{(l)} = \sum_{i=2}^{n} s_{ki}^{(l)} T_i\,,$$

em que s_{ki} é o número de vezes em que T_i aparece nos ramos de tamanho k.

O comprimento médio dos ramos de tamanho k de todos os sítios é, portanto:

$$L_k = \frac{1}{M}\sum_{l=1}^{M} L_k^{(l)} = \sum_{i=2}^{n} s_{ki} T_i\,,$$

em que

$$s_{ki} = \frac{1}{M}\sum_{l=1}^{M} s_{ki}^{(l)}\,.$$

Quando não há recombinação, todas as M genealogias são idênticas e, conseqüentemente, $s_{ki} = s_{ki}^{(l)} = \ldots = s_{ki}^{(M)}$.

Desde que ν é definido como a taxa de mutação por seqüência por geração, a taxa de mutação por sítio por geração é, então, ν/M. Assuma que o número de mutações em um ramo da genealogia de um sítio é uma variável aleatória com

Tabela 4.2: s_{ki} para a topologia da Figura 4.1.

	i		
k	2	3	4
1	0	2	4
2	2	1	0
3	0	0	0

distribuição Poisson($L\nu/M$) em que L é o comprimento do ramo. Então, o número de mutações nos ramos de tamanho i em todas as M genealogias de um sítio é uma variável aleatória com distribuição Poisson($L\nu$). Suponha que no total existam T genealogias para uma amostra de n seqüências de DNA e seja $r_i^{(t)}$ a soma das mutações que ocorreram nos ramos de tamanho i para a genealogia t. Tem-se:

$$P(r_i^{(t)} = k \mid l_i) = \frac{e^{l_i\nu}(l_i\nu)^k}{k!}, \quad k = 0, 1, \ldots$$

De maneira equivalente à (4.5), tem-se que:

$$\begin{aligned} \mathrm{E}(r_i^{(t)}) &= \theta x_i^{(t)} \\ \mathrm{Var}(r_i^{(t)}) &= x_i\theta + z_{ii}^{(t)}\theta^2 \\ \mathrm{Cov}(r_i^{(t)}, r_j^{(t)}) &= z_{ij}^{(t)}\theta^2 \\ \mathrm{E}(r_i^{(t)} r_j^{(t)}) &= \theta^2(z_{ij}^{(t)} + x_i^{(t)} x_j^{(t)}) \end{aligned} \tag{4.17}$$

em que

$$x_i^{(t)} = \sum_k \frac{s_{ik}}{k(k-1)}\,,$$

$$z_{ij}^{(t)} = \sum_k \frac{s_{ik}s_{jk}}{k^2(k-1)^2}\,.$$

Definem-se:

$$\alpha_i = \sum_t^T x_i^{(t)} p_t\,, \tag{4.18}$$

$$\sigma_{ij} = \sum_t^T (z_{ij}^{(t)} + x_i^{(t)} x_j^{(t)}) p_t - \alpha_i\alpha_j\,. \tag{4.19}$$

A probabilidade de observar-se a genealogia t, p_t, está apresentada em (2.28).

Aplicando a teoria de modelos lineares nos resultados de (4.17), (4.18) e (4.19), tem-se:

$$\begin{cases} \mathbf{r} = \theta\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \mathrm{E}(\mathbf{r}) = \theta\boldsymbol{\alpha} \\ \mathrm{Var}(\boldsymbol{\epsilon}) = \mathrm{Var}(\mathbf{r}) = \theta\mathbf{D}_\alpha + \theta^2\boldsymbol{\Sigma} \\ \boldsymbol{\alpha} \quad \text{com posto completo } n-1 \\ \mathrm{Var}(\mathbf{r}) \quad \text{positiva definida} \end{cases} \tag{4.20}$$

em que $\mathbf{D}_\alpha = \mathrm{diag}(\boldsymbol{\alpha})$ e $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, i, $j = 1, 2, \ldots, n-1$.

O BLUE de θ, pelo Teorema 3, é:

$$\widehat{\theta}_r = \mathbf{u}'\mathbf{r}\,, \tag{4.21}$$

em que $\mathbf{u}' = (u_1, \ldots, u_{n-1})$ é

$$\mathbf{u}' = (\boldsymbol{\alpha}'(\mathbf{D}_\alpha + \theta\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}'(\mathbf{D}_\alpha + \theta\boldsymbol{\Sigma})^{-1}. \tag{4.22}$$

A variância é dada por:

$$\mathrm{Var}(\widehat{\theta}_r) = \theta(\boldsymbol{\alpha}'(\mathbf{D}_\alpha + \theta\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha})^{-1}\,. \tag{4.23}$$

No entanto, (4.21) não fornece diretamente uma estimativa de θ, pois depende do próprio parâmetro desconhecido. O problema é resolvido através de um procedimento iterativo. Pode-se obter uma série de $\mathbf{u}_{(k)}$ e uma série de $\theta_{(k)}$ através de:

$$\mathbf{u}'_{(k)} = (\boldsymbol{\alpha}'(\mathbf{D}_\alpha + \theta_{(k-1)}\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}'(\mathbf{D}_\alpha + \theta_{(k-1)}\boldsymbol{\Sigma})^{-1}\,,$$

$$\theta_{(k)} = \mathbf{u}'_{(k)}\mathbf{r}\,.$$

O limite da seqüência, $\theta_{(\infty)}$, é tomado como o BLUE de θ a partir de $\mathbf{r}$, denotado por $\widetilde{\theta}_r$.

Tabela 4.3: s_{ik} para a topologia da Figura 2.2a.

	k		
i	2	3	4
1	1	2	4
2	0	1	0
3	1	0	0

Tabela 4.4: s_{ik} para a topologia da Figura 2.2b.

	k		
i	2	3	4
1	0	2	4
2	2	1	0
3	0	0	0

De maneira análoga à Seção anterior, tem-se:

$$V_c = \widehat{\mathrm{Var}(\widetilde{\theta}_r)} = a\widetilde{\theta}_r + \frac{\widetilde{\theta}_r(\widetilde{\theta}_r - a)b}{1+b}\,, \tag{4.24}$$

em que $a = \mathbf{u}'\mathbf{D}_\alpha\mathbf{u}$ e $b = \mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}$.

Deve-se notar, porém, que obter $\boldsymbol{\alpha}$ e $\boldsymbol{\Sigma}$ não é simples, pois T é um número grande mesmo para pequenas amostras e p_k nem sempre é facilmente calculado. Soluções analíticas para o modelo de neutralidade de Wright-Fisher estão em Fu (1995). Quando estas soluções analíticas não estão disponíveis, pode-se estimar $\boldsymbol{\alpha}$ e $\boldsymbol{\Sigma}$ da seguinte maneira. Suponha que se tem um algoritmo para gerar genealogias de amostras sob um certo modelo e seja G o número total de genealogias geradas aleatoriamente. Assim, de acordo com a teoria de integração de Monte-Carlo, pode-se estimar α_i e σ_{ij} por:

$$\hat{\alpha}_i = \frac{1}{G}\sum_{k}^{G} x_i^{(k)} \quad \hat{\sigma}_{ij} = \frac{1}{G}\sum_{k}^{G}(z_{ij}^{(k)} + x_i^{(k)}x_j^{(k)}) - \hat{\alpha}_i\hat{\alpha}_j\,,$$

respectivamente.

Na Figura 2.2, apresentam-se as relações evolucionárias esperadas entre 4 genes amostrados. As Tabelas 4.3 e 4.4 apresentam os valores de s_{ik} para a topologia a e

b, respectivamente. A probabilidade de ocorrência de cada topologia é obtida por (2.28).

Calculam-se:

$$\begin{aligned}
x_1^{(a)} &= \frac{s_{12}}{2} + \frac{s_{13}}{6} + \frac{s_{14}}{12} = \frac{7}{6} \\
x_2^{(a)} &= \frac{s_{22}}{2} + \frac{s_{23}}{6} + \frac{s_{24}}{12} = \frac{1}{6} \\
x_3^{(a)} &= \frac{s_{32}}{2} + \frac{s_{33}}{6} + \frac{s_{34}}{12} = \frac{1}{2} \\
x_1^{(b)} &= \frac{s_{12}}{2} + \frac{s_{13}}{6} + \frac{s_{14}}{12} = \frac{2}{3} \\
x_2^{(b)} &= \frac{s_{22}}{2} + \frac{s_{23}}{6} + \frac{s_{24}}{12} = \frac{7}{6} \\
x_3^{(b)} &= \frac{s_{32}}{2} + \frac{s_{33}}{6} + \frac{s_{34}}{12} = 0
\end{aligned}$$

$$\begin{aligned}
\alpha_1 &= x_1^{(a)} p_a + x_1^{(b)} p_b = \frac{7}{6} \cdot \frac{2}{3} + \frac{2}{3} \cdot \frac{1}{3} = 1 \\
\alpha_2 &= x_2^{(a)} p_a + x_2^{(b)} p_b = \frac{1}{2} \\
\alpha_3 &= x_3^{(a)} p_a + x_3^{(b)} p_b = \frac{1}{3}
\end{aligned}$$

$$\begin{aligned}
z_{11}^{(a)} &= \frac{17}{36}, \quad z_{12}^{(a)} = \frac{1}{18} \\
z_{13}^{(a)} &= \frac{1}{4}, \quad z_{22}^{(a)} = \frac{1}{36} \\
z_{23}^{(a)} &= 0, \quad z_{33}^{(a)} = \frac{1}{4}
\end{aligned}$$

$$\begin{aligned}
z_{11}^{(b)} &= \frac{2}{9}, \quad z_{12}^{(b)} = \frac{1}{18} \\
z_{13}^{(b)} &= 0, \quad z_{22}^{(b)} = \frac{37}{36} \\
z_{23}^{(b)} &= 0, \quad z_{33}^{(b)} = 0
\end{aligned}$$

$$\begin{aligned}
\sigma_{11} &= \left(z_{11}^{(a)} + x_1^{(a)}x_1^{(a)}\right)p_a + \left(z_{11}^{(b)} + x_1^{(b)}x_1^{(b)}\right)p_b - \alpha_1^2 = \frac{4}{9} \\
\sigma_{12} &= \left(z_{12}^{(a)} + x_1^{(a)}x_2^{(a)}\right)p_a + \left(z_{12}^{(b)} + x_1^{(b)}x_2^{(b)}\right)p_b - \alpha_1\alpha_2 = -\frac{1}{18} \\
\sigma_{13} &= \left(z_{13}^{(a)} + x_1^{(a)}x_3^{(a)}\right)p_a + \left(z_{13}^{(b)} + x_1^{(b)}x_3^{(b)}\right)p_b - \alpha_1\alpha_3 = \frac{2}{9} \\
\sigma_{22} &= \left(z_{22}^{(a)} + x_2^{(a)}x_2^{(a)}\right)p_a + \left(z_{22}^{(b)} + x_2^{(b)}x_2^{(b)}\right)p_b - \alpha_2^2 = \frac{7}{12} \\
\sigma_{23} &= \left(z_{23}^{(a)} + x_2^{(a)}x_3^{(a)}\right)p_a + \left(z_{23}^{(b)} + x_2^{(b)}x_3^{(b)}\right)p_b - \alpha_2\alpha_3 = -\frac{1}{9} \\
\sigma_{33} &= \left(z_{33}^{(a)} + x_3^{(a)}x_3^{(a)}\right)p_a + \left(z_{33}^{(b)} + x_3^{(b)}x_3^{(b)}\right)p_b - \alpha_3^2 = \frac{2}{9}
\end{aligned}$$

Fu (1995) mostrou que, para o modelo de Wright-Fisher:

$$\begin{aligned}
\mathrm{E}(r_i) &= \frac{1}{i}\,\theta\,, \\
\mathrm{Var}(r_i) &= \frac{1}{i}\,\theta + \sigma_{ii}\theta^2\,, \\
\mathrm{Cov}(r_i, r_j) &= \sigma_{ij}\,\theta^2\,,
\end{aligned} \tag{4.25}$$

em que

$$\sigma_{ii} = \begin{cases} \beta_n(i+1) & \text{se } i < n/2, \\ 2\frac{a_n - a_i}{n-i} - \frac{1}{i^2} & \text{se } i = n/2\,, \\ \beta_n(i) - \frac{1}{i^2} & \text{se } i > n/2\,. \end{cases}$$

e $\sigma_{ij}(i > j)$ é dado por:

$$\sigma_{ij} = \begin{cases} \frac{\beta_n(i+1) - \beta_n(i)}{2} & \text{se } i + j < n, \\ \frac{a_n - a_i}{n-i} + \frac{a_n - a_j}{n-j} - \frac{\beta_n(i) + \beta_n(j+1)}{2} - \frac{1}{ij} & \text{se } i + j = n\,, \\ \frac{\beta_n(j) - \beta_n(j+1)}{2} - \frac{1}{ij} & \text{se } i + j > n\,. \end{cases}$$

em que $a_n = \sum_{j=1}^{n-1} 1/j$ e $\beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}$.

Outro vetor de informação utilizado na estimação de θ é:

$$\mathbf{v} = (v_1, \ldots, v_{[n/2]})' ,$$

cujos elementos são dados por:

$$v_i = \frac{r_i + r_{n-i}}{1 + \delta_{i,n-i}} ,$$

em que $\delta_{i,j}$ é igual a 1 se $i = j$ e 0 caso contrário e $[x]$ é o menor inteiro maior ou igual a x. Sob o modelo de sítios infinitos, define-se v_i como o número de sítios segregantes nos quais as freqüências de dois nucleotídeos segregantes são i e $n - i$, respectivamente. Tal sítio segregante é chamado de *sítio do tipo i* ou *sítio segregante i*. Assim, v_1 é o número de *singletones*. De (4.25), tem-se:

$$\begin{aligned} \mathrm{E}(v_i) &= \phi_i \theta , \\ \mathrm{Var}(v_i) &= \phi_i \theta + \lambda_{ii} \theta^2 , \\ \mathrm{Cov}(v_i, v_j) &= \lambda_{ij} \theta^2 , \end{aligned} \tag{4.26}$$

em que

$$\phi_i = \frac{1}{1 + \delta_{i,n-i}} \left(\frac{1}{i} + \frac{1}{n-i} \right) , \tag{4.27}$$

$$\lambda_{ij} = \frac{\sigma_{ij} + \sigma_{i,n-j} + \sigma_{n-i,j} + \sigma_{n-i,n-j}}{(1 + \delta_{i,n-i})(1 + \delta_{j,n-j})} . \tag{4.28}$$

Tem-se, assim, o seguinte modelo linear:

$$\begin{cases} \mathbf{v} = \boldsymbol{\phi}\, \theta + \boldsymbol{\epsilon} \\ \mathrm{E}(\mathbf{v}) = \boldsymbol{\phi}\, \theta \\ \mathrm{Var}(\boldsymbol{\epsilon}) = \mathrm{Var}(\mathbf{v}) = \theta \mathbf{D}_\phi + \theta^2 \boldsymbol{\Lambda} \\ \boldsymbol{\phi} \quad \text{com posto completo } [n/2] \\ \mathrm{Var}(\mathbf{v}) \quad \text{positiva definida} \end{cases} \tag{4.29}$$

em que $\mathbf{D}_\phi = \text{diag}(\boldsymbol{\phi})$ e $\boldsymbol{\Lambda} = \{\lambda_{ij}\}$, i, $j = 1, 2, \ldots, [n/2]$.

O BLUE de θ é dado por:

$$\widehat{\theta}_v = \mathbf{u}'\mathbf{v}\,, \tag{4.30}$$

em que

$$\mathbf{u}' = (\boldsymbol{\phi}'(\mathbf{D}_\phi + \theta\boldsymbol{\Lambda})^{-1}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'(\mathbf{D}_\phi + \theta\boldsymbol{\Lambda})^{-1}$$

e sua variância é:

$$\text{Var}(\widehat{\theta}_v) = \theta(\boldsymbol{\phi}'(\mathbf{D}_\phi + \theta\boldsymbol{\Lambda})^{-1}\boldsymbol{\phi})^{-1}\,. \tag{4.31}$$

Para este modelo também é necessário o procedimento iterativo e o estimador obtido é denotado por $\widetilde{\theta}_v$. A estimativa de sua variância é dada por:

$$V_c = \widehat{\text{Var}(\widetilde{\theta}_v)} = a\widetilde{\theta}_v + \frac{\widetilde{\theta}_v(\widetilde{\theta}_r - a)b}{1+b}\,, \tag{4.32}$$

em que $a = \mathbf{u}'\mathbf{D}_\phi\mathbf{u}$ e $b = \mathbf{u}'\boldsymbol{\Lambda}\mathbf{u}$.

Comparando as variâncias dos três modelos, definidas em (4.14), (4.23) e (4.31), assim como suas respectivas variâncias estimadas calculadas em (4.16), (4.24) e (4.32), observa-se que para os dois últimos modelos esses valores são constantes uma vez fixados n e θ. No caso do modelo apresentado em (4.11), esses valores dependem da topologia da árvore, uma vez que x_i e z_{ij} são calculados a partir dos c_{ik}'s.

Os estimadores apresentados no Capítulo 2 também são funções lineares dos vetores $\mathbf{m}$, $\mathbf{r}$ e $\mathbf{v}$, de fato:

$$\mathcal{T}_1 = \frac{1}{a_n}\sum_{i=1}^{n-1} r_i = \frac{1}{a_n}\sum_{i=1}^{[n/2]} v_i = \frac{1}{a_n}\sum_{i=1}^{2(n-1)} m_i\,,$$

$$\mathcal{T}_2 = \sum_{i=1}^{n-1}\frac{2i(n-i)}{n(n-1)}r_i = \sum_{i=1}^{[n/2]}\frac{2i(n-i)}{n(n-1)}v_i = \sum_{i=1}^{2(n-1)}\frac{2a_i(n-a_i)m_i}{n(n-1)}\,,$$

$$\mathcal{T}_3 = \frac{n-1}{n} v_1 \,,$$

em que a_i é o tamanho do ramo i. No entanto, seus coeficientes são constantes pré-determinadas uma vez que n é escolhido e, em geral, estimadores com constantes pré-determinadas não são os melhores dentre os lineares. É interessante observar que quando θ é próximo de zero, tem-se que:

$$\begin{aligned} \widetilde{\theta}_m &\approx (\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{m} = \mathcal{T}_1 \,, \\ \widetilde{\theta}_r &\approx (\boldsymbol{\alpha}'\mathbf{D}_\alpha^{-1}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}'\mathbf{D}_\alpha^{-1}\mathbf{r} = \mathcal{T}_1 \,, \\ \widetilde{\theta}_v &\approx (\boldsymbol{\phi}'\mathbf{D}_\phi^{-1}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}'\mathbf{D}_\phi^{-1}\mathbf{v} = \mathcal{T}_1 \,, \end{aligned} \tag{4.33}$$

ou seja, para pequenos valores de θ, o estimador de Watterson é aproximadamente o melhor estimador linear de θ.

Reescrever $\mathcal{T}_1$ em termos de $\mathbf{m}$ facilita na comparação da variância deste estimador com a variância de $\widetilde{\theta}_m$.

$$\mathrm{Var}(\mathcal{T}_1) = \mathrm{Var}[(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{m}] = \theta(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{V}_\theta\mathbf{D}_x^{-1}\mathbf{x}(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1} \,.$$

$$\mathrm{Var}(\widetilde{\theta}_m) = \theta(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1} \,.$$

$$\begin{aligned} \mathbf{V}_\theta &= \mathbf{D}_x + \theta\mathbf{Z} = \mathbf{D}_x + \mathbf{z}\mathbf{z}' \,. \\ \mathbf{V}_\theta^{-1} &= \mathbf{D}_x^{-1} - \frac{\mathbf{D}_x^{-1}\mathbf{z}\mathbf{z}'\mathbf{D}_x^{-1}}{1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z}} \,. \\ \mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x} &= \mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x} - \frac{\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{z}\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{x}}{1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z}} \\ &= \mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x} - \mathbf{b}\mathbf{b}' \,, \end{aligned}$$

em que $\mathbf{b} = \frac{\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{z}}{\sqrt{1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z}}}$. Note que $\mathbf{b}\mathbf{b}' = \frac{\theta\sum_i\sum_j z_{ij}}{1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z}}$.

$$(\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1} = (\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1} - \underbrace{\frac{(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{b}\mathbf{b}'(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}}{1+\mathbf{b}'(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{b}}}_{\mathbf{a}}.$$

$$\begin{aligned}
\mathbf{a} &= \frac{\theta\sum_i\sum_j z_{ij}}{1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z}}\frac{1}{a_n^2}\left[1+\frac{\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{x}\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{z}}{a_n(1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z})}\right]^{-1} \\
&= \frac{\theta\sum_i\sum_j z_{ij}}{1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z}}\frac{1}{a_n^2}\left[\frac{a_n(1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z})+\mathbf{z}'\begin{pmatrix}1 & \dots & 1\\ \vdots & \dots & \vdots\\ 1 & \dots & 1\end{pmatrix}\mathbf{z}}{a_n(1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z})}\right]^{-1} \\
&= \frac{\theta\sum_i\sum_j z_{ij}}{a_n(1+\mathbf{z}'\mathbf{D}_x^{-1}\mathbf{z})+\mathbf{z}'\begin{pmatrix}1 & \dots & 1\\ \vdots & \dots & \vdots\\ 1 & \dots & 1\end{pmatrix}\mathbf{z}}.
\end{aligned}$$

Se $\mathrm{Var}(\mathcal{T}_1) - \mathrm{Var}(\widetilde{\theta}_m) \geq 0$, tem-se que $\widetilde{\theta}_m$ é mais preciso do que $\mathcal{T}_1$. Portanto, deve-se verificar se a diferença a seguir é não negativa:

$$
\begin{aligned}
&(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{V}_\theta\mathbf{D}_x^{-1}\mathbf{x}(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1} - (\mathbf{x}'\mathbf{V}_\theta^{-1}\mathbf{x})^{-1} = \\
&= (\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{V}_\theta\mathbf{D}_x^{-1}\mathbf{x}(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1} - [(\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1} - \mathbf{a}] \\
&= (\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\left[\frac{\begin{pmatrix} 1 & \dots & 1 \end{pmatrix}\mathbf{V}_\theta\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}{a_n} - 1\right] + \mathbf{a} \\
&= (\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\left[\frac{a_n + \frac{\theta\sum_i\sum_j\sum_k c_{ik}c_{jk}}{k^2(k-1)^2}}{a_n} - 1\right] + \mathbf{a} \\
&= (\mathbf{x}'\mathbf{D}_x^{-1}\mathbf{x})^{-1}\frac{\theta\sum_i\sum_j\sum_k c_{ik}c_{jk}}{a_n k^2(k-1)^2} + \mathbf{a}\,,
\end{aligned}
$$

que é uma quantidade não negativa. Assim, tem-se que:

$$\mathrm{Var}(\widetilde{\theta}_m) \leq \mathrm{Var}(\mathcal{T}_1)\,.$$

Além disso, $\widetilde{\theta}_m$ é um estimador consistente. De fato:

$$P(\mid \widetilde{\theta}_m - \theta \mid > \varepsilon) \leq \frac{\mathrm{Var}(\widetilde{\theta}_m)}{\varepsilon^2} \leq \frac{\mathrm{Var}(\mathcal{T}_1)}{\varepsilon^2} \to 0, \quad \text{quando n} \to \infty\,.$$

No Capítulo 5, apresentam-se as distribuições empíricas dos três estimadores estudados neste capítulo.

5 *Simulação Computacional e Aplicação*

5.1 Introdução

A eficiência dos estimadores apresentados nos Capítulos 2 e 4 pode ser medida através da comparação de suas variâncias com o limite inferior da variância de todos os estimadores não viesados possíveis de θ, obtido em (4.3):

$$V_{min} = \frac{\theta}{\sum_{k=1}^{n-1} \frac{1}{\theta+k}} .$$

Utilizam-se amostras simuladas para avaliar o desempenho dos procedimentos de estimação. Simula-se um número de genealogias de acordo com os valores de θ e n escolhidos e a teoria de coalescência (TAJIMA, 1983).

5.2 Simulação de árvores genealógicas

Tajima (1989) propôs um método para simular árvore genealógicas sob o modelo de neutralidade seletiva que consiste no seguinte.

Primeiro, gera-se a árvore genealógica das seqüências de DNA. Quando tem-se n seqüências de DNA, escolhe-se aleatoriamente duas destas seqüências para coalescerem. Assim, após este procedimento tem-se $n-1$ seqüências de DNA. A Figura 5.1 mostra um exemplo deste processo. No caso de 5 seqüências de DNA (*A,*

B, C, D e *E*), se *B* e *C* são escolhidas,obtêm-se quatro novas seqüências de DNA (*A, BC, D* e *E*). Depois, após a coalescência de *D* e *E*, tem-se três seqüências (*A, BC* e *DE*). Quando *A* e *BC* são escolhidas,obtém-se a relação genealógica de cinco seqüências de DNA mostrada na Figura 5.1. Desta maneira, pode-se obter muitas árvores genealógicas de n seqüências de DNA.

O próximo passo é gerar o número de mutações em cada ramo. Seja Y_{ij}, definido na Seção 4.1.1, o número de mutações ocorrendo no j-ésimo ramo dentre os i ramos existentes existentes durante o tempo de coalescência T_i (Figura 5.1), e seja Y_i o número total de mutações em i ramos.

$$Y_i = \sum_{j=1}^{i} Y_{ij}, \quad i = 2, 3, \ldots \tag{5.1}$$

Em (4.2) foi visto que Y_i segue uma distribuição Geométrica.

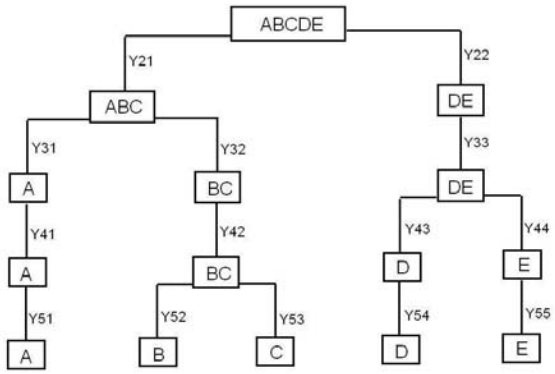


Figura 5.1: Relação evolucionária entre cinco genes
usada para explicar o processo de simulação.

A distribuição conjunta de $Y_{i1}, Y_{i2}, \ldots, Y_{ii}$ para um certo valor de Y_i é uma Multinomial:

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{ii} = y_{ii} \mid Y_i = y_i) = \frac{y_i!}{\prod_{j=1}^{i} y_{ij}!} \left(\frac{1}{i}\right)^{y_i}. \tag{5.2}$$

Primeiro, gera-se Y_i de acordo com (4.2). Depois, Y_{ij}'s são obtidos de acordo com

(5.2).

5.3 Distribuição Empírica da Estatística do Teste de Tajima

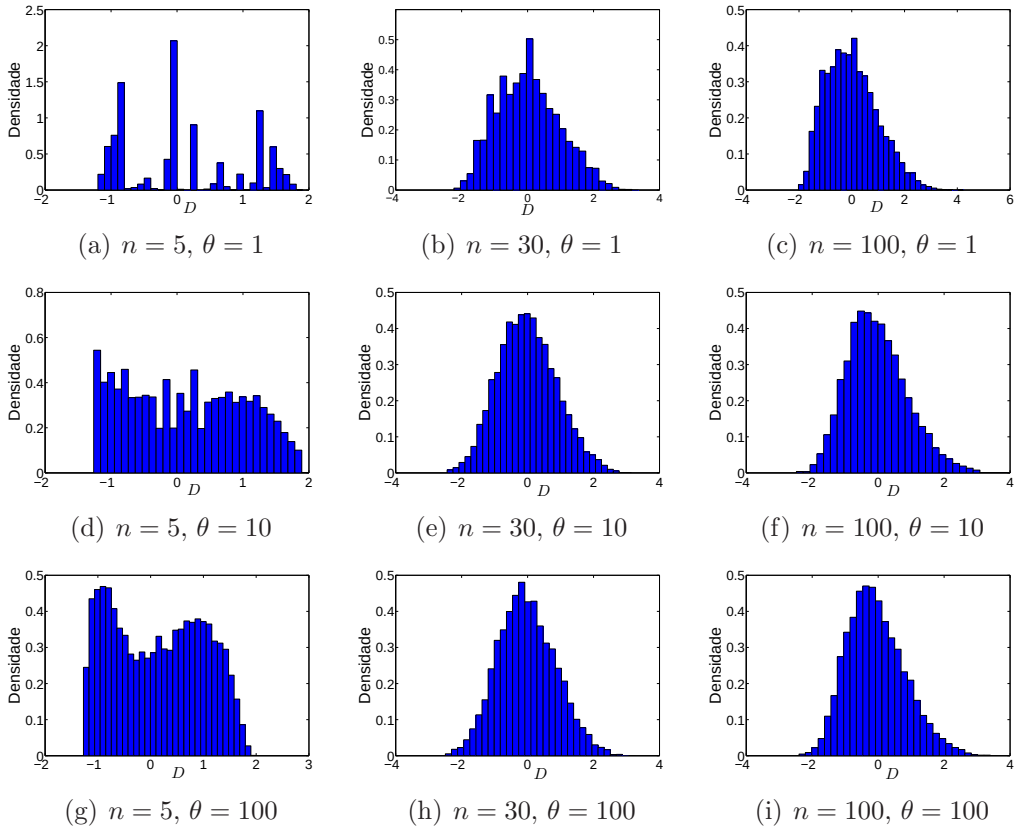
Como visto na Seção 2.5.2, a estatística do teste de Tajima para neutralidade de mutações é:

$$D = \frac{\mathcal{T}_2 - \mathcal{T}_1}{\sqrt{\widehat{\mathrm{Var}}(\mathcal{T}_2 - \mathcal{T}_1)}}\,.$$

Com a árvore pronta e as mutações em cada ramo simuladas, calcula-se o número de sítios segregantes ($V_n = Y_2 + \ldots + Y_n$) e o número médio de diferenças de nucleotídeos ($\bar{K}$).

Na simulação, foram utilizados os mesmos valores de θ (1, 10 e 100) usados por Tajima (1989) e os valores de n (5, 30 e 100). Foram gerados 10000 valores da estatística D para cada um destes casos.

Os gráficos com a distribuição de D sob H_0 para diversos valores de n e θ obtidos através das simulações estão apresentados na Figura 5.2.



(a) $n = 5$, $\theta = 1$ (b) $n = 30$, $\theta = 1$ (c) $n = 100$, $\theta = 1$

(d) $n = 5$, $\theta = 10$ (e) $n = 30$, $\theta = 10$ (f) $n = 100$, $\theta = 10$

(g) $n = 5$, $\theta = 100$ (h) $n = 30$, $\theta = 100$ (i) $n = 100$, $\theta = 100$

Figura 5.2: Distribuição de D.

As médias e os desvios padrão da estatística D para cada caso são apresentadas na Tabela 5.1, indicando que, sob H_0, a estatística D possui variância inferior a 1. Deve-se observar que a variância em geral diminui com o aumento do parâmetro θ.

Através dos histogramas, percebe-se que a distribuição de D não apresenta simetria, portanto, não seguiria a distribuição Normal.

5.4 Distribuição Empírica dos Estimadores quando a Genealogia é Conhecida

Os estimadores apresentados no Capítulo 2 não requerem o conhecimento da genealogia da amostra. Aqueles apresentados no Capítulo 4, no entanto, exigem o

Tabela 5.1: Médias e Desvios-Padrão da estatística D.

n	Estatísticas	θ		
		1	10	100
5	Média	0,1065	0,1279	0,1443
	Desvio Padrão	0,8872	0,8866	0,8610
30	Média	0,0166	-0,0064	-0,0223
	Desvio Padrão	0,9484	0,8819	0,8704
100	Média	0,0101	-0,0259	-0,0600
	Desvio Padrão	0,9726	0,9097	0,8756

conhecimento da estrutura da árvore genealógica.

O procedimento apresentado nas seções 5.2 e 5.3 são suficientes para a obtenção dos estimadores $\mathcal{T}_1$, baseado no número de sítios segregantes, $\mathcal{T}_2$, número médio de diferenças em pares de nucleotídeos e $\mathcal{T}_3$, baseado no número de *singletones*.

Para obter a distribuição empírica dos estimadores baseados nos modelos lineares: $\widetilde{\theta}_m$, $\widetilde{\theta}_r$ e $\widetilde{\theta}_v$, é preciso obter, para cada árvore simulada, os valores de c_{ik} e s_{ik}, definidos em (4.4) e na Seção 4.2.2, respectivamente. Calculadas essas quantidades, pode-se obter m_i, o número de mutações em cada ramo; r_i, o número de mutações nos ramos de tamanho i e v_i, o número de sítios segregantes nos quais as freqüências de dois nucleotídeos segregantes são i e $n-i$, respectivamente.

Note que, no caso do modelo linear para $\mathbf{m}$, o vetor $\mathbf{x}$ e a matriz $\mathbf{Z}$ envolvem os valores de c_{ik}, sendo diferentes para diferentes topologias. Nos demais modelos lineares, os valores dos vetores $\boldsymbol{\alpha}$ e $\boldsymbol{\phi}$ e das matrizes $\boldsymbol{\Sigma}$ e $\boldsymbol{\Lambda}$ dependem apenas de n. Assim, para este modelos, uma vez fixados n e θ, obtém-se diretamente a variância do estimador. No modelo que envolve $\mathbf{m}$, a variância tomada é a média das variâncias obtidas através das simulações.

As Tabelas 5.2-5.6 apresentam os resultados das simulações. Foram geradas 10000 árvores para cada tamanho amostral n e θ considerados. Para cada árvore simulada, calcularam-se as estimativas: $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, $\widetilde{\theta}_m$, $\widetilde{\theta}_r$ e $\widetilde{\theta}_v$. As estimativas apresentadas são as médias das estimativas obtidas nas simulações. As variâncias teóricas apresentadas nas tabelas são dadas por (2.26), (2.63), (2.91), pela média de

(4.14) obtidas nas simulações, por (4.23) e (4.31), respectivamente. As variâncias teóricas estão apresentadas graficamente na Figura 5.8. O cálculo da variância estimada é feito pela substituição de θ por $\mathcal{T}_1$ e de θ^2 por $S(S-1)/(a_n^2+b_n)$ em (2.26), (2.63), (2.91) no caso dos estimadores $\mathcal{T}_1$, $\mathcal{T}_2$ e $\mathcal{T}_3$ para cada amostra simulada e depois suas respectivas médias foram tomadas. A escolha de $\mathcal{T}_1$ para estimar θ no cálculo das variâncias desses estimadores se dá porque $\mathcal{T}_1$ tem a menor variância dentre esses três estimadores considerados (Figura 5.8). No caso dos estimadores baseados em modelos lineares, as variâncias estimadas são as médias do cálculo de (4.16), (4.24) e (4.32) para cada amostra simulada.

As Figuras 5.3-5.7 apresentam a distribuição empírica dos estimadores. Para facilitar a comparação do comportamento destes estimadores, foram utilizadas densidades estimadas. Nota-se que o estimador $\mathcal{T}_3$ é o que apresenta cauda mais longa e $\widetilde{\theta}_m$, a mais curta.

Tabela 5.2: Estatísticas Sumárias dos Estimadores para $\theta = 2$

$\theta = 2$								
n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	V_{min}
5	Estimativa	2,01	2,07	1,79	2,02	2,01	2,00	
	Variância Teórica	2,27	2,47	5,01	2,13	2,19	2,26	2,11
	Variância estimada	2,29	2,49	5,05	2,15	2,19	2,25	
	Variância amostral	2,28	2,64	3,57	2,15	2,17	2,23	
10	Estimativa	1,99	2,04	1,81	1,99	1,99	1,93	
	Variância Teórica	1,48	1,93	4,22	1,33	1,38	1,46	1,32
	Variância estimada	1,46	1,91	4,17	1,32	1,36	1,39	
	Variância amostral	1,44	1,98	3,27	1,32	1,34	1,42	
30	Estimativa	2,02	2,07	1,91	2,02	2,01	1,99	
	Variância Teórica	0,92	1,67	3,09	0,80	0,83	0,88	0,79
	Variância estimada	0,93	1,69	3,12	0,80	0,83	0,87	
	Variância amostral	0,93	1,77	2,78	0,82	0,84	0,89	
50	Estimativa	2,00	2,05	1,92	2,00	2,00	1,98	
	Variância Teórica	0,77	1,62	2,74	0,67	0,69	0,73	0,66
	Variância estimada	0,77	1,62	2,74	0,66	0,69	0,72	
	Variância amostral	0,76	1,68	2,40	0,65	0,67	0,70	
100	Estimativa	1,99	2,03	1,94	1,99	1,99	1,97	
	Variância Teórica	0,63	1,59	2,43	0,54	0,56	0,59	0,54
	Variância estimada	0,62	1,57	2,41	0,54	0,56	0,58	
	Variância amostral	0,62	1,59	2,27	0,54	0,56	0,58	

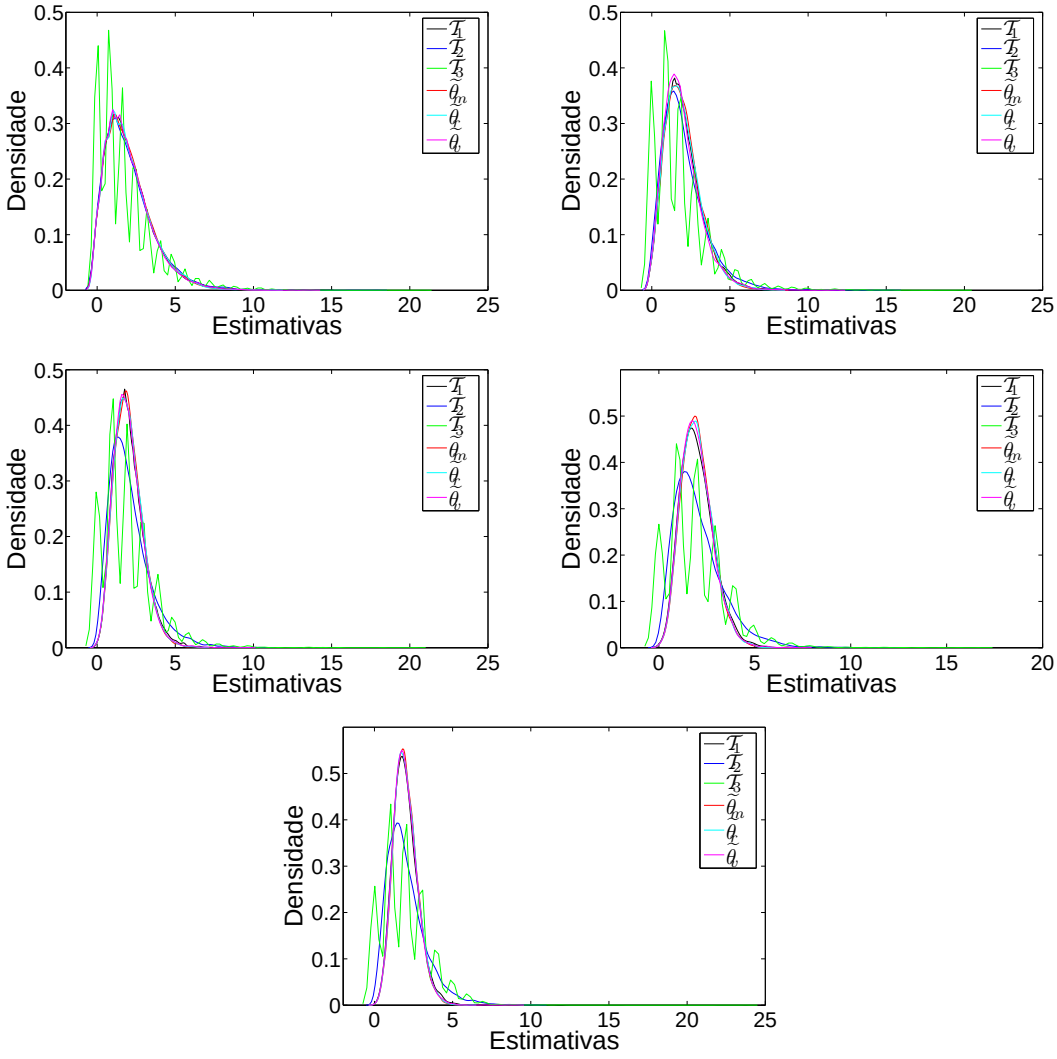


Figura 5.3: Comportamento dos estimadores quando $\theta = 2$ e n=5, 10, 30, 50 e 100, respectivamente.

Tabela 5.3: Estatísticas Sumárias dos Estimadores para $\theta = 5$

n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	V_{min}
5	Estimativa	5,01	5,14	4,50	5,03	5,01	4,97	
	Variância Teórica	10,60	11,67	21,92	9,28	9,67	10,54	9,16
	Variância estimada	10,61	11,68	21,94	9,33	9,62	10,39	
	Variância amostral	10,52	12,20	17,08	9,27	9,48	10,26	
10	Estimativa	5,01	5,16	4,49	5,00	4,98	4,85	
	Variância Teórica	6,58	9,01	18,04	5,24	5,58	6,40	5,16
	Variância estimada	6,61	9,06	18,12	5,22	5,51	6,13	
	Variância amostral	6,67	9,71	13,79	5,25	5,51	6,40	
30	Estimativa	5,02	5,14	4,76	5,01	5,00	4,95	
	Variância Teórica	3,83	7,74	11,53	2,76	2,97	3,51	2,73
	Variância estimada	3,84	7,77	11,57	2,76	2,95	3,45	
	Variância amostral	3,81	8,07	9,99	2,78	2,89	3,43	
50	Estimativa	5,02	5,14	4,87	5,01	5,00	4,96	
	Variância Teórica	3,14	7,52	9,48	2,21	2,37	2,79	2,18
	Variância estimada	3,15	7,55	9,51	2,21	2,36	2,76	
	Variância amostral	3,09	7,75	8,94	2,21	2,34	2,76	
100	Estimativa	5,01	5,11	4,91	5,02	5,00	4,98	
	Variância Teórica	2,49	7,37	7,61	1,72	1,83	2,13	1,70
	Variância estimada	2,50	7,39	7,63	1,72	1,83	2,11	
	Variância amostral	2,50	7,58	7,42	1,74	1,84	2,17	

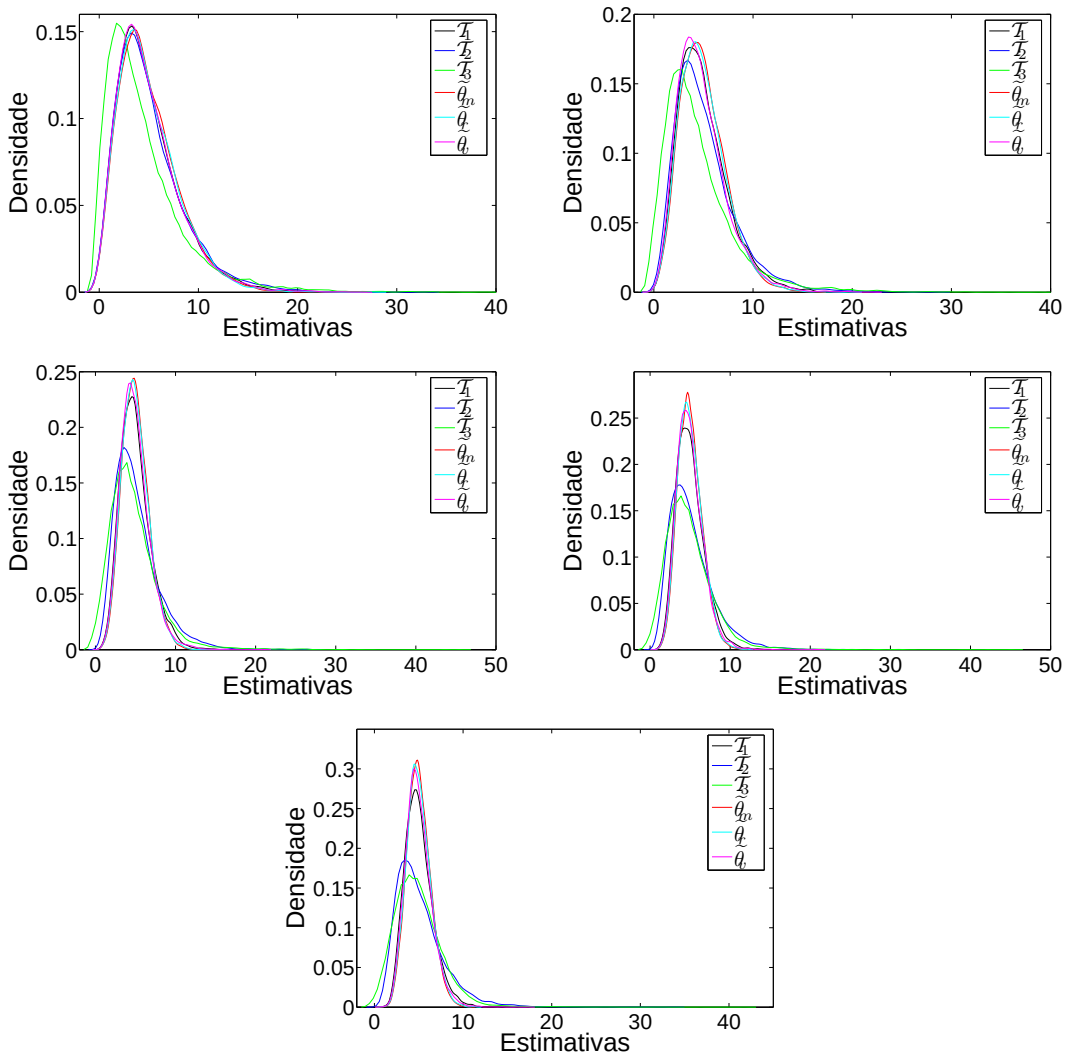


Figura 5.4: Comportamento dos estimadores quando $\theta = 5$ e n=5, 10, 30, 50 e 100, respectivamente.

Tabela 5.4: Estatísticas Sumárias dos Estimadores para $\theta = 10$

n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	V_{min}
5	Estimativa	10,00	10,29	8,86	10,00	9,98	9,91	
	Variância Teórica	37,60	41,67	75,17	31,33	32,63	37,34	31,00
	Variância estimada	34,65	41,73	75,27	31,29	32,33	36,68	
	Variância amostral	37,76	44,42	60,21	31,25	32,09	36,61	
10	Estimativa	9,96	10,22	9,05	9,99	9,93	9,67	
	Variância Teórica	22,77	31,98	61,03	16,41	17,59	22,01	16,16
	Variância estimada	22,70	31,87	60,83	16,36	17,36	20,98	
	Variância amostral	23,25	33,94	49,39	16,65	17,67	22,58	
30	Estimativa	9,98	10,24	9,36	9,99	9,93	9,81	
	Variância Teórica	12,79	27,39	35,76	7,67	8,46	11,33	7,55
	Variância estimada	12,76	27,22	35,56	7,64	8,35	10,95	
	Variância amostral	12,48	27,28	30,46	7,77	8,28	11,03	
50	Estimativa	10,02	10,31	9,60	10,00	9,96	9,88	
	Variância Teórica	10,33	26,63	27,72	5,85	6,47	8,67	5,77
	Variância estimada	10,37	26,74	27,82	5,83	6,42	8,50	
	Variância amostral	10,42	28,14	25,33	5,79	6,32	8,58	
100	Estimativa	9,98	10,15	9,83	9,99	9,96	9,91	
	Variância Teórica	8,03	26,08	20,34	4,32	4,76	6,25	4,27
	Variância estimada	8,00	25,96	20,27	4,30	4,73	6,17	
	Variância amostral	7,92	26,06	19,83	4,39	4,77	6,34	

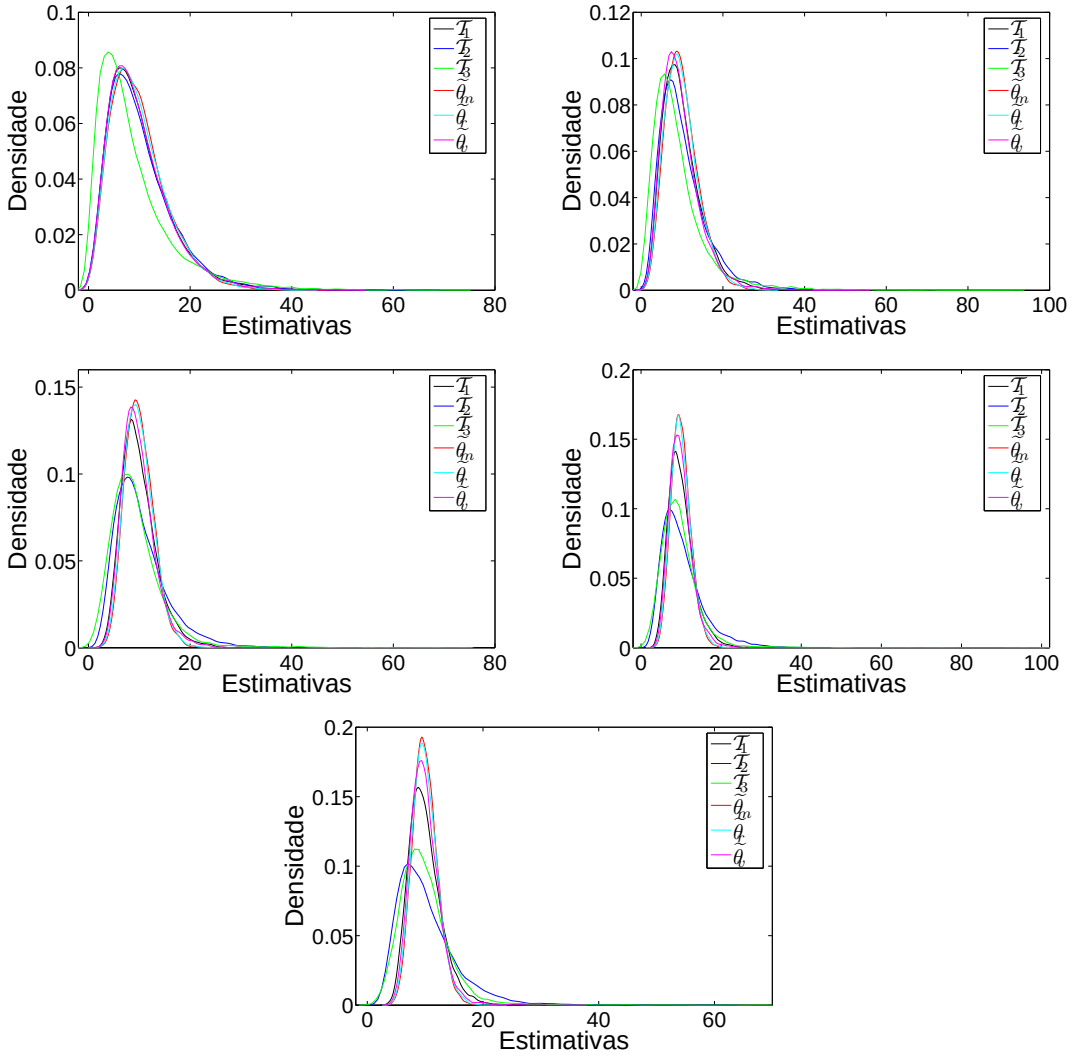


Figura 5.5: Comportamento dos estimadores quando $\theta = 10$ e n=5, 10, 30, 50 e 100, respectivamente.

Tabela 5.5: Estatísticas Sumárias dos Estimadores para $\theta = 20$

n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	V_{min}
5	Estimativa	20,06	20,62	17,81	20,10	20,06	19,87	
	Variância Teórica	140,80	156,67	275,67	113,06	116,88	139,72	112,22
	Variância estimada	140,89	156,76	275,86	113,80	116,71	137,09	
	Variância amostral	138,62	163,28	225,99	112,74	114,33	134,40	
10	Estimativa	19,89	20,39	18,21	19,94	19,84	19,32	
	Variância Teórica	84,03	119,75	221,91	55,65	59,64	80,79	54,96
	Variância estimada	82,93	118,17	219,03	55,31	58,57	76,17	
	Variância amostral	81,63	122,85	179,40	55,35	57,73	79,29	
30	Estimativa	20,06	20,49	19,18	20,06	19,95	19,79	
	Variância Teórica	46,11	102,45	122,37	23,07	26,20	39,73	22,69
	Variância estimada	46,43	103,18	123,19	23,14	26,03	39,01	
	Variância amostral	46,99	107,32	108,45	22,89	25,48	39,71	
50	Estimativa	20,04	20,49	19,36	20,06	19,97	19,79	
	Variância Teórica	36,86	99,57	90,48	16,65	19,22	29,51	16,38
	Variância estimada	37,05	100,09	90,52	16,71	19,16	28,98	
	Variância amostral	37,59	104,50	81,58	17,00	19,29	29,46	
100	Estimativa	20,01	20,49	19,57	20,00	19,92	19,80	
	Variância Teórica	28,26	97,51	61,17	11,51	13,38	20,22	11,35
	Variância estimada	28,28	97,60	61,21	11,50	13,28	19,88	
	Variância amostral	28,41	102,35	61,60	11,77	13,37	20,56	

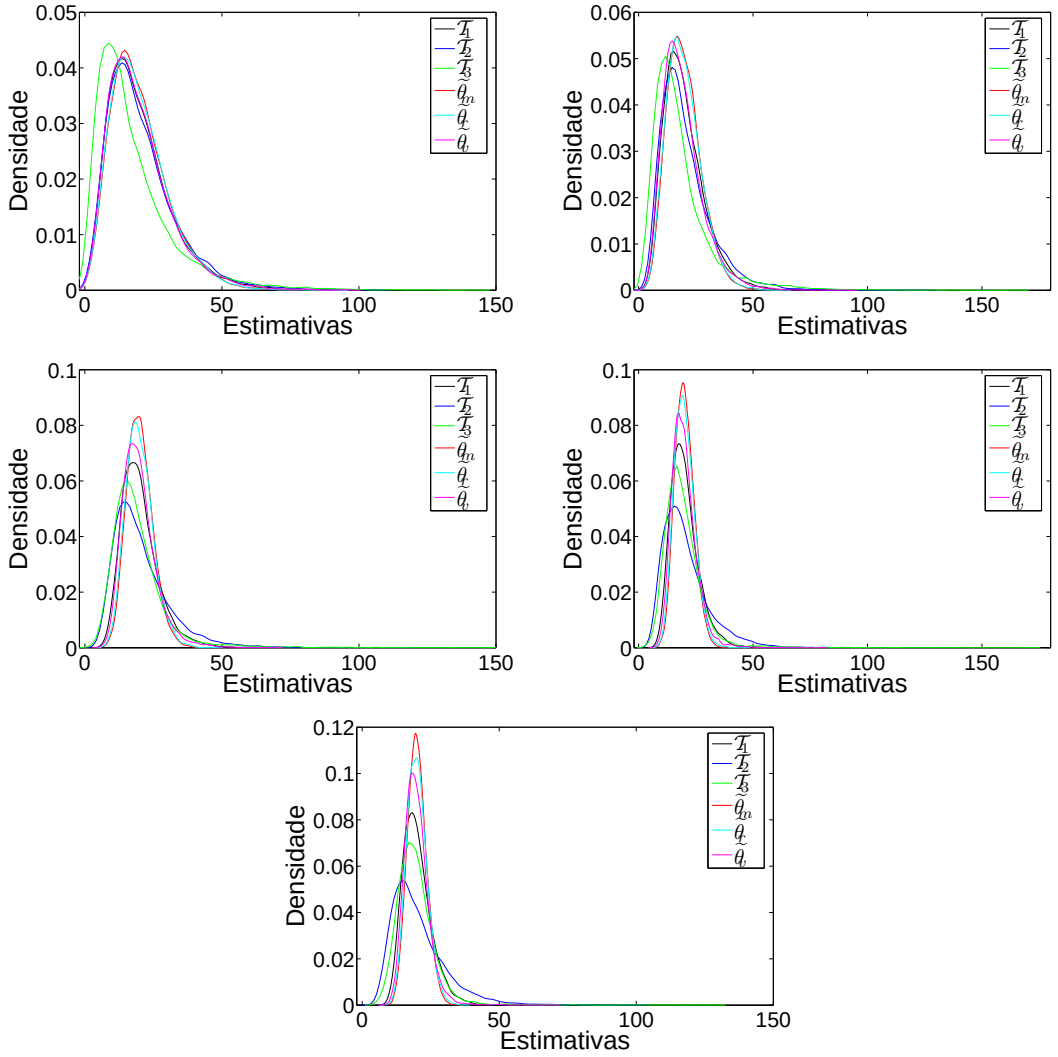


Figura 5.6: Comportamento dos estimadores quando $\theta = 20$ e n=5, 10, 30, 50 e 100, respectivamente.

Tabela 5.6: Estatísticas Sumárias dos Estimadores para $\theta = 30$

n		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	V_{min}
5	Estimativa	30,14	30,93	26,96	29,96	29,89	29,87	
	Variância Teórica	309,60	345,00	601,50	244,85	251,89	307,13	243,46
	Variância estimada	311,55	347,17	605,26	243,08	248,82	303,42	
	Variância amostral	309,01	362,65	522,08	238,95	244,93	300,36	
10	Estimativa	29,77	30,50	27,14	29,97	29,80	28,88	
	Variância Teórica	183,76	263,33	482,63	117,22	125,47	176,34	116,03
	Variância estimada	180,69	258,90	474,61	117,01	123,87	165,90	
	Variância amostral	179,09	260,72	403,15	118,07	124,18	177,03	
30	Estimativa	30,07	30,90	28,68	30,08	29,86	29,58	
	Variância Teórica	99,96	225,17	259,81	45,61	52,80	85,15	44,90
	Variância estimada	100,52	226,46	261,23	45,80	52,35	83,14	
	Variância amostral	101,60	239,64	234,40	46,46	52,60	85,79	
50	Estimativa	30,06	30,81	29,11	29,96	29,82	29,68	
	Variância Teórica	79,58	218,82	188,28	31,84	37,89	62,42	31,32
	Variância estimada	79,88	219,64	188,95	31,75	37,47	61,26	
	Variância amostral	79,69	229,62	187,03	32,12	37,25	62,67	
100	Estimativa	29,94	30,74	29,16	29,95	29,82	29,58	
	Variância Teórica	60,69	214,30	122,47	21,08	25,55	41,72	20,75
	Variância estimada	60,52	213,70	122,14	21,00	25,28	40,64	
	Variância amostral	61,80	225,59	112,48	21,39	25,31	40,34	

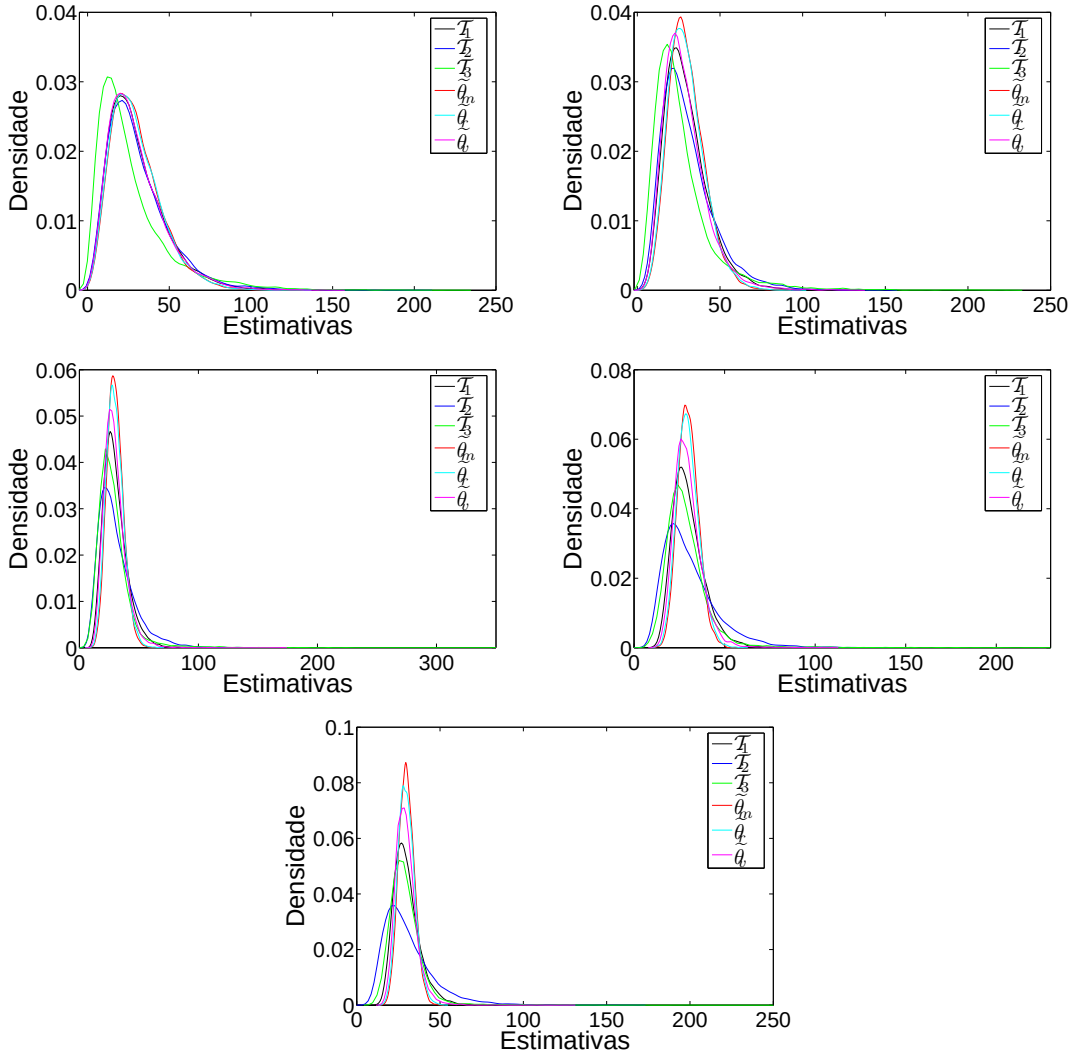


Figura 5.7: Comportamento dos estimadores quando $\theta = 30$ e n=5, 10, 30, 50 e 100, respectivamente.

Observando as Figuras 5.3-5.7, percebe-se que $\mathcal{T}_3$ é o estimador que apresenta o pior comportamento, com cauda muito longa em todos os tamanhos amostrais considerados. O comportamento dos demais estimadores se diferencia à medida que o tamanho amostral cresce. Nos casos em que θ é pequeno, percebe-se que $\mathcal{T}_1$ tem comportamento parecido com aquele dos estimadores baseados em modelos lineares. Além disso, pelas tabelas e figuras, observa-se que $\widetilde{\theta}_m$ é o estimador que apresenta melhor comportamento, sendo bem mais eficaz do que aqueles baseados nos métodos de momentos.

Os gráficos da Figura 5.8 mostram que as variâncias teóricas dos estimadores $\mathcal{T}_1$, $\widetilde{\theta}_m$, $\widetilde{\theta}_r$ e $\widetilde{\theta}_v$ são próximas quando o parâmetro θ é pequeno, o que era esperado (ver equação (4.33)). Uma diferença mais evidente na eficiência destes estimadores se dá quando o parâmetro é maior.

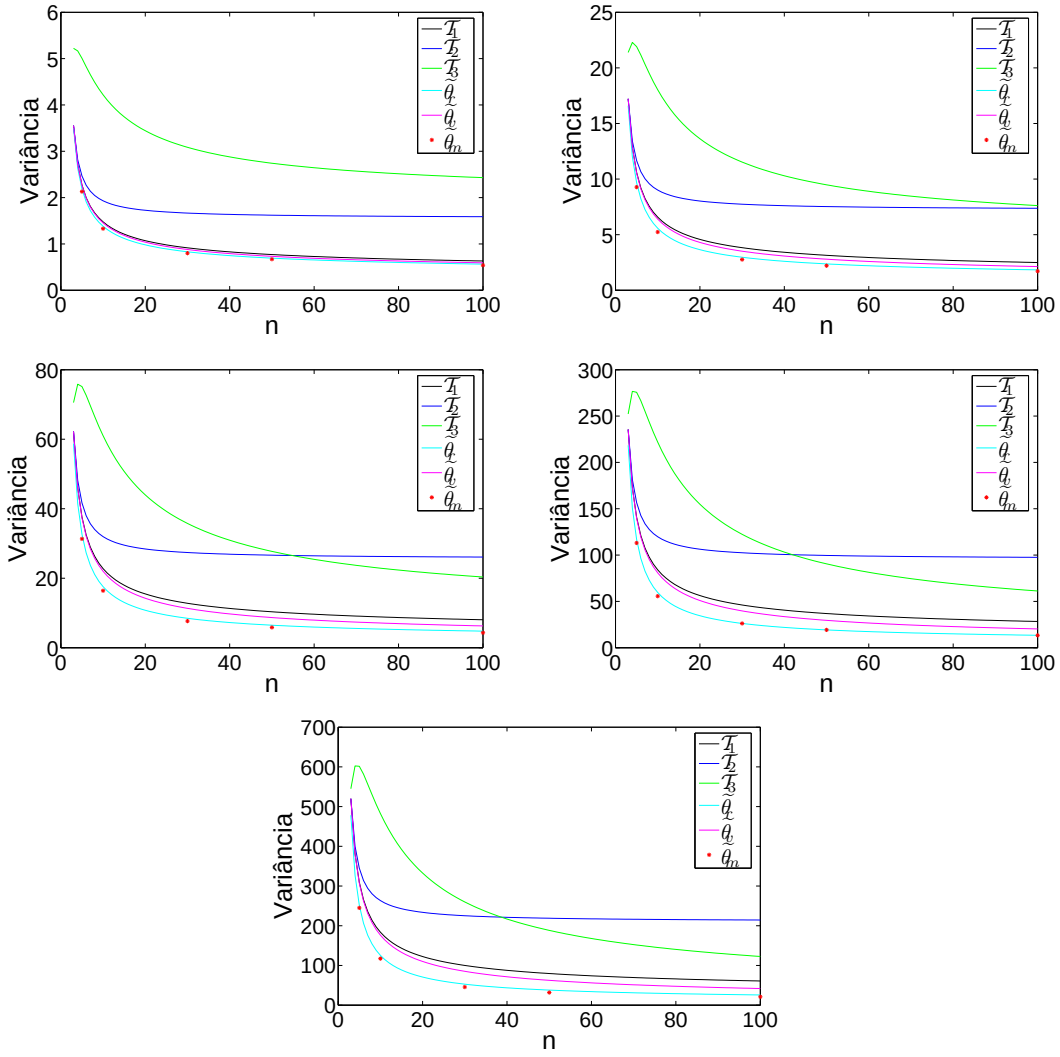


Figura 5.8: Variâncias Teóricas em função do tamanho amostral, n para $\theta = 2$, 5, 10, 20 e 30, respectivamente.

5.5 Distribuição dos Estimadores quando a Genealogia é Desconhecida

No caso dos estimadores obtidos através do método dos momentos ($\mathcal{T}_1$, $\mathcal{T}_2$ e $\mathcal{T}_3$), a genealogia não precisa ser conhecida. Na Seção 4.2 foram estudados o comportamento dos estimadores baseados em modelos lineares quando a genealogia é conhecida, além do comportamento dos estimadores baseados nos métodos dos momentos. Como a genealogia de uma amostra de genes é em geral desconhecida, deve-se estimar a genealogia para depois efetuar os cálculos dos estimadores $\widetilde{\theta}_m$, $\widetilde{\theta}_r$ e $\widetilde{\theta}_v$. Os erros na reconstrução da árvore provocam um viés na estimação de θ. A utilidade dos estimadores baseados em modelos lineares depende, desta maneira, da possibilidade de se corrigir este viés. O viés na estimação de θ depende do método utilizado para reconstruir a genealogia de uma amostra. O método UPGMA, apresentado na Seção 3.6, é um método simples e, portanto, computacionalmente vantajoso. Este método é eficiente sob a suposição de taxa constante de evolução, o que é válido sob o modelo de Wright-Fisher.

Simulações foram feitas para estudar o comportamento dos estimadores quando a genealogia é estimada. Para certos valores de n e θ, foram simuladas amostras e, para cada uma, calculou-se a matriz de distância utilizada no método UPGMA. A matriz de distâncias contém o número de diferenças em pares de nucleotídeos. Para cada árvore obtida pelo método UPGMA, calculam-se o número de mutações que ocorreu no ramo i (m_i), x_i e z_{ij} através de (4.6) e (4.7) para o modelo linear definido em (4.11); o número de mutações que ocorreu nos ramos de tamanho i (r_i), α_i e σ_{ij} através de (4.18) e (4.19) para o modelo linear definido em (4.20); o número de sítios segregantes nos quais as freqüências de dois nucleotídeos segregantes são i e $n-i$ (v_i), ϕ_i e λ_{ij} definidos em (4.27) e (4.28) para o modelo linear apresentado em (4.29). O procedimento iterativo é então aplicado para cada um dos modelos, obtendo-se, dessa maneira, as respectivas estimativas.

A Figura 5.9 apresenta os resultados das simulações para o modelo linear baseado no número de mutações que ocorreu em cada ramo, $\mathbf{m}$. Para cada valor de θ e n

considerados, foram geradas 2000 amostras. A Figura 5.9 mostra que $\widetilde{\theta}_m$ subestima θ. Uma análise de regressão mostra que a seguinte equação de regressão descreve a relação entre θ, n e a média de $\widetilde{\theta}_m$:

$$\widetilde{\theta}_m = \left(-0,03425\sqrt{n-2} + 1,00575\sqrt{\theta}\right)^2 . \tag{5.3}$$

Assim, uma estimativa aproximadamente não viesada para θ seria:

$$\widetilde{\theta}_{m,\,c} = \left(0,03405\sqrt{n-2} + 0,99428\sqrt{\widetilde{\theta}_m}\right)^2 . \tag{5.4}$$

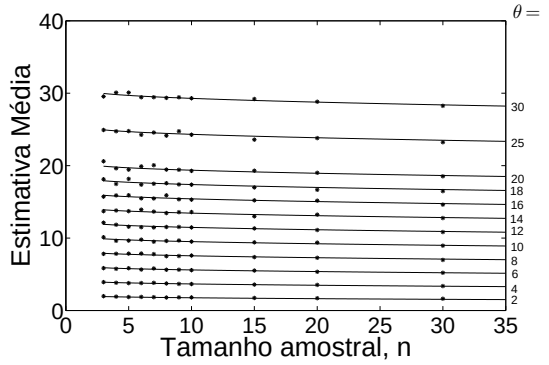


Figura 5.9: Relação entre θ, n e a média de $\widetilde{\theta}_m$.
Cada * é a média de $\widetilde{\theta}_m$ nas 2000 simulações. As curvas são dadas por (5.3), fixando θ.

A Figura 5.10 apresenta os resultados das simulações para o modelo linear baseado no número de mutações que ocorreu no ramo de tamanho i, r_i. O procedimento foi o mesmo aplicado no modelo para $\mathbf{m}$. Uma análise de regressão mostra que a seguinte equação de regressão descreve a relação entre θ, n e a média de $\widetilde{\theta}_r$:

$$\widetilde{\theta}_r = \left(-0,02478\sqrt{n-2} + 1,00380\sqrt{\theta}\right)^2 . \tag{5.5}$$

Assim, uma estimativa aproximadamente não viesada para θ seria:

$$\widetilde{\theta}_{r,\,c} = \left(0,02468\sqrt{n-2} + 0,99621\sqrt{\widetilde{\theta}_r}\right)^2 . \tag{5.6}$$

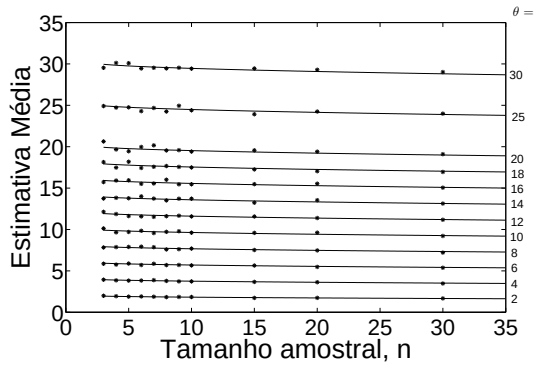


Figura 5.10: Relação entre θ, n e a média de $\widetilde{\theta}_r$.
Cada * é a média de $\widetilde{\theta}_r$ nas 2000 simulações. As curvas são dadas por (5.5), fixando θ.

Aplicando o mesmo procedimento para o estimador $\widetilde{\theta}_v$, tem-se a equação de regressão seguinte:

$$\widetilde{\theta}_v = \left(-0,00813\sqrt{n-2} + 0,98481\sqrt{\theta}\right)^2 . \tag{5.7}$$

A Figura 5.11 mostra que para tamanhos amostrais pequenos (até 10) o ajuste não é satisfatório, com exceção dos casos em que θ é pequeno. As estimativas corrigidas através da equação:

$$\widetilde{\theta}_{v,\,c} = \left(0,00826\sqrt{n-2} + 1,01542\sqrt{\widetilde{\theta}_v}\right)^2 . \tag{5.8}$$

estão apresentadas nas Tabelas 5.7-5.11.

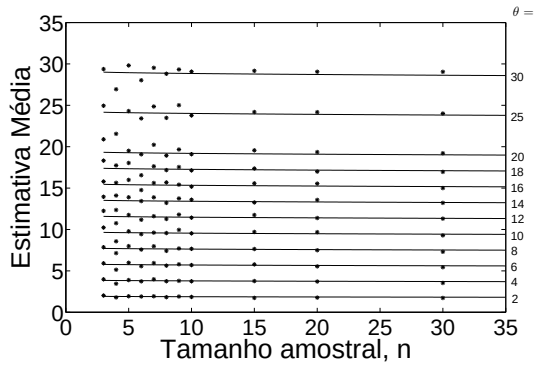


Figura 5.11: Relação entre θ, n e a média de $\widetilde{\theta}_v$.
Cada * é a média de $\widetilde{\theta}_v$ nas 2000 simulações. As curvas são dadas por (5.7), fixando θ.

Desta maneira, temos os seguintes passos para estimar θ:

1. Calcular a matriz de distâncias.
2. Obter a genealogia através do método UPGMA.
3. Obter $\widetilde{\theta}_m$ ($\widetilde{\theta}_r$, $\widetilde{\theta}_v$).
4. Usar a equação (5.4) ((5.6), (5.8)) para obter uma estimativa aproximadamente não viesada para θ.
5. Calcular a variância da estimativa através da equação (4.16) ((4.24),(4.32)) substituindo $\widetilde{\theta}_m$ ($\widetilde{\theta}_r$, $\widetilde{\theta}_v$) por $\widetilde{\theta}_{m,c}$ ($\widetilde{\theta}_{r,c}$, $\widetilde{\theta}_{v,c}$) dado por (5.4) ((5.6), (5.8)).

As Tabelas 5.7-5.11 apresentam as estatísticas sumárias das estimativas baseadas em 10000 amostras simuladas para os tamanhos amostrais 5 e 10 e 2000 amostras para os demais tamanhos amostrais devido ao custo computacional. As três primeiras linhas de cada tamanho amostral considerado referem-se às estatísticas antes da correção do viés.

Tabela 5.7: Estatísticas Sumárias dos Estimadores após UPGMA para $\theta = 2$

n		$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	Variância Mínima
5	Estimativa	1,91	1,94	1,98	
	Variância estimada	1,98	2,07	2,22	
	Variância amostral	2,02	2,05	2,22	
	Estimativa Corrigida	2,04	2,03	2,08	
	Variância estimada	2,17	2,21	2,40	2,11
	Variância amostral	2,16	2,15	2,41	
10	Estimativa	1,80	1,84	1,86	
	Variância estimada	1,15	1,22	1,31	
	Variância amostral	1,19	1,23	1,37	
	Estimativa Corrigida	2,04	2,01	1,98	
	Corrigida 1 a 1	1,35	1,38	1,44	1,32
	Variância amostral	1,34	1,35	1,50	
30	Estimativa	1,63	1,67	1,73	
	Variância estimada	0,61	0,65	0,72	
	Variância amostral	0,60	0,63	0,74	
	Estimativa Corrigida	2,09	2,00	1,89	
	Variância estimada	0,84	0,82	0,82	0,79
	Variância amostral	0,76	0,76	0,83	
50	Estimativa	1,61	1,64	1,69	
	Variância estimada	0,51	0,54	0,58	
	Variância amostral	0,50	0,53	0,60	
	Estimativa Corrigida	2,22	2,08	1,89	
	Variância estimada	0,76	0,72	0,68	0,66
	Variância amostral	0,69	0,67	0,70	

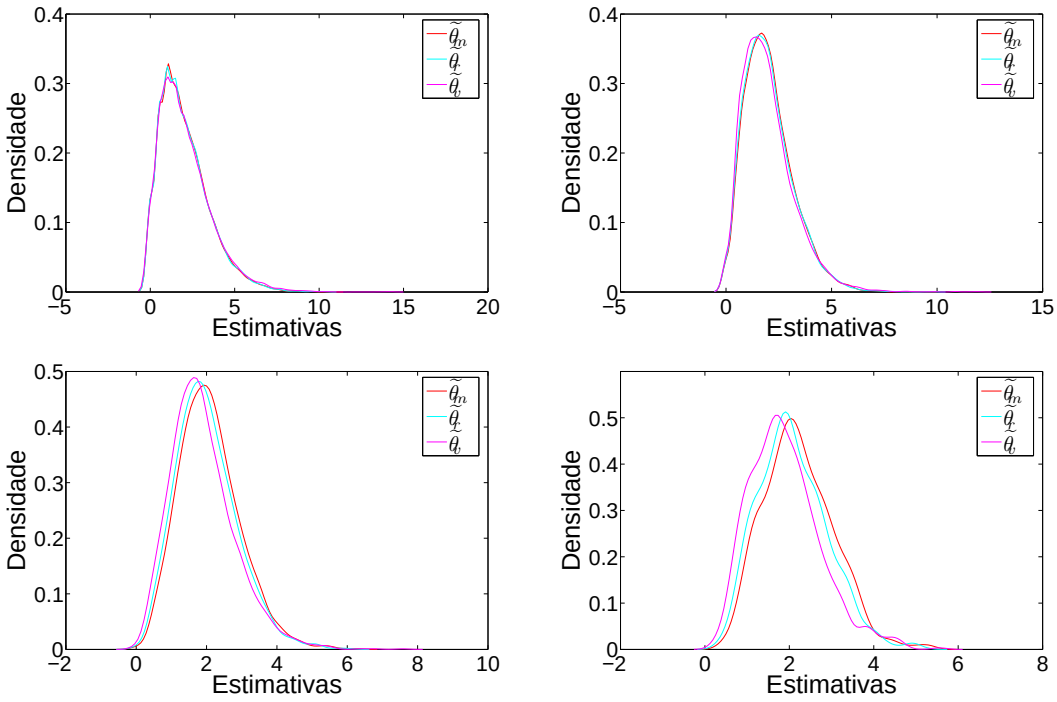


Figura 5.12: Comportamento dos estimadores quando $\theta = 2$ e n=5, 10, 30 e 50, respectivamente.

Tabela 5.8: Estatísticas Sumárias dos Estimadores após UPGMA para $\theta = 5$

n		$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	Variância Mínima
5	Estimativa	4,79	4,84	4,91	
	Variância estimada	8,71	9,11	10,21	
	Variância amostral	9,04	9,16	10,23	
	Estimativa Corrigida	4,99	4,98	5,12	
	Variância estimada	9,23	9,52	11,00	9,16
	Variância amostral	9,31	9,37	11,01	
10	Estimativa	4,62	4,70	4,70	
	Variância estimada	4,65	5,03	5,84	
	Variância amostral	5,00	5,18	6,32	
	Estimativa Corrigida	4,97	4,96	4,94	
	Variância estimada	5,20	5,46	6,37	5,16
	Variância amostral	5,34	5,43	6,86	
30	Estimativa	4,34	4,45	4,51	
	Variância estimada	2,26	2,51	3,01	
	Variância amostral	2,54	2,77	3,43	
	Estimativa Corrigida	5,05	4,98	4,84	
	Variância estimada	2,80	2,94	3,36	2,73
	Variância amostral	2,93	3,07	3,79	
50	Estimativa	4,16	4,28	4,35	
	Variância estimada	1,72	1,90	2,25	
	Variância amostral	1,76	1,89	2,27	
	Estimativa Corrigida	5,11	4,98	4,73	
	Variância estimada	2,27	2,34	2,56	2,18
	Variância amostral	2,15	2,18	2,54	

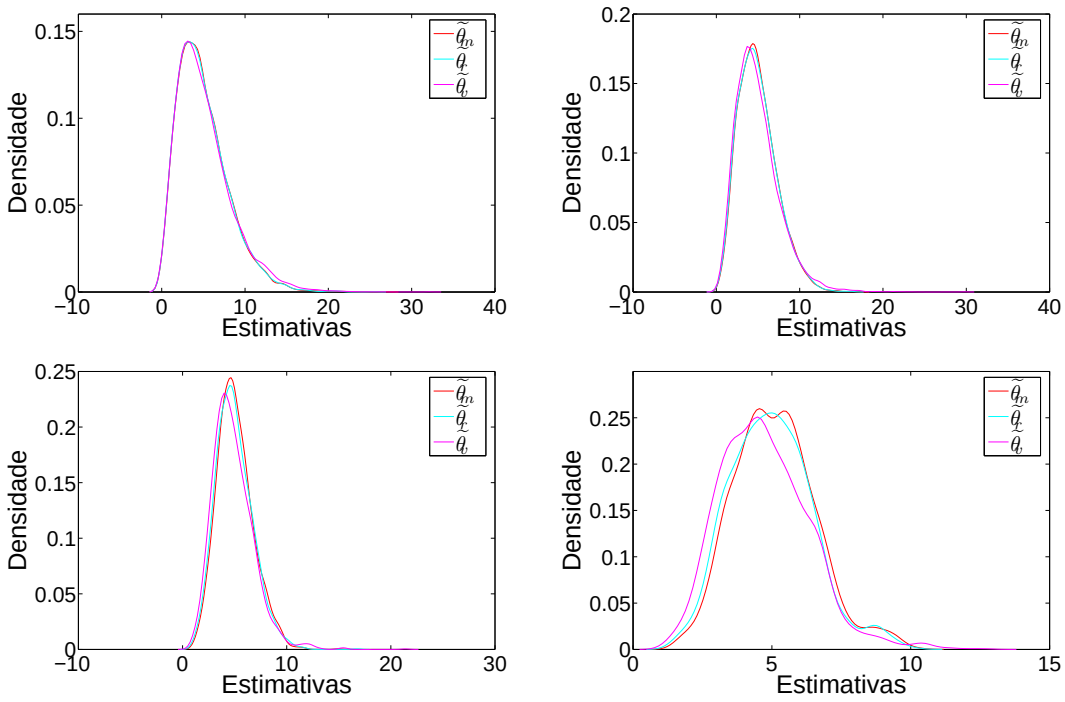


Figura 5.13: Comportamento dos estimadores quando $\theta = 5$ e n=5, 10, 30 e 50, respectivamente.

Tabela 5.9: Estatísticas Sumárias dos Estimadores após UPGMA para $\theta = 10$

n		$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	Variância Mínima
5	Estimativa	9,74	9,78	9,83	
	Variância estimada	30,03	31,20	36,03	
	Variância amostral	30,75	31,00	35,57	
	Estimativa Corrigida	9,98	9,96	10,23	
	Variância estimada	31,20	32,11	38,74	31,00
	Variância amostral	31,18	31,36	38,14	
10	Estimativa	9,48	9,60	9,52	
	Variância estimada	15,07	16,38	20,45	
	Variância amostral	16,15	16,72	22,33	
	Estimativa Corrigida	9,95	9,95	9,96	
	Variância estimada	16,30	17,35	22,18	16,16
	Variância amostral	16,78	17,21	24,08	
30	Estimativa	8,95	9,21	9,28	
	Variância estimada	6,51	7,45	10,03	
	Variância amostral	7,30	8,02	11,00	
	Estimativa Corrigida	9,94	9,94	9,84	
	Variância estimada	7,61	8,37	11,09	7,55
	Variância amostral	8,02	8,59	12,02	
50	Estimativa	8,87	9,23	9,33	
	Variância estimada	4,93	5,75	7,81	
	Variância amostral	5,74	6,55	9,65	
	Estimativa Corrigida	10,20	10,21	9,97	
	Variância estimada	6,04	6,68	8,74	5,77
	Variância amostral	6,53	7,19	10,62	

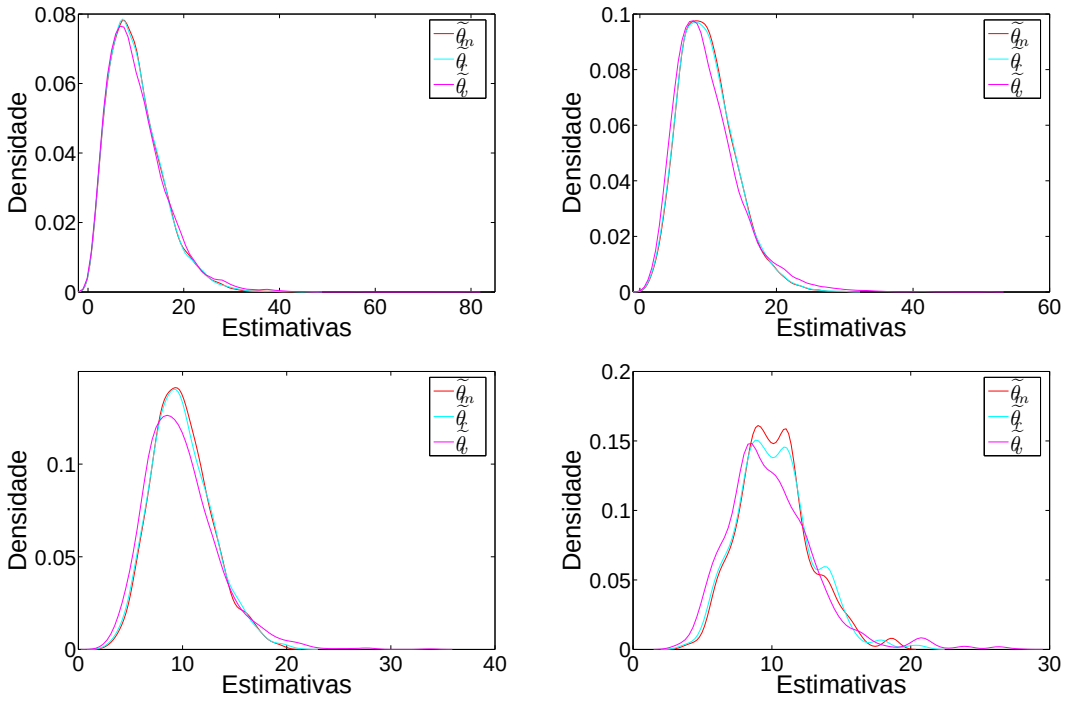


Figura 5.14: Comportamento dos estimadores quando $\theta = 10$ e n=5, 10, 30 e 50, respectivamente.

Tabela 5.10: Estatísticas Sumárias dos Estimadores após UPGMA para $\theta = 20$

n		$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	Variância Mínima
5	Estimativa	19,66	19,681	19,69	
	Variância estimada	109,12	112,17	133,60	
	Variância amostral	107,91	108,004	127,56	
	Estimativa Corrigida	19,94	19,899	20,43	
	Variância estimada	111,56	114,128	143,29	112,22
	Variância amostral	108,24	108,401	136,44	
10	Estimativa	19,45	19,59	19,19	
	Variância estimada	53,24	57,42	75,71	
	Variância amostral	56,51	58,05	81,25	
	Estimativa Corrigida	20,07	20,05	19,99	
	Variância estimada	56,12	59,72	81,75	54,96
	Variância amostral	57,65	58,97	87,23	
30	Estimativa	18,53	19,10	19,21	
	Variância estimada	20,39	24,22	36,96	
	Variância amostral	21,84	24,47	38,12	
	Estimativa Corrigida	19,88	20,10	20,20	
	Variância estimada	22,84	26,36	40,51	22,69
	Variância amostral	23,17	25,56	41,28	
50	Estimativa	17,96	18,96	18,57	
	Variância estimada	14,14	17,27	25,82	
	Variância amostral	16,57	19,66	26,23	
	Estimativa Corrigida	19,79	20,05	19,65	
	Variância estimada	16,42	19,32	28,57	16,38
	Variância amostral	18,08	20,93	28,60	

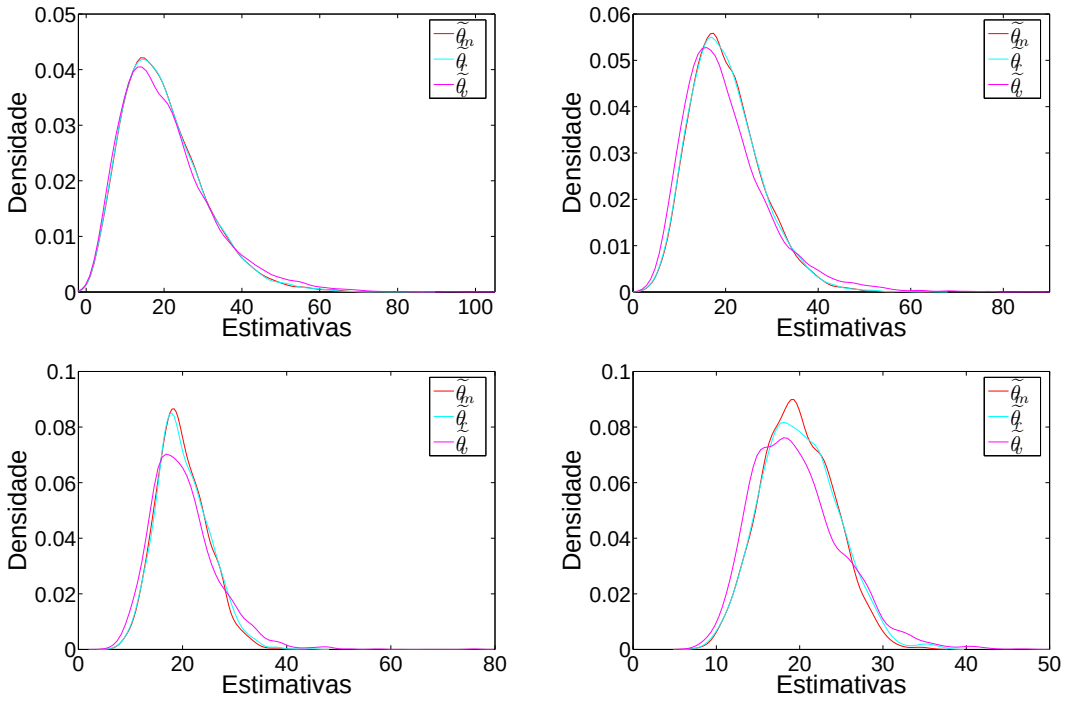


Figura 5.15: Comportamento dos estimadores quando $\theta = 20$ e n=5, 10, 30 e 50, respectivamente.

Tabela 5.11: Estatísticas Sumárias dos Estimadores após UPGMA para $\theta = 30$

n		$\widetilde{\theta}_m$	$\widetilde{\theta}_r$	$\widetilde{\theta}_v$	Variância Mínima
5	Estimativa	29,55	29,55	29,41	
	Variância estimada	236,98	242,51	292,47	
	Variância amostral	233,81	234,67	283,49	
	Estimativa Corrigida	29,83	29,78	30,47	
	Variância estimada	240,41	245,37	313,27	243,46
	Variância amostral	233,40	234,70	302,86	
10	Estimativa	29,34	29,46	28,87	
	Variância estimada	112,75	120,99	165,89	
	Variância amostral	115,54	119,05	177,70	
	Estimativa Corrigida	30,04	29,99	30,02	
	Variância estimada	117,39	124,73	178,70	116,03
	Variância amostral	116,97	120,27	190,42	
30	Estimativa	28,26	29,02	29,05	
	Variância estimada	41,22	49,74	80,42	
	Variância amostral	42,44	49,32	84,17	
	Estimativa Corrigida	29,86	30,21	30,43	
	Variância estimada	45,21	53,30	87,73	44,90
	Variância amostral	44,36	50,95	90,84	
50	Estimativa	27,53	28,39	28,35	
	Variância estimada	27,75	34,49	56,12	
	Variância amostral	30,76	35,14	55,63	
	Estimativa Corrigida	29,72	30,01	29,85	
	Variância estimada	31,39	37,86	61,74	31,32
	Variância amostral	32,83	36,87	60,36	

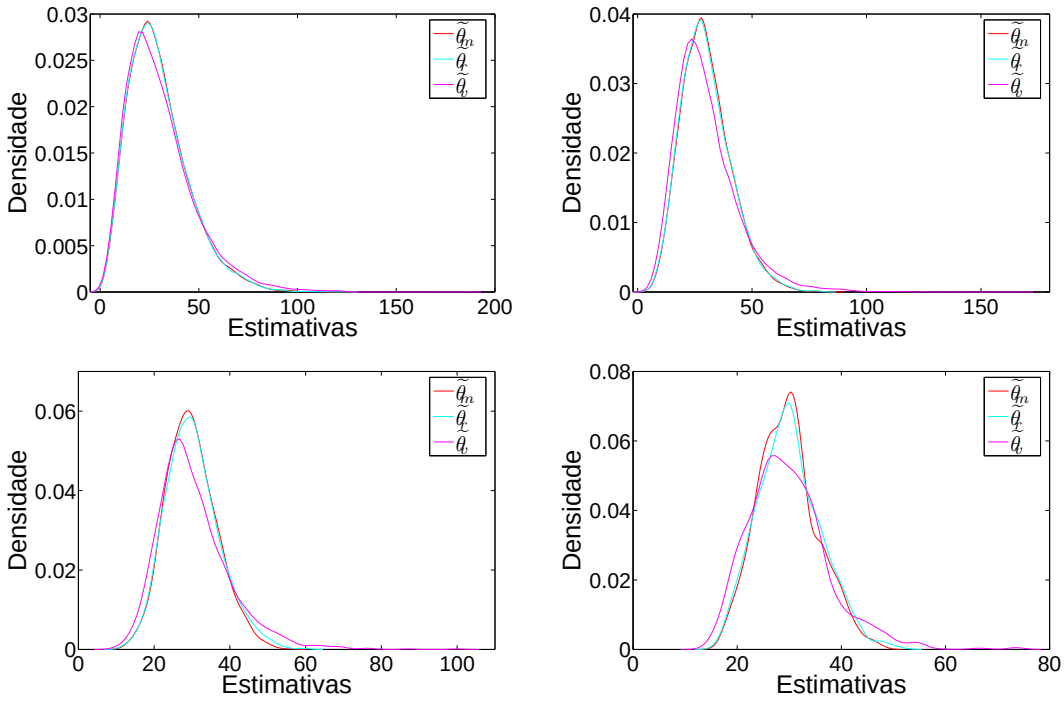


Figura 5.16: Comportamento dos estimadores quando $\theta = 30$ e n=5, 10, 30 e 50, respectivamente.

5.6 Estimador de Máxima Verossimilhança

Foi visto na Seção 4.1 que o estimador de máxima verossimilhança para θ é a solução da equação:

$$\sum_{i=2}^{n} \frac{y_i + 1}{\widehat{\theta}_m + i - 1} = \frac{V_n}{\widehat{\theta}_m}.$$

Para encontrar tal estimador pode-se utilizar o método do *score* de Fisher. As iterações são calculadas através de:

$$\theta_{(k+1)} = \theta_{(k)} + \frac{\theta_{(k)}}{\sum_{i=1}^{n-1} \frac{1}{\theta_{(k)}+i}} \sum_{i=2}^{n} \left(\frac{y_i}{\theta_{(k)}} - \frac{y_i + 1}{\theta_{(k)} + i - 1} \right).$$

A Tabela 5.12 apresenta resultados obtidos através da simulação. Foram geradas 10000 árvores para diversos tamanhos de amostra e $\theta = 2$. Considera-se aqui o caso em que a genealogia é conhecida. Verifica-se que a variância amostral das estimativas é bem próxima ao LICR.

Tabela 5.12: Estatísticas Sumárias do Estimador de Máxima Verossimilhança para $\theta = 2$

	n				
	5	10	30	50	100
Média	1,9995	2,0132	2,0029	2,0029	2,0017
Variância amostral	2,1046	1,3366	0,7856	0,6490	0,5590

Os gráficos com a distribuição do estimador de máxima verossimilhança para diversos valores de n e $\theta = 2$ obtidos através das simulações quando a genealogia é conhecida estão apresentados na Figura 5.17.

A Tabela 5.13 apresenta resultados obtidos através da simulação para diversos valores de n e $\theta = 20$. Considera-se aqui o caso em que a genealogia é conhecida.

Os gráficos com a distribuição do estimador de máxima verossimilhança para diversos valores de n e $\theta = 20$ obtidos através das simulações quando a genealogia é conhecida estão apresentados na Figura 5.18.

Tabela 5.13: Estatísticas Sumárias do Estimador de Máxima Verossimilhança para $\theta = 20$

	n				
	5	10	30	50	100
Média	20,0199	20,0223	20,0514	19,9570	19,9960
Variância amostral	112,3705	55,4026	22,6455	15,9563	11,3140

5.7 Aplicação

Nesta Seção apresentam-se exemplos da aplicação do teste de neutralidade seletiva proposto por Tajima (1989).

Um dos conjuntos de dados estudados é uma amostra da população de cágados da espécie *Hydromedusa maximiliani*. Esta espécie habita rios e riachos de águas rasas e claras de regiões montanhosas e é endêmica da floresta Atlântica. Estes ecossistemas apresentam uma das mais elevadas taxas de destruição e fragmentação e, como a espécie é altamente sensível a efeitos estocásticos demográficos e ambientais, esta encontra-se ameaçada de extinção. Quando animais da mesma espécie se encontram isolados geograficamente (por rios, montanhas, etc) tendem a se diferenciar, pois a mudança de ambiente favorece a ação da seleção natural, o que pode levar a uma mudança na composição genética.

O conjunto de dados consiste de seqüências da região do DNA mitocondrial: o gene do citocromo b, dos 48 cágados amostrados. Na região gene do citocromo b, temos 262 posições seqüenciadas.

Primeiramente, aplica-se o teste de Tajima, para testar a hipótese de neutralidade seletiva. Tem-se 48 cágados (n=48) e 262 sítios, destes apenas 9 são sítios polimórficos, ou seja, $S = 9$ e 7 são *singletones* ($\mathcal{S}^* = 7$). O número de seqüências únicas nesta amostra é 11. Para uma melhor aproximação do modelo de sítios infinitos, utilizam-se somente essas seqüências. O número médio de diferenças de nucleotídeos $\bar{K}$ é 2,1455.

A estatística D de Tajima obtida para esta amostra é -1,2682.

Para gerar a distribuição empírica da estatística D, sob H_0, para esta amostra, utilizamos o método *bootstrap*; o procedimento está descrito a seguir:

Passo 1: Identificar os sítios polimórficos e utilizar apenas estes. Temos S sítios polimórficos.

Passo 2: Calcular as freqüências relativas de cada tipo de nucleotídeo (A = adenina, C = citosina, T = timina e G = guanina) presente em cada sítio.

Passo 3: Gerar S vetores aleatórios de dimensão 11 (número de seqüências utilizadas), onde cada elemento é gerado a partir de uma distribuição Multinomial com parâmetros sendo as freqüências relativas calculadas no passo 2.

Passo 4: Contar o número de sítios polimórficos da amostra *bootstrap* (S), calcular $\bar{K}$ e, finalmente, a estatística D de Tajima. Para calcular $\bar{K}$, Tajima (1989) sugere o seguinte método:

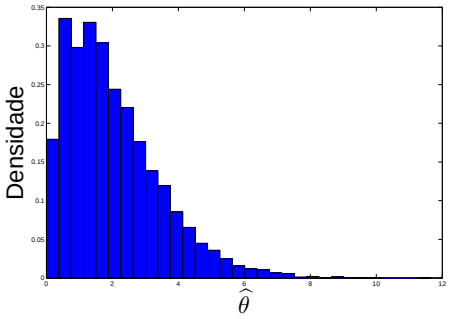
$$\bar{K} = \sum_{i=1}^{S} h_i,$$

onde h_i é a estimativa não viesada da heterozigosidade (ou diversidade de nucleotídeo), que é dada por:

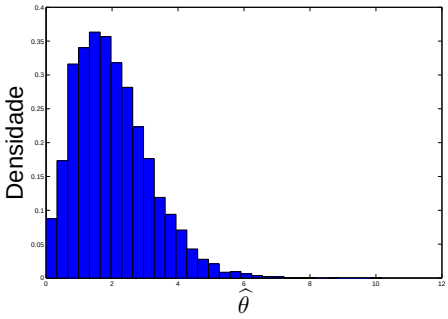
$$h_i = \frac{n(1 - \sum_j x_{ji}^2)}{n-1},$$

onde x_{ji} é a freqüência amostral do j-ésimo nucleotídeo no i-ésimo sítio polimórfico.

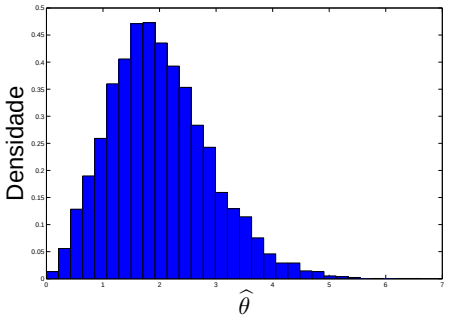
Passo 5: Repetir os passos 3 e 4 10000 vezes.



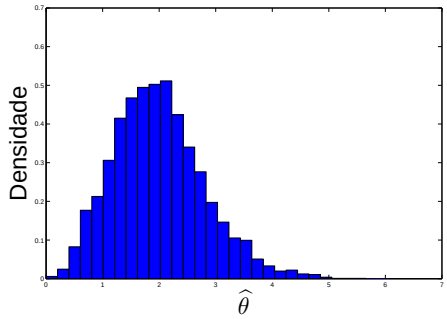
(a) $n = 5$



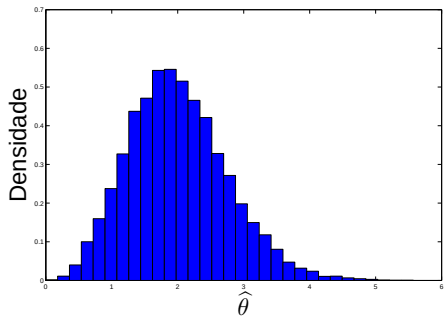
(b) $n = 10$



(c) $n = 30$

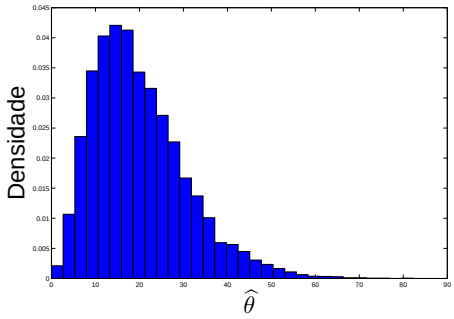


(d) $n = 50$

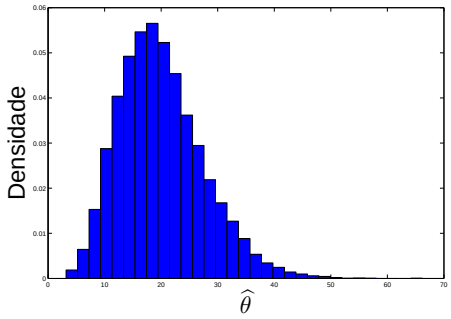


(e) $n = 100$

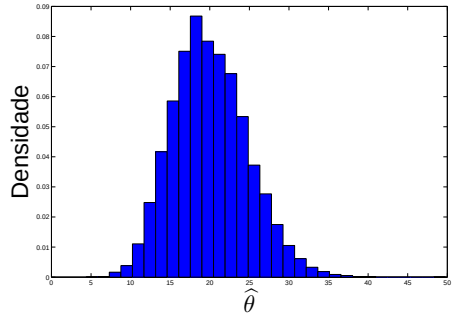
Figura 5.17: Distribuição dos Estimadores de Máxima Verossimilhança para $\theta = 2$.



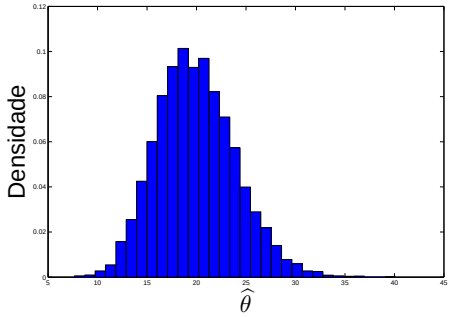
(a) $n = 5$



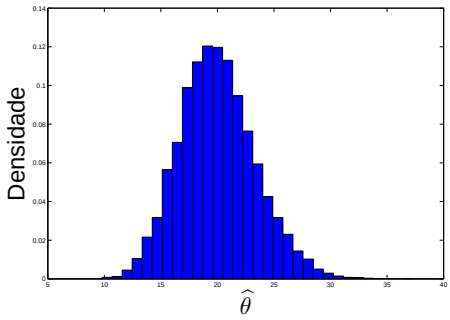
(b) $n = 10$



(c) $n = 30$



(d) $n = 50$



(e) $n = 100$

Figura 5.18: Distribuição dos Estimadores de Máxima Verossimilhança para $\theta = 20$.

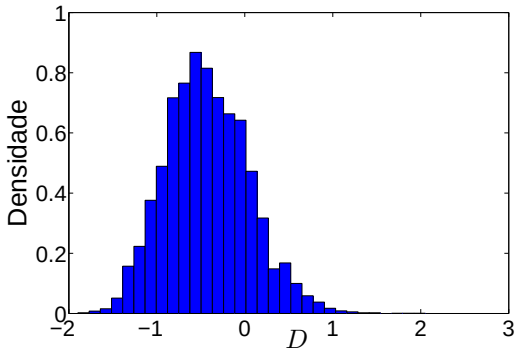


Figura 5.19: Distribuição empírica de D para a região do gene citocromo b.

Através da distribuição empírica de D, obtêm-se os percentis. O p-valor para a estatística do teste D observada (-1,2682) é 0,0397, desta maneira, encontram-se evidências estatísticas significantes para rejeitar a hipótese de neutralidade seletiva a um nível de 5% de significância. Assim, temos evidências de que o modelo de Wright-Fisher parece não ser válido para esta amostra. As estimativas obtidas e suas respectivas variâncias estimadas são:

	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_{m,\,c}$	$\widetilde{\theta}_{r,\,c}$	$\widetilde{\theta}_{v,\,c}$
Estimativa	3,0728	2,1455	6,3636	4,2044	3,8023	3,3177
Variância Estimada	4,0157	5,5772	11,4217	3,3938	3,5488	3,9173

As variâncias estimadas foram calculadas a partir de (2.26), (2.63) e (2.91), substituindo θ por $\widetilde{\theta}_{m,\,c}$ e θ^2 pela estimativa não viesada dada em (4.15) aplicando a correção:

$$\widehat{\theta^2} = \frac{\widetilde{\theta}_{m,\,c}(\widetilde{\theta}_{m,\,c} - \mathbf{u'D}_x\mathbf{u})}{1 + \mathbf{u'Zu}}\,.$$

Para o estimador $\widetilde{\theta}_{m,\,c}$, a variância é calculada por (4.16), substituindo $\widetilde{\theta}_m$ por $\widetilde{\theta}_{m,\,c}$. Para os demais modelos:

$$\widehat{\mathrm{Var}(\widetilde{\theta}_{r,\,c})} = (\mathbf{u'D}_\alpha\mathbf{u})\widetilde{\theta}_{m,\,c} + (\mathbf{u'\Sigma u})\frac{\widetilde{\theta}_{m,\,c}(\widetilde{\theta}_{m,\,c} - \mathbf{u'D}_x\mathbf{u})}{1 + \mathbf{u'Zu}}\,.$$

$$\widehat{\mathrm{Var}(\widetilde{\theta}_{v,\,c})} = (\mathbf{u'D}_\phi\mathbf{u})\widetilde{\theta}_{m,\,c} + (\mathbf{u'\Lambda u})\frac{\widetilde{\theta}_{m,\,c}(\widetilde{\theta}_{m,\,c} - \mathbf{u'D}_x\mathbf{u})}{1 + \mathbf{u'Zu}}\,.$$

O cálculos substituindo os parâmetros θ e θ^2 pelas respectivas estimativas baseadas em $\widetilde{\theta}_{m,\,c}$ se justifica pelo bom comportamento deste estimador, conforme observado nas simulações.

O segundo conjunto de dados consiste de 12 seqüências amostradas de um paciente com HIV estudado por Holmes e Brown (1992). O paciente não apresentava sintomas da doença e não foi submetido a nenhuma terapia antiviral. Para cada seqüência, foram seqüenciados 234 sítios. Têm-se 22 sítios polimórficos, dos quais 11 são *singletones*. O número médio de diferenças em pares de nucleotídeos é 6,97.

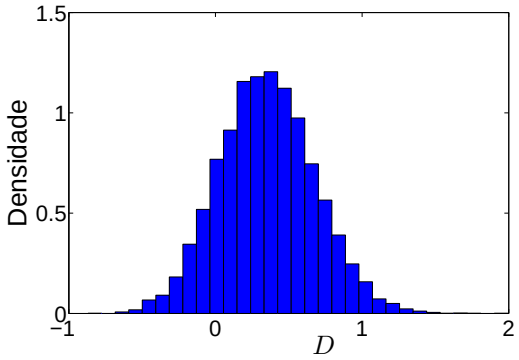


Figura 5.20: Distribuição empírica de D para as seqüências de HIV.

A estatística D de Tajima obtida para esta amostra é -0,1914, cujo p-valor, obtido através do método *bootstrap* descrito anteriormente é 0,7298. Assim, não foram encontradas evidências estatísticas significantes para rejeitar a hipótese de neutralidade seletiva. A Figura 5.20 apresenta a distribuição empírica de D sob H_0 para esta amostra.

As estimativas obtidas e suas respectivas variâncias estimadas são:

	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\widetilde{\theta}_{m,\,c}$	$\widetilde{\theta}_{r,\,c}$	$\widetilde{\theta}_{v,\,c}$
Estimativa	7,2851	6,9697	10,0833	19,1348	16,5045	12,7470
Variância Estimada	65,4138	96,5464	176,1505	40,2015	43,1519	58,9570

O estimador de máxima verossimilhança, obtido através da maximização de (4.2), para o conjunto de dados dos cágados é 4,0. A Figura 5.21 apresenta a função de verossimilhança para o conjunto de dados considerado. É importante ressaltar que neste caso a genealogia foi estimada através do método UPGMA. Para os estimadores baseados em modelos lineares, foi visto que após a estimação da genealogia a estimativa fica viesada. Simulações preliminares mostraram que o mesmo ocorre com o estimador de máxima verossimilhança. Estudos para a correção do viés devem ser feitos futuramente. No entanto, o estimador sem correção foi calculado para ilustração.

O estimador de máxima verossimilhança para o conjunto de dados do HIV é

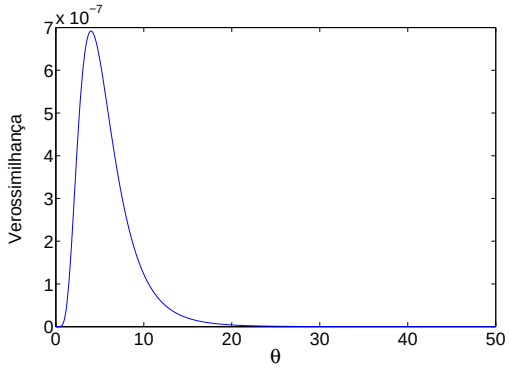


Figura 5.21: Função de verossimilhança para os dados dos cágados.

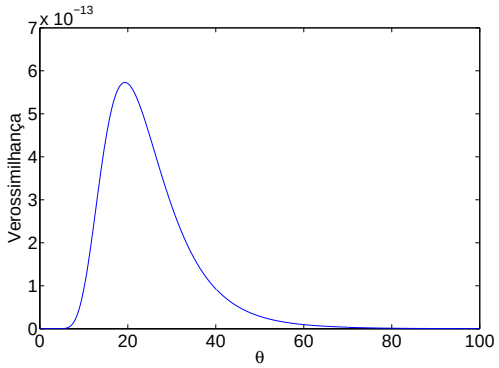


Figura 5.22: Função de verossimilhança para os dados do HIV.

19,40. A Figura 5.22 apresenta a função de verossimilhança para o conjunto de dados considerado.

Para o primeiro conjunto de dados, considera-se como estimativa do parâmetro θ o valor 4,20, cuja variância é 3,39. Para o conjunto de dados do paciente portador de HIV, a estimativa do parâmetro obtida é 19,13 com variância 40,20. As estimativas escolhidas foram aquelas fornecidas pelo modelo linear que envolve $\mathbf{m}$, o número de mutações em cada ramo.

6 Considerações Finais

Os resultados apresentados no Capítulo 5 mostraram que os modelos lineares aplicados na estimação de θ se mostraram eficientes. O ganho em eficiência na estimação ao se utilizar modelos lineares que incorporem a informação da árvore filogenética dos dados é grande. Mesmo com a utilização de um método de estimação de árvore bem simples, como o UPGMA, obtêm-se estimadores com variâncias muito próximas ao limite inferior de Cramer-Rao. Isso se explica pelo fato da estimação através destes modelos depender da determinação ou do número de mutações (m_i), ou do tamanho dos ramos (r_i) ou ainda dos tipos de ramos (v_i). A topologia da árvore não depende de θ. Viu-se que, segundo o modelo considerado, a construção da topologia se dá por bifurcações aleatórias, independentes do parâmetro de interesse. Assim, a topologia de uma árvore genealógica não contém informação sobre θ. A informação sobre este parâmetro está nos vetores $\mathbf{m}$, $\mathbf{r}$ e $\mathbf{v}$. Desta maneira, erros de reconstrução das árvore ocorrem, mas não afetam de maneira significativa a qualidade de estimação de θ, visto que não é a topologia em si que fornece informação sobre o parâmetro.

Dentre os modelos lineares, aquele que considera o vetor $\mathbf{v}$, ou seja, os tipos de sítios segregantes, é o menos eficaz, com maior variância. No caso deste estimador, após estimar a árvore por UPGMA, não se tem uma equação de regressão satisfatória para a correção do viés. O comportamento dos estimadores $\widetilde{\theta}_m$ e $\widetilde{\theta}_r$ é parecido, sendo que o segundo apresenta uma cauda um pouco mais longa. Assim, o melhor estimador seria $\widetilde{\theta}_m$.

É importante ressaltar que para valores pequenos de θ, $\mathcal{T}_1$ se aproxima dos esti-

madores baseados em modelos lineares. Pode-se notar isso na aplicação da metodologia para o conjunto de dados dos cágados. A estimativa de $\mathcal{T}_1$ está próxima àquela de $\widetilde{\theta}_v$, o mesmo pode-se observar com relação às suas variâncias. Um valor pequeno para θ é esperado para esse tipo de conjunto de dados, pois as seqüências são provenientes da região do citocromo b do DNA mitocondrial. Sabe-se que o DNA mitocondrial é aquele que apresenta menos variabilidade, desta maneira, espera-se que a estimativa para θ seja pequena. Com relação ao teste de neutralidade seletiva de Tajima, a estatística do teste observada apresentou valor negativo, estatisticamente significativo. Uma razão típica para isto pode ser tempo insuficiente desde o gargalo populacional para restaurar o equilíbrio entre mutação e deriva genética aleatória.

No segundo conjunto de dados apresentado, o valor da estimativa para θ é bem mais alto, coerente com o fato de estar-se lidando com seqüências de vírus, que apresentam taxa de mutação bem mais alta. É interessante observar que a diferença entre as variâncias das estimativas é bem maior nesse caso, com θ maior, como foi observado também nas simulações.

O estimador de máxima verossimilhança, apresentado na Seção 4.1, é mencionado na literatura apenas para a obtenção do LICR. Estudos utilizando este estimador diretamente não são apresentados. Um estudo preliminar deste estimador foi apresentado na Seção 5.6. O comportamento do estimador de máxima verossimilhança, obtido iterativamente através do método do *score* de Fisher, é muito similar ao comportamento do melhor estimador obtido através de modelos lineares, $\widetilde{\theta}_m$. No entanto, obter o estimador de máxima verossimilhança é um pouco mais simples, pois não é necessário inverter matrizes como no caso dos modelos lineares. O método do *score* de Fisher converge com pouquíssimas iterações quando a estimativa de $\mathcal{T}_1$ é utilizada como valor inicial. Os estudos aqui apresentados, no entanto, referem-se apenas ao caso em que a genealogia é conhecida. Simulações foram feitas no caso em que a genealogia é desconhecida e estimada através de UPGMA, indicando um viés, como no caso dos demais estimadores. É interessante que se estude futuramente o comportamento dessas estimativas no caso em que a genealogia é desconhecida e a

possibilidade de correção deste viés.

Referências

CAVALLI-SFORZA, L. L.; EDWARDS, A. W. Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet.*, v. 19, n. 3, p. 233–257, 1967.

DURRETT, R. *Probability Models for DNA Sequence Evolution*. [S.l.]: Springer, 2002.

FISHER, R. A. *The genetical theory of natural selection*. 1ª ed. Oxford: Clarendon Press, 1930.

FITCH, W. M. On the problem of discovering the most parsimonius tree. *Am. Nat.*, n. 111, p. 223–257, 1977.

FU, Y. X. A phylogenetic estimator of effective population size or mutation rate. *Genetics*, n. 136, p. 685–692, 1994.

FU, Y.-X. Statistical properties of segregating sites. *Theoretical Population Biology*, n. 48, p. 172–197, 1995.

FU, Y. X.; LI, W. H. Maximum likelihood estimation of population parameters. *Genetics*, n. 134, p. 1261–1270, 1993.

FU, Y. X.; LI, W. H. Statistical tests of neutrality of mutations. *Genetics*, n. 133, p. 693–709, 1993.

GIBSON, G.; MUSE, S. V. *A Primer of Genome Science*. 2nd ed. [S.l.]: Sinauer, 2004.

GRIFFITHS, A. J. F. et al. *An Introduction to Genetic Analysis*. New York: W. H. Freeman, 2000.

HARTL, D.; CLARK, A. *Principles of Population Genetics*. 3rd ed. Sunderland - Massachusetts: Sinauer Associates, Inc., 1997.

HOLMES, E. C.; BROWN, A. J. Convergent and divergent sequence evolution in the surface envelope of glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *PNAS*, n. 89, p. 4835–4839, 1992.

JUKES, T. H.; CANTOR, C. R. Evolution of protein molecules. In: *H. N. Munro.* New York: Academic Press, 1969, (Mammalian Protein Metabolism). p. 21–132.

KIMURA, M. Evolutionary rate at the molecular level. *Nature*, v. 217, p. 624–626, 1968.

KIMURA, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, v. 61, p. 624–626, 1969.

LI, W.-H.; GRAUR, D. *Fundamentals of Molecular Evolution*. 1st ed. Sunderland - Massachusetts: Sinauer Associates, Inc., 1991.

LOCKHART, P. J. et al. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, v. 11, n. 4, p. 605–612, 1994.

NAGYLAKI, T. *Introduction to Theoretical Population Genetics*. [S.l.]: Springer-Verlag, 1992.

NEI, M. *Molecular Population Genetics and Evolution*. Amsterdam: North-Holland, 1975.

PINHEIRO, H.; SEILLIER-MOISEIWITSCH, F.; SEN, P. K. Genomic sequences and quasi-multivariate catanova. In: *Handbook of Statistics*. [S.l.]: Elsevier Science Publishers, 2000, (Vol. 18). p. 713–746.

SAITOU, N.; NEI, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, v. 4, n. 4, p. 406–425, 1987.

SEILLIER-MOISEIWITSCH, F.; MARGOLIN, B. H.; SWANSTROM, R. Genetic variability of human immunodeficiency virus: Statistical and biological issues. *Annual Review of Genetics*, v. 28, p. 559–596, 1994.

SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, v. 28, p. 1409–1438, 1958.

TAJIMA, F. Evolutionary relationship of dna sequences in finite populations. *Genetics*, n. 105, p. 437–460, 1983.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, n. 123, p. 585–595, 1989.

TAKAHATA, N. Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. *Genet. Res.*, n. 39, p. 63–77, 1982.

WATSON, J. et al. *Recombinant DNA*. 2nd ed. New York: Scientific American Books, 1992.

WATTERMAN, M. S. *Introduction to Computational Biology: Maps, Sequences and Genomes*. [S.l.: s.n.], 1995.

WATTERSON, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, n. 7, p. 256–276, 1975.

WRIGHT, S. Evolution in mendelian populations. *Genetics*, v. 16, p. 97–159, 1931.